

RTGWG  
Internet-Draft  
Intended status: Informational  
Expires: January 16, 2014

S. Ning  
Tata Communications  
D. McDysan  
Verizon  
E. Osborne  
Cisco  
L. Yong  
Huawei USA  
C. Villamizar  
Outer Cape Cod Network  
Consulting  
July 15, 2013

Advanced Multipath Framework in MPLS  
draft-ietf-rtgwg-cl-framework-04

Abstract

This document specifies a framework for support of Advanced Multipath in MPLS networks. As defined in this framework, an Advanced Multipath consists of a group of homogenous or non-homogenous links that have the same forward adjacency (FA) and can be considered as a single TE link or an IP link when advertised into IGP routing.

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 16, 2014.

Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal

Provisions Relating to IETF Documents  
(<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

## Table of Contents

1. Introduction . . . . .	4
1.1. Background . . . . .	4
1.2. Architecture Summary . . . . .	4
1.3. Conventions used in this document . . . . .	5
1.4. Terminology . . . . .	5
1.5. Document Issues . . . . .	5
2. Advanced Multipath Key Characteristics . . . . .	7
2.1. Flow Identification . . . . .	7
2.1.1. Flow Identification Granularity . . . . .	8
2.1.2. Flow Identification Summary . . . . .	9
2.1.3. Flow Identification Using Entropy Label . . . . .	9
2.2. Advanced Multipath in Control Plane . . . . .	10
2.3. Advanced Multipath in Data Plane . . . . .	13
3. Architecture Tradeoffs . . . . .	14
3.1. Scalability Motivations . . . . .	14
3.2. Reducing Routing Information and Exchange . . . . .	15
3.3. Reducing Signaling Load . . . . .	15
3.3.1. Reducing Signaling Load using LDP MPTP . . . . .	16
3.3.2. Reducing Signaling Load using Hierarchy . . . . .	16
3.3.3. Using Both LDP MPTP and RSVP-TE Hierarchy . . . . .	17
3.4. Reducing Forwarding State . . . . .	17
3.5. Avoiding Route Oscillation . . . . .	17
4. New Challenges . . . . .	18
4.1. Control Plane Challenges . . . . .	19
4.1.1. Delay and Jitter Sensitive Routing . . . . .	19
4.1.2. Local Control of Traffic Distribution . . . . .	20
4.1.3. Path Symmetry Requirements . . . . .	20
4.1.4. Requirements for Contained LSP . . . . .	21
4.1.5. Retaining Backwards Compatibility . . . . .	21
4.2. Data Plane Challenges . . . . .	22
4.2.1. Very Large LSP . . . . .	22
4.2.2. Very Large Microflows . . . . .	23
4.2.3. Traffic Ordering Constraints . . . . .	23
4.2.4. Accounting for IP and LDP Traffic . . . . .	23
4.2.5. IP and LDP Limitations . . . . .	24
5. Existing Mechanisms . . . . .	25

5.1. Link Bundling . . . . .	25
5.2. Classic Multipath . . . . .	26
6. Mechanisms Proposed in Other Documents . . . . .	27
6.1. Loss and Delay Measurement . . . . .	27
6.2. Link Bundle Extensions . . . . .	28
6.3. Pseudowire Flow and MPLS Entropy Labels . . . . .	28
6.4. Multipath Extensions . . . . .	29
7. Required Protocol Extensions and Mechanisms . . . . .	29
7.1. Brief Review of Requirements . . . . .	29
7.2. Proposed Document Coverage . . . . .	30
7.2.1. Component Link Grouping . . . . .	31
7.2.2. Delay and Jitter Extensions . . . . .	31
7.2.3. Path Selection and Admission Control . . . . .	32
7.2.4. Dynamic Multipath Balance . . . . .	32
7.2.5. Frequency of Load Balance . . . . .	33
7.2.6. Inter-Layer Communication . . . . .	33
7.2.7. Packet Ordering Requirements . . . . .	33
7.2.8. Minimally Disruption Load Balance . . . . .	34
7.2.9. Path Symmetry . . . . .	34
7.2.10. Performance, Scalability, and Stability . . . . .	35
7.2.11. IP and LDP Traffic . . . . .	35
7.2.12. LDP Extensions . . . . .	35
7.2.13. Pseudowire Extensions . . . . .	36
7.2.14. Multi-Domain Advanced Multipath . . . . .	36
7.3. Framework Requirement Coverage by Protocol . . . . .	36
7.3.1. OSPF-TE and ISIS-TE Protocol Extensions . . . . .	37
7.3.2. PW Protocol Extensions . . . . .	37
7.3.3. LDP Protocol Extensions . . . . .	37
7.3.4. RSVP-TE Protocol Extensions . . . . .	37
7.3.5. RSVP-TE Path Selection Changes . . . . .	37
7.3.6. RSVP-TE Admission Control and Preemption . . . . .	37
7.3.7. Flow Identification and Traffic Balance . . . . .	37
8. IANA Considerations . . . . .	38
9. Security Considerations . . . . .	38
10. Acknowledgments . . . . .	38
11. References . . . . .	39
11.1. Normative References . . . . .	39
11.2. Informative References . . . . .	39
Authors' Addresses . . . . .	42

## 1. Introduction

Advanced Multipath functional requirements are specified in [I-D.ietf-rtgwg-cl-requirement]. Advanced Multipath use cases are described in [I-D.ietf-rtgwg-cl-use-cases]. This document specifies a framework to meet these requirements.

This document describes an Advanced Multipath framework in the context of MPLS networks using an IGP-TE and RSVP-TE MPLS control plane with GMPLS extensions [RFC3209] [RFC3630] [RFC3945] [RFC5305].

Specific protocol solutions are outside the scope of this document, however a framework for the extension of existing protocols is provided. Backwards compatibility is best achieved by extending existing protocols where practical rather than inventing new protocols. The focus is on examining where existing protocol mechanisms fall short with respect to [I-D.ietf-rtgwg-cl-requirement] and on the types of extensions that will be required to accommodate functionality that is called for in [I-D.ietf-rtgwg-cl-requirement].

### 1.1. Background

Classic multipath, including Ethernet Link Aggregation has been widely used in today's MPLS networks [RFC4385][RFC4928]. Classic multipath using non-Ethernet links are often advertised using MPLS Link bundling. A link bundle [RFC4201] bundles a group of homogeneous links as a TE link to make IGP-TE information exchange and RSVP-TE signaling more scalable. An Advanced Multipath allows bundling non-homogenous links together as a single logical link.

An Advanced Multipath is a single logical link in MPLS network that contains multiple parallel component links between two MPLS LSR. Unlike a link bundle [RFC4201], the component links in an Advanced Multipath can have different properties such as cost, capacity, delay, or jitter.

### 1.2. Architecture Summary

Networks aggregate information, both in the control plane and in the data plane, as a means to achieve scalability. A tradeoff exists between the needs of scalability and the needs to identify differing path and link characteristics and differing requirements among flows contained within further aggregated traffic flows. These tradeoffs are discussed in detail in Section 3.

Some aspects of Advanced Multipath requirements present challenges for which multiple solutions may exist. In Section 4 various challenges and potential approaches are discussed.

A subset of the functionality called for in [I-D.ietf-rtgwg-cl-requirement] is available through MPLS Link Bundling [RFC4201]. Link bundling and other existing standards applicable to Advanced Multipath are covered in Section 5.

The most straightforward means of supporting Advanced Multipath requirements is to extend MPLS protocols and protocol semantics and in particular to extend link bundling. Extensions which have already been proposed in other documents which are applicable to Advanced Multipath are discussed in Section 6.

A goal of most new protocol work within IETF is to reuse existing protocol encapsulations and mechanisms where they meet requirements and extend existing mechanisms. This approach minimizes additional complexity while meeting requirements and tends to preserve backwards compatibility to the extent it is practical to do so. These goals are considered in proposing a framework for further protocol extensions and mechanisms in Section 7.

### 1.3. Conventions used in this document

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

### 1.4. Terminology

Terminology defined in [I-D.ietf-rtgwg-cl-requirement] is used in this document. The additional terms defined in [I-D.ietf-rtgwg-cl-use-cases] are also used.

The abbreviation IGP-TE is used as a shorthand indicating either OSPF-TE [RFC3630] or ISIS-TE [RFC5305].

### 1.5. Document Issues

This subsection exists solely for the purpose of focusing the RTGWG meeting and mailing list discussions on areas within this document that need attention in order for the document to achieve the level of quality necessary to advance the document through the IETF process. This subsection will be removed before work group last call.

The following issues need to be resolved.

1. The feasibility of symmetric paths for all flows is questionable. The only case where this is practical is where LSP are smaller than component links and where classic link bundling (not using the all-ones component) is used. Perhaps the emphasis on this

(mis)feature should be reduced in the requirements document. See Section 4.1.3.

2. There is a tradeoff between supporting delay optimized routing and avoiding oscillation. This may be sufficiently covered, but a careful review by others and comments would be beneficial.
3. Any measurement of jitter (delay variation) that is used in route decision is likely to cause oscillation. Trying to optimize a path to reduce jitter may be a fools errand. How do we say this in the draft or does the existing text cover it adequately?
4. RTGWG needs to consider the possibility of using multi-topology IGP extensions in IP and LDP routing where the topologies reflect differing requirements (see Section 4.2.5). This idea is similar to TOS routing, which has been discussed for decades but has never been deployed. One possible outcome of discussion would be to declare TOS routing out of scope in the requirements document.
5. The following referenced drafts have expired:
  - A. [I-D.ospf-cc-stlv]
  - B. [I-D.villamizar-mppls-multipath-extn]

A replacement for [I-D.ospf-cc-stlv] is expected to be submitted. [I-D.villamizar-mppls-multipath-extn] is expected to emerge in a simplified form, removing extensions for which existing workarounds are considered adequate based on feedback at a prior IETF.
6. Clarification of what we intend to do with Multi-Domain Advanced Multipath is needed in Section 7.2.14.
7. The following topics in the requirements document are not addressed. Since they are explicitly mentioned in the requirements document some mention of how they are supported is needed in this document.
  - A. Migration (incremental deployment) may not be adequately covered in Section 4.1.5. It might also be necessary to say more here on performance, scalability, and stability as it related to migration. Comments on this from co-authors or the WG?
  - B. We may need a performance section in this document to specifically address #DR6 (fast convergence), and #DR7 (fast worst case failure convergence). We do already have

scalability discussion and make a recommendation for a separate document. At the very least the performance section would have to say "no worse than before, except were there was no alternative to make it very slightly worse" (in a bit more detail than that). It might also be helpful to better define the nature of the performance criteria implied by #DR6 and #DR7.

The above list has been in this document for the better part of a year with very little discussion (or none) of the above issues on the RTGWG mailing list.

## 2. Advanced Multipath Key Characteristics

[I-D.ietf-rtgwg-cl-requirement] defines external behavior of Advanced Multipath. The overall framework approach involves extending existing protocols in a backwards compatible manner and reusing ongoing work elsewhere in IETF where applicable, defining new protocols or semantics only where necessary. Given the requirements, and this approach of extending MPLS, Advanced Multipath key characteristics can be described in greater detail than given requirements alone.

### 2.1. Flow Identification

Traffic mapping to component links is a data plane operation. Control over how the mapping is done may be directly dictated or constrained by the control plane or by the management plane. When unconstrained by the control plane or management plane, distribution of traffic is entirely a local matter. Regardless of constraints or lack of constraints, the traffic distribution is required to keep packets belonging to individual flows in sequence and meet QoS criteria specified per LSP by either signaling or management [RFC2475] [RFC3260].

Key objectives of the traffic distribution are to not overload any component link, and to be able to perform local recovery when a subset of component links fails.

The network operator may have other objectives such as placing a bidirectional flow or LSP on the same component link in both direction, bounding delay and/or jitter, Advanced Multipath energy saving, and etc. These new requirements are described in [I-D.ietf-rtgwg-cl-requirement].

Examples of means to identify a flow may in principle include:

1. an LSP identified by an MPLS label,
2. a pseudowire (PW) [RFC3985] identified by an MPLS PW label,
3. a flow or group of flows within a pseudowire (PW) [RFC6391] identified by an MPLS flow label,
4. a flow or flow group in an LSP [RFC6790] identified by an MPLS entropy label,
5. all traffic between a pair of IP hosts, identified by an IP source and destination pair,
6. a specific connection between a pair of IP hosts, identified by an IP source and destination pair, protocol, and protocol port pair,
7. a layer-2 conversation within a pseudowire (PW), where the identification is PW payload type specific, such as Ethernet MAC addresses and VLAN tags within an Ethernet PW [RFC4448]. This is feasible but not practical (see below).

Although in principle a layer-2 conversation within a pseudowire (PW), may be identified by PW payload type specific information, in practice this is impractical at LSP midpoints when PW are carried. The PW ingress may provide equivalent information in a PW flow label [RFC6391]. Therefore, in practice, item #8 above is covered by [RFC6391] and may be dropped from the list.

#### 2.1.1.1. Flow Identification Granularity

An LSR must at least be capable of identifying flows based on MPLS labels. Most MPLS LSP do not require that traffic carried by the LSP are carried in order. MPLS-TP is a recent exception. If it is assumed that no LSP require strict packet ordering of the LSP itself (only of flows within the LSP), then the entire label stack can be used as flow identification. If some LSP may require strict packet ordering but those LSP cannot be distinguished from others, then only the top label can be used as a flow identifier. If only the top label is used (for example, as specified by [RFC4201] when the "all-ones" component described in [RFC4201] is not used), then there may not be adequate flow granularity to accomplish well balanced traffic distribution and it will not be possible to carry LSP that are larger than any individual component link.

The number of flows can be extremely large. This may be the case when the entire label stack is used and is always the case when IP addresses are used in provider networks carrying Internet traffic.



Current practice for native IP load balancing at the time of writing were documented in [RFC2991] and [RFC2992]. These practices as described, make use of IP addresses.

The common practices described in [RFC2991] and [RFC2992] were extended to include the MPLS label stack and the common practice of looking at IP addresses within the MPLS payload. These extended practices require that pseudowires use a PWE3 Control Word and are described in [RFC4385] and [RFC4928]. Additional detail on current multipath practices can be found in the appendices of [I-D.ietf-rtgwg-cl-use-cases].

Using only the top label supports too coarse a traffic balance. Prior to MPLS Entropy Label [RFC6790] using the full label stack was also too coarse. Using the full label stack and IP addresses as flow identification provides a sufficiently fine traffic balance, but is capable of identifying such a high number of distinct flows, that a technique of grouping flows, such as hashing on the flow identification criteria, becomes essential to reduce the stored state, and is an essential scaling technique. Other means of grouping flows may be possible.

#### 2.1.2. Flow Identification Summary

In summary:

1. Load balancing using only the MPLS label stack provides too coarse a granularity of load balance.
2. Tracking every flow is not scalable due to the extremely large number of flows in provider networks.
3. Existing techniques, IP source and destination hash in particular, have proven in over two decades of experience to be an excellent way of identifying groups of flows.
4. If a better way to identify groups of flows is discovered, then that method can be used.
5. IP address hashing is not required, but use of this technique is strongly encouraged given the technique's long history of successful deployment.

#### 2.1.3. Flow Identification Using Entropy Label

MPLS Entropy Label [RFC6790] provides a means of making use of the entropy from information that would require deeper packet inspection, such as inspection of IP addresses, and putting that entropy in the

form of a hashed value into the label stack. Midpoint LSR that understand the Entropy Label Indicator can make use of only label stack information but still obtain a fine load balance granularity.

## 2.2. Advanced Multipath in Control Plane

An Advanced Multipath is advertised as a single logical interface between two connected routers, which forms forwarding adjacency (FA) between the routers. The FA is advertised as a TE-link in a link state IGP, using either OSPF-TE or ISIS-TE. The IGP-TE advertised interface parameters for the Advanced Multipath can be preconfigured by the network operator or be derived from its component links. Advanced Multipath advertisement requirements are specified in [I-D.ietf-rtgwg-cl-requirement].

In IGP-TE, an Advanced Multipath is advertised as a single TE link between two connected routers. This is similar to a link bundle [RFC4201]. Link bundle applies to a set of homogenous component links. Advanced Multipath allows homogenous and non-homogenous component links. Due to the similarity, and for backwards compatibility, extending link bundling is viewed as both simple and as the best approach.

In order for a route computation engine to calculate a proper path for a LSP, it is necessary for Advanced Multipath to advertise the summarized available bandwidth as well as the maximum bandwidth that can be made available for single flow (or single LSP where no finer flow identification is available). If an Advanced Multipath contains some non-homogeneous component links, the Advanced Multipath also should advertise the summarized bandwidth and the maximum bandwidth for single flow per each homogeneous component link group.

Both LDP [RFC5036] and RSVP-TE [RFC3209] can be used to signal a LSP over an Advanced Multipath. LDP cannot be extended to support traffic engineering capabilities [RFC3468].

When an LSP is signaled using RSVP-TE, the LSP MUST be placed on the component link that meets the LSP criteria indicated in the signaling message.

When an LSP is signaled using LDP, the LSP MUST be placed on the component link that meets the LSP criteria, if such a component link is available. LDP does not support traffic engineering capabilities, imposing restrictions on LDP use of Advanced Multipath. See Section 4.2.5 for further details.

If the Advanced Multipath solution is based on extensions to IGP-TE and RSVP-TE, then in order to meet requirements defined in

[I-D.ietf-rtgwg-cl-requirement], the following derived requirements MUST be met.

1. An Advanced Multipath MAY contain non-homogeneous component links. The route computing engine MAY select one group of component links for a LSP. The The route computing engine MUST accommodate service objectives for a given LSP when selecting a group of component links for a LSP.
2. The routing protocol MUST make a grouping of component links available in the TE-LSDB, such that within each group all of the component links have similar characteristics (the component links are homogeneous within a group).
3. The route computation used in RSVP-TE MUST be extended to include only the capacity of groups within an Advanced Multipath which meet LSP criteria.
4. The signaling protocol MUST be able to indicate either the criteria, or which groups may be used.
5. An Advanced Multipath MUST place each LSP on a component link or group which meets or exceeds the LSP criteria.

Advanced Multipath capacity is aggregated capacity. LSP capacity MAY be larger than individual component link capacity. Any aggregated LSP can determine a bounds on the largest microflow that could be carried and this constraint can be handled as follows.

1. If no information is available through signaling, management plane, or configuration, the largest microflow is bound by one of the following:
  - A. the largest single LSP if most traffic is RSVP-TE signaled and further aggregated,
  - B. the largest pseudowire if most traffic is carrying pseudowire payloads that are aggregated within RSVP-TE LSP,
  - C. or the largest interface or component link capacity carrying IP or LDP if a large amount of IP or LDP traffic is contained within the aggregate.

If a very large amount of traffic being aggregated is IP or LDP, then the largest microflow is bound by the largest component link on which IP traffic can arrive. For example, if an LSR is acting as an LER and IP and LDP traffic is arriving on 10 Gb/s edge interfaces, then no microflow larger than 10 Gb/s will be present

on the RSVP-TE LSP that aggregate traffic across the core, even if the core interfaces are 100 Gb/s interfaces.

2. The prior conditions provide a bound on the largest microflow when no signaling extensions indicate a bounds. If an LSP is aggregating smaller LSP for which the largest expected microflow carried by the smaller LSP is signaled, then the largest microflow expected in the containing LSP (the aggregate) is the maximum of the largest expected microflow for any contained LSP. For example, RSVP-TE LSP may be large but aggregate traffic for which the source or sink are all 1 Gb/s or smaller interfaces (such as in mobile applications in which cell sites backhauls are no larger than 1 Gb/s). If this information is carried in the LSP originated at the cell sites, then further aggregates across a core may make use of this information.
3. The IGP must provide the bounds on the largest microflow that an Advanced Multipath can accommodate, which is the maximum capacity on a component link that can be made available by moving other traffic. This information is needed by the ingress LER for path determination.
4. A means to signal an LSP whose capacity is larger than individual component link capacity is needed [I-D.ietf-rtgwg-cl-requirement] and also signal the largest microflow expected to be contained in the LSP. If a bounds on the largest microflow is not signaled there is no means to determine if an LSP which is larger than any component link can be subdivided into flows and therefore should be accepted by admission control.

When a bidirectional LSP request is signaled over an Advanced Multipath, if the request indicates that the LSP must be placed on the same component link, the routers of the Advanced Multipath MUST place the LSP traffic in both directions on a same component link. This is particularly challenging for aggregated capacity which makes use of the label stack for traffic distribution. The two requirements are mutually exclusive for any one LSP. No one LSP may be both larger than any individual component link and require symmetrical paths for every flow. Both requirements can be accommodated by the same Advanced Multipath for different LSP, with any one LSP requiring no more than one of these two features.

Individual component link may fail independently. Upon component link failure, an Advanced Multipath MUST support a minimally disruptive local repair, preempting any LSP which can no longer be supported. Available capacity in other component links MUST be used to carry impacted traffic. The available bandwidth after failure MUST be advertised immediately to avoid looped crankback.

When an Advanced Multipath is not able to transport all flows, it preempts some flows based upon holding priority and informs the control plane of these preempted flows. To minimize impact on traffic, the Advanced Multipath MUST support soft preemption [RFC5712]. The network operator SHOULD enable soft preemption. This action ensures the remaining traffic is transported properly. FR#10 requires that the traffic be restored. FR#12 requires that any change be minimally disruptive. These two requirements are interpreted to include preemption among the types of changes that must be minimally disruptive.

### 2.3. Advanced Multipath in Data Plane

The data plane must identify groups of flows. Flow identification is covered in Section 2.1. Having identified groups of flows the groups must be placed on individual component links. This step following flow group identification is called traffic distribution or traffic placement. The two steps together are known as traffic balancing or load balancing.

Traffic distribution may be determined by or constrained by control plane or management plane. Traffic distribution may be changed due to component link status change, subject to constraints imposed by either the management plane or control plane. The distribution function is local to the routers in which an Advanced Multipath belongs to and its implementation is not specified here.

When performing traffic placement, an Advanced Multipath does not differentiate multicast traffic vs. unicast traffic.

In order to maintain scalability, existing data plane forwarding retains state associated with the top label only. Using UHP (UHP is the absence of the more common PHP), zero or more labels may be POPed and packet and byte counters incremented prior to processing what becomes the top label after the POP operations are completed. Flow group identification may be a parallel step in the forwarding process. Data plane forwarding makes use of the top label to select an Advanced Multipath, or a group of components within an Advanced Multipath or for the case where an LSP is pinned (see [RFC4201]), a specific component link. For those LSP for which the LSP selects only the Advanced Multipath or a group of components within an Advanced Multipath, the load balancing makes use of the set of component links selected based on the top label, and makes use of the flow group identification to select among that group.

The simplest traffic placement techniques uses a modulo operation after computing a hash. This techniques has significant disadvantages. The most common traffic placement techniques uses the

a flow group identification as an index into a table. The table provides an indirection. The number of bits of hash is constrained to keep table size small. While this is not the best technique, it is the most common. Better techniques exist but they are outside the scope of this document and some are considered proprietary.

Requirements to limit frequency of load balancing can be adhered to by keeping track of when a flow group was last moved and imposing a minimum period before that flow group can be moved again. This is straightforward for a table approach. For other approaches it may be less straightforward.

### 3. Architecture Tradeoffs

Scalability and stability are critical considerations in protocol design where protocols may be used in a large network such as today's service provider networks. Advanced Multipath is applicable to networks which are large enough to require that traffic be split over multiple paths. Scalability is a major consideration for networks that reach a capacity large enough to require Advanced Multipath.

Some of the requirements of Advanced Multipath could potentially have a negative impact on scalability. This section is about architectural tradeoffs, many motivated by the need to maintain scalability and stability, a need which is reflected in [I-D.ietf-rtgwg-cl-requirement], specifically in DR#6 and DR#7.

#### 3.1. Scalability Motivations

In the interest of scalability, information is aggregated in situations where information about a large amount of network capacity or a large amount of network demand provides is adequate to meet requirements. Routing information is aggregated to reduce the amount of information exchange related to routing and to simplify route computation (see Section 3.2).

In an MPLS network large routing changes can occur when a single fault occurs. For example, a single fault may impact a very large number of LSP traversing a given link. As new LSP are signaled to avoid the fault, resources are consumed elsewhere, and routing protocol announcements must flood the resource changes. If protection is in place, there is less urgency to converging quickly. If multiple faults occur that are not covered by shared risk groups (SRG), then some protection may fail, adding urgency to converging quickly even where protection is deployed.

Reducing the amount of information allows the exchange of information

during a large routing change to be accomplished more quickly and simplifies route computation. Simplifying route computation improves convergence time after very significant network faults which cannot be handled by preprovisioned or precomputed protection mechanisms. Aggregating smaller LSP into larger LSP is a means to reduce path computation load and reduce RSVP-TE signaling (see Section 3.3).

Neglecting scaling issues can result in performance issues, such as slow convergence. Neglecting scaling in some cases can result in networks which perform so poorly as to become unstable.

### 3.2. Reducing Routing Information and Exchange

Link bundling provides a means of aggregating control plane information. Even where the all-ones component link supported by link bundling is not used, the amount of control information is reduced by the number of component links in a bundle.

Fully deaggregating link bundle information would negate this benefit. If there is a need to deaggregate, such as to distinguish between groups of links within specified ranges of delay, then no more deaggregation than is necessary should be done.

For example, in supporting the requirement for heterogeneous component links, it makes little sense to fully deaggregate link bundles when adding support for groups of component links with common attributes within a link bundle can maintain most of the benefit of aggregation while adequately supporting the requirement to support heterogeneous component links.

Routing information exchange is also reduced by making sensible choices regarding the amount of change to link parameters that require link readvertisement. For example, if delay measurements include queuing delay, then a much more coarse granularity of delay measurement would be called for than if the delay does not include queuing and is dominated by geographic delay (speed of light delay).

### 3.3. Reducing Signaling Load

Aggregating traffic into very large hierarchical LSP in the core very substantially reduces the number of LSP that need to be signaled and the number of path computations any given LSR will be required to perform when a network fault occurs.

In the extreme, applying MPLS to a very large network without hierarchy could exceed the 20 bit label space. For example, in a network with 4,000 nodes, with 2,000 on either side of a cutset, would have 4,000,000 LSP crossing the cutset. Even in a degree four

cutset, an uneven distribution of LSP across the cutset, or the loss of one link would result in a need to exceed the size of the label space. Among provider networks, 4,000 access nodes is not at all large. Hierarchy is an absolute requirement if all access nodes were interconnected in such a network.

In less extreme cases, having each node terminate hundreds of LSP to achieve a full mesh creates a very large computational load. Computational complexity is a function of the number of nodes ( $N$ ) and links ( $L$ ) in a topology, and the number of LSP that need to be set up. In the common case where  $L$  is proportional to  $N$  (relatively constant node degree with growth), the time complexity of one CSPF computation is  $\text{order}(N \log N)$ . If each node must perform  $\text{order}(N)$  computations when a fault occurs, then the computational load increases as  $\text{order}(N^2 \log N)$  as the number of nodes increases (where  $^$  is the power of operator and  $N^2$  is read "N-squared"). In practice at the time of writing, this imposes a limit of a few hundred nodes in a full mesh of MPLS LSP before the computational load is sufficient to result in unacceptable convergence times.

Two solutions are applied to reduce the amount of RSVP-TE signaling. Both involve subdividing the MPLS domain into a core and a set of regions.

#### 3.3.1. Reducing Signaling Load using LDP MPTP

LDP can be used for edge-to-edge LSP, using RSVP-TE to carry the LDP intra-core traffic and also optionally also using RSVP-TE to carry the LDP intra-region traffic within each region. LDP does not support traffic engineering, but does support multipoint-to-point (MPTP) LSP, which require less signaling than edge-to-edge RSVP-TE point-to-point (PTP) LSP. A drawback of this approach is the inability to use RSVP-TE protection (FRR or GMPLS protection) against failure of the border LSR sitting at a core/region boundary.

#### 3.3.2. Reducing Signaling Load using Hierarchy

When the number of nodes grows too large, the amount of RSVP-TE signaling can be reduced using the MPLS PSC hierarchy [RFC4206]. A core within the hierarchy can divide the topology into  $M$  regions of on average  $N/M$  nodes. Within a region the computational load is reduced by more than  $M^2$ . Within the core, the computational load generally becomes quite small since  $M$  is usually a fairly small number (a few tens of regions) and each region is generally attached to the core in typically only two or three places on average.

Using hierarchy improves scaling but has two consequences. First, hierarchy effectively forces the use of platform label space. When a



containing LSP is rerouted, the labels assigned to the contained LSP cannot be changed but may arrive on a different interface. Second, hierarchy results in much larger LSP. These LSP today are larger than any single component link and therefore force the use of the all-ones component in link bundles.

### 3.3.3. Using Both LDP MPTP and RSVP-TE Hierarchy

It is also possible to use both LDP and RSVP-TE hierarchy. MPLS networks with a very large number of nodes may benefit from the use of both LDP and RSVP-TE hierarchy. The two techniques are certainly not mutually exclusive.

### 3.4. Reducing Forwarding State

Both LDP and MPLS hierarchy have the benefit of reducing the amount of forwarding state. Using the example from Section 3.3, and using MPLS hierarchy, the worst case generally occurs at borders with the core.

For example, consider a network with approximately 1,000 nodes divided into 10 regions. At the edges, each node requires 1,000 LSP to other edge nodes. The edge nodes also require 100 intra-region LSP. Within the core, if the core has only 3 attachments to each region the core LSR have less than 100 intra-core LSP. At the border cutset between the core and a given region, in this example there are 100 edge nodes with inter-region LSP crossing that cutset, destined to 900 other edge nodes. That yields forwarding state for on the order of 90,000 LSP at the border cutset. These same routers need only reroute well under 200 LSP when a multiple fault occurs, as long as only links are affected and a border LSR does not go down.

Interior to the core, the forwarding state is greatly reduced. If inter-region LSP have different characteristics, it makes sense to make use of aggregates with different characteristics. Rather than exchange information about every inter-region LSP within the intra-core LSP it makes more sense to use multiple intra-core LSP between pairs of core nodes, each aggregating sets of inter-region LSP with common characteristics or common requirements.

### 3.5. Avoiding Route Oscillation

Networks can become unstable when a feedback loop exists such that moving traffic to a link causes a metric such as delay to increase, which then causes traffic to move elsewhere. For example, the original ARPANET routing used a delay based cost metric and proved prone to route oscillations [DBP].

Delay may be used as a constraint in routing for high priority traffic, when this high priority traffic makes a minor contribution to total load, such that the movement of the high priority traffic has a small impact on the delay experienced by other high priority traffic. The safest way to measure delay is to make measurements based on traffic which is prioritized such that it is queued ahead of the lower priority traffic which will be affected if high priority traffic is moved. The amount of high priority traffic must be constrained to consume a fraction of link capacities with the remaining capacity available to lower priority traffic.

Any measurement of jitter (delay variation) that is used in route decision is likely to cause oscillation. Jitter that is caused by queuing effects and cannot be measured using a very high priority measurement traffic flow.

It may be possible to find links with constrained queuing delay or jitter using a theoretical maximum or a probability based bound on queuing delay or jitter at a given priority based on the types and amounts of traffic accepted and combining that theoretical limit with a measured delay at very high priority. Using delay or jitter as path metrics without creating oscillations is challenging.

Instability can occur due to poor performance and interaction with protocol timers. In this way a computational scaling problem can become a stability problem when a network becomes sufficiently large.

#### 4. New Challenges

New technical challenges are posed by [I-D.ietf-rtgwg-cl-requirement] in both the control plane and data plane.

Among the more difficult challenges are the following.

1. The requirements related to delay or jitter conflict with requirements for scalability and stability (see Section 4.1.1),
2. The combination of ingress control over LSP placement and retaining an ability to move traffic as demands dictate can pose challenges and such requirements can even be conflicting (see Section 4.1.2),
3. Path symmetry requires extensions and is particularly challenging for very large LSP (see Section 4.1.3),
4. Accommodating a very wide range of requirements among contained LSP can lead to inefficiency if the most stringent requirements

are reflected in aggregates, or reduce scalability if a large number of aggregates are used to provide a too fine a reflection of the requirements in the contained LSP (see Section 4.1.4),

5. Backwards compatibility is somewhat limited due to the need to accommodate legacy multipath interfaces which provide too little information regarding their configured default behavior, and legacy LSP which provide too little information regarding their LSP requirements (see Section 4.1.5),
6. Data plane challenges include those of accommodating very large LSP, large microflows, traffic ordering constraints imposed by a subset of LSP, and accounting for IP and LDP traffic (see Section 4.2).

#### 4.1. Control Plane Challenges

Some of the control plane requirements are particularly challenging. Handling large flows which aggregate smaller flows must be accomplished with minimal impact on scalability. Potentially conflicting are requirements for jitter and requirements for stability. Potentially conflicting are the requirements for ingress control of a large number of parameters, and the requirements for local control needed to achieve traffic balance across an Advanced Multipath. These challenges and potential solutions are discussed in the following sections.

##### 4.1.1. Delay and Jitter Sensitive Routing

Delay and jitter sensitive routing are called for in [I-D.ietf-rtgwg-cl-requirement] in requirements FR#2, FR#7, FR#8, FR#9, FR#15, FR#16, FR#17, FR#18. Requirement FR#17 is particularly problematic, calling for constraints on jitter.

A tradeoff exists between scaling benefits of aggregating information, and potential benefits of using a finer granularity in delay reporting. To maintain the scaling benefit, measured link delay for any given Advanced Multipath SHOULD be aggregated into a small number of delay ranges. IGP-TE extensions MUST be provided which advertise the available capacities for each of the selected ranges.

For path selection of delay sensitive LSP, the ingress SHOULD bias link metrics based on available capacity and select a low cost path which meets LSP total path delay criteria. To communicate the requirements of an LSP, the ERO MUST be extended to indicate the per link constraints. To communicate the type of resource used, the RRO SHOULD be extended to carry an identification of the group that is

used to carry the LSP at each link bundle hop.

#### 4.1.2. Local Control of Traffic Distribution

Many requirements in [I-D.ietf-rtgwg-cl-requirement] suggest that a node immediately adjacent to a component link should have a high degree of control over how traffic is distributed, as long as network performance objectives are met. Particularly relevant are FR#18 and FR#19.

The requirements to allow local control are potentially in conflict with requirement FR#21 which gives full control of component link select to the LSP ingress. While supporting this capability is mandatory, use of this feature is optional per LSP.

A given network deployment will have to consider this set of conflicting requirements and make appropriate use of local control of traffic placement and ingress control of traffic placement to best meet network requirements.

#### 4.1.3. Path Symmetry Requirements

Requirement FR#21 in [I-D.ietf-rtgwg-cl-requirement] includes a provision to bind both directions of a bidirectional LSP to the same component. This is easily achieved if the LSP is directly signaled across an Advanced Multipath. This is not as easily achieved if a set of LSP with this requirement are signaled over a large hierarchical LSP which is in turn carried over an Advanced Multipath. The basis for load distribution in such a case is the label stack. The labels in either direction are completely independent.

This could be accommodated if the ingress, egress, and all midpoints of the hierarchical LSP make use of an entropy label in the distribution, and the ingress use a fixed value per contained LSP in the entropy label. A solution for this problem may add complexity with very little benefit. There is little or no true benefit of using symmetrical paths rather than component links of identical characteristics.

Traffic symmetry and large LSP capacity are a second pair of conflicting requirements. Any given LSP can meet one of these two requirements but not both. A given network deployment will have to make appropriate use of each of these features to best meet network requirements.

#### 4.1.4. Requirements for Contained LSP

[I-D.ietf-rtgwg-cl-requirement] calls for new LSP constraints. These constraints include frequency of load balancing rearrangement, delay and jitter, packet ordering constraints, and path symmetry.

When LSP are contained within hierarchical LSP, there is no signaling available at midpoint LSR which identifies the contained LSP let alone providing the set of requirements unique to each contained LSP. Defining extensions to provide this information would severely impact scalability and defeat the purpose of aggregating control information and forwarding information into hierarchical LSP. For the same scalability reasons, not aggregating at all is not a viable option for large networks where scalability and stability problems may occur as a result.

As pointed out in Section 4.1.3, the benefits of supporting symmetric paths among LSP contained within hierarchical LSP may not be sufficient to justify the complexity of supporting this capability.

A scalable solution which accommodates multiple sets of LSP between given pairs of LSR is to provide multiple hierarchical LSP for each given pair of LSR, each hierarchical LSP aggregating LSP with common requirements and a common pair of endpoints. This is a network design technique available to the network operator rather than a protocol extension. This technique can accommodate multiple sets of delay and jitter parameters, multiple sets of frequency of load balancing parameters, multiple sets of packet ordering constraints, etc.

#### 4.1.5. Retaining Backwards Compatibility

Backwards compatibility and support for incremental deployment requires considering the impact of legacy LSR in the role of LSP ingress, and considering the impact of legacy LSR advertising ordinary links, advertising Ethernet LAG as ordinary links, and advertising link bundles.

Legacy LSR in the role of LSP ingress cannot signal requirements which are not supported by their control plane software. The additional capabilities supported by other LSR has no impact on these LSR. These LSR however, being unaware of extensions, may try to make use of scarce resources which support specific requirements such as low delay. To a limited extent it may be possible for a network operator to avoid this issue using existing mechanisms such as link administrative attributes and attribute affinities [RFC3209].

Legacy LSR advertising ordinary links will not advertise attributes

needed by some LSP. For example, there is no way to determine the delay or jitter characteristics of such a link. Legacy LSR advertising Ethernet LAG pose additional problems. There is no way to determine that packet ordering constraints would be violated for LSP with strict packet ordering constraints, or that frequency of load balancing rearrangement constraints might be violated.

Legacy LSR advertising link bundles have no way to advertise the configured default behavior of the link bundle. Some link bundles may be configured to place each LSP on a single component link and therefore may not be able to accommodate an LSP which requires bandwidth in excess of the size of a component link. Some link bundles may be configured to spread all LSP over the all-ones component. For LSR using the all-ones component link, there is no documented procedure for correctly setting the "Maximum LSP Bandwidth". There is currently no way to indicate the largest microflow that could be supported by a link bundle using the all-ones component link.

Having received the RRO, it is possible for an ingress to look for the all-ones component to identify such link bundles after having signaled at least one LSP. Whether any LSR collects this information on legacy LSR and makes use of it to set defaults, is an implementation choice.

#### 4.2. Data Plane Challenges

Flow identification is briefly discussed in Section 2.1. Traffic distribution is briefly discussed in Section 2.3. This section discusses issues specific to particular requirements specified in [I-D.ietf-rtgwg-cl-requirement].

##### 4.2.1. Very Large LSP

Very large LSP may exceed the capacity of any single component of an Advanced Multipath. In some cases contained LSP may exceed the capacity of any single component. These LSP may make use of the equivalent of the all-ones component of a link bundle, or may use a subset of components which meet the LSP requirements.

Very large LSP can be accommodated as long as they can be subdivided (see Section 4.2.2). A very large LSP cannot have a requirement for symmetric paths unless complex protocol extensions are proposed (see Section 2.2 and Section 4.1.3).

#### 4.2.2. Very Large Microflows

Within a very large LSP there may be very large microflows. A very large microflow is one which cannot be further subdivided and contributes a very large amount of capacity. Flows which cannot be subdivided must be no larger than the capacity of any single component link.

Current signaling provides no way to specify the largest microflow that can be supported on a given link bundle in routing advertisements. Extensions which address this are discussed in Section 6.4. Absent extensions of this type, traffic containing microflows that are too large for a given Advanced Multipath may be present. There is no data plane solution for this problem that would not require reordering traffic at the Advanced Multipath egress.

Some techniques are susceptible to statistical collisions where an algorithm to distribute traffic is unable to disambiguate traffic among two or more very large microflow where their sum is in excess of the capacity of any single component. Hash based algorithms which use too small a hash space are particularly susceptible and require a change in hash seed in the event that this were to occur. A change in hash seed is highly disruptive, causing traffic reordering among all traffic flows over which the hash function is applied.

#### 4.2.3. Traffic Ordering Constraints

Some LSP have strict traffic ordering constraints. Most notable among these are MPLS-TP LSP. In the absence of aggregation into hierarchical LSP, those LSP with strict traffic ordering constraints can be placed on individual component links if there is a means of identifying which LSP have such a constraint. If LSP with strict traffic ordering constraints are aggregated in hierarchical LSP, the hierarchical LSP capacity may exceed the capacity of any single component link. In such a case the load balancing may be constrained through the use of an entropy label [RFC6790]. This and related issues are discussed further in Section 6.4.

#### 4.2.4. Accounting for IP and LDP Traffic

Networks which carry RSVP-TE signaled MPLS traffic generally carry low volumes of native IP traffic, often only carrying control traffic as native IP. There is no architectural guarantee of this, it is just how network operators have made use of the protocols.

[I-D.ietf-rtgwg-cl-requirement] requires that native IP and native LDP be accommodated (DR#2 and DR#3). In some networks, a subset of services may be carried as native IP or carried as native LDP. Today

this may be accommodated by the network operator estimating the contribution of IP and LDP and configuring a lower set of available bandwidth figures on the RSVP-TE advertisements.

The only improvement that Advanced Multipath can offer is that of measuring the IP and LDP traffic levels and automatically reducing the available bandwidth figures on the RSVP-TE advertisements. The measurements would have to be filtered. This is similar to a feature in existing LSR, commonly known as "autobandwidth" with a key difference. In the "autobandwidth" feature, the bandwidth request of an RSVP-TE signaled LSP is adjusted in response to traffic measurements. In this case the IP or LDP traffic measurements are used to reduce the link bandwidth directly, without first encapsulating in an RSVP-TE LSP.

This may be a subtle and perhaps even a meaningless distinction if Advanced Multipath is used to form a Sub-Path Maintenance Element (SPME). A SPME is in practice essentially an unsignaled single hop LSP with PHP enabled [RFC5921]. An Advanced Multipath SPME looks very much like classic multipath, where there is no signaling, only management plane configuration creating the multipath entity (of which Ethernet Link Aggregation is a subset).

#### 4.2.5. IP and LDP Limitations

IP does not offer traffic engineering. LDP cannot be extended to offer traffic engineering [RFC3468]. Therefore there is no traffic engineered fallback to an alternate path for IP and LDP traffic if resources are not adequate for the IP and/or LDP traffic alone on a given link in the primary path. The only option for IP and LDP would be to declare the link down. Declaring a link down due to resource exhaustion would reduce traffic to zero and eliminate the resource exhaustion. This would cause oscillations and is therefore not a viable solution.

Congestion caused by IP or LDP traffic loads is a pathologic case that can occur if IP and/or LDP are carried natively and there is a high volume of IP or LDP traffic. This situation can be avoided by carrying IP and LDP within RSVP-TE LSP.

It is also not possible to route LDP traffic differently for different FEC. LDP traffic engineering is specifically disallowed by [RFC3468]. It may be possible to support multi-topology IGP extensions to accommodate more than one set of criteria. If so, the additional IGP could be bound to the forwarding criteria, and the LDP FEC bound to a specific IGP instance, inheriting the forwarding criteria. Alternately, one IGP instance can be used and the LDP SPF can make use of the constraints, such as delay and jitter, for a



given LDP FEC.

## 5. Existing Mechanisms

In MPLS the one mechanism which supports explicit signaling of multiple parallel links is Link Bundling [RFC4201]. The set of techniques known as "classis multipath" support no explicit signaling, except in two cases. In Ethernet Link Aggregation the Link Aggregation Control Protocol (LACP) coordinates the addition or removal of members from an Ethernet Link Aggregation Group (LAG). The use of the "all-ones" component of a link bundle indicates use of classis multipath, however the ability to determine if a link bundle makes use of classis multipath is not yet supported.

### 5.1. Link Bundling

Link bundling supports advertisement of a set of homogenous links as a single route advertisement. Link bundling supports placement of an LSP on any single component link, or supports placement of an LSP on the all-ones component link. Not all link bundling implementations support the all-ones component link. There is no way for an ingress LSR to tell which potential midpoint LSR support this feature and use it by default and which do not. Based on [RFC4201] it is unclear how to advertise a link bundle for which the all-ones component link is available and used by default. Common practice is to violate the specification and set the Maximum LSP Bandwidth to the Available Bandwidth. There is no means to determine the largest microflow that could be supported by a link bundle that is using the all-ones component link.

[RFC6107] extends the procedures for hierarchical LSP but also extends link bundles. An LSP can be explicitly signaled to indicate that it is an LSP to be used as a component of a link bundle. Prior to that the common practice was to simply not advertise the component link LSP into the IGP, since only the ingress and egress of the link bundle needed to be aware of their existence, which they would be aware of due to the RSVP-TE signaling used in setting up the component LSP.

While link bundling can be the basis for Advanced Multipath, a significant number of small extension needs to be added.

1. To support link bundles of heterogeneous links, a means of advertising the capacity available within a group of homogeneous links needs to be provided.

2. Attributes need to be defined to support the following parameters for the link bundle or for a group of homogeneous links.
  - A. delay range
  - B. jitter (delay variation) range
  - C. group metric
  - D. all-ones component capable
  - E. capable of dynamically balancing load
  - F. largest supportable microflow
  - G. support for entropy label
3. For each of the prior extended attributes, the constraint based routing path selection needs to be extended to reflect new constraints based on the extended attributes.
4. For each of the prior extended attributes, LSP admission control needs to be extended to reflect new constraints based on the extended attributes.
5. Dynamic load balance must be provided for flows within a given set of links with common attributes such that Performance Objectives are not violated including frequency of load balance adjustment for any given flow.

## 5.2. Classic Multipath

Classic multipath is described in [I-D.ietf-rtgwg-cl-use-cases].

Classic multipath refers to the most common current practice in implementation and deployment of multipath. The most common current practice makes use of a hash on the MPLS label stack and if IPv4 or IPv6 are indicated under the label stack, makes use of the IP source and destination addresses [RFC4385] [RFC4928].

Classic multipath provides a highly scalable means of load balancing. Dynamic multipath has proven value in assuring an even loading on component link and an ability to adapt to change in offered load that occurs over periods of hundreds of milliseconds or more. Classic multipath scalability is due to the ability to effectively work with an extremely large number of flows (IP host pairs) using relatively little resources (a data structure accessed using a hash result as a key or using ranges of hash results).

Classic multipath meets a small subset of Advanced Multipath requirements. Due to scalability of the approach, classic multipath seems to be an excellent candidate for extension to meet the full set of Advanced Multipath forwarding requirements.

Additional detail can be found in [I-D.ietf-rtgwg-cl-use-cases].

## 6. Mechanisms Proposed in Other Documents

A number of documents which at the time of writing are works in progress address parts of the requirements of Advanced Multipath, or assist in making some of the goals achievable.

### 6.1. Loss and Delay Measurement

Procedures for measuring loss and delay are provided in [RFC6374]. These are OAM based measurements. This work could be the basis of delay measurements and delay variation measurement used for metrics called for in [I-D.ietf-rtgwg-cl-requirement].

Currently there are three documents that address delay and delay variation metrics.

draft-ietf-ospf-te-metric-extensions

[I-D.ietf-ospf-te-metric-extensions] provides a set of OSPF-TE extension to support delay, jitter, and loss. Stability is not adequately addressed and some minor issues remain.

I-D.previdi-isis-te-metric-extensions

[I-D.previdi-isis-te-metric-extensions] provides the set of extensions for ISIS that [I-D.ietf-ospf-te-metric-extensions] provides for OSPF. This draft mirrors [I-D.ietf-ospf-te-metric-extensions] sometimes lagging for a brief period when the OSPF version is updated.

I-D.atlas-mppls-te-express-path

[I-D.atlas-mppls-te-express-path] provides information on the use of OSPF and ISIS extensions defined in [I-D.ietf-ospf-te-metric-extensions] and [I-D.previdi-isis-te-metric-extensions] and a modified CSPF path selection to meet LSP performance criteria such as minimal delay paths or bounded delay paths.

Delay variance, loss, residual bandwidth, and available bandwidth extensions are particular prone to network instability. The question as to whether queuing delay and delay variation should be considered, and if so for which diffserv Per-Hop Service Class (PSC) is not

adequately addressed in the current versions of these drafts. These drafts are actively being discussed and updated and remaining issues are expected to be resolved.

## 6.2. Link Bundle Extensions

A set of extension are needed to indicate a group of component links in the ERO or RRO, where the group is given an interface identification like the bundle itself. The extensions could also be further extended to support specification of the all-ones component link in the ERO or RRO.

[I-D.ospf-cc-stlv] provides a baseline draft for extending link bundling to advertise components. A new component TLV (C-TLV) is proposed, which must reference an Advanced Multipath Link TLV. [I-D.ospf-cc-stlv] is intended for the OSPF WG and submitted for the "Experimental" track. The 00 version expired in February 2012. A replacement is expected that will be submitted for consideration on the standards track.

## 6.3. Pseudowire Flow and MPLS Entropy Labels

Two documents provide a means to add entropy for the purpose of improving load balance. MPLS encapsulation can bury information that is needed to identify microflows. These two documents allow a pseudowire ingress and LSP ingress respectively to add a label solely for the purpose of providing a finer granularity of microflow groups.

[RFC6391] allows pseudowires which carry a large volume of traffic, where microflows can be identified to be load balanced across multiple members of an Ethernet LAG or an MPLS link bundle. This is accomplished by adding a flow label below the pseudowire label in the MPLS label stack. For this to be effective the link bundle load balance must make use of the label stack up to and including this flow label.

[RFC6790] provides a means for a LER to put an additional label known as an entropy label on the MPLS label stack. Only the LER can add the entropy label. The LER of a PSC LSP would have to add a entropy label for contained LSPs for which it is a midpoint LSR.

Core LSR acting as LER for aggregated LSP can add entropy labels based on deep packet inspection and place an entropy label indicator (ELI) and entropy label (EL) just below the label being acted on. This would be helpful in situations where the label stack depth to which load distribution can operate is limited by implementation or is limited for other reasons such as carrying both MPLS-TP and MPLS with entropy labels within the same hierarchical LSP.

#### 6.4. Multipath Extensions

The multipath extensions drafts address the issue of accommodating LSP which have strict packet ordering constraints in a network containing multipath. MPLS-TP has become the one important instance of LSP with strict packet ordering constraints and has driven this work.

[I-D.ietf-mpls-multipath-use] proposed to use MPLS Entropy Label [RFC6790] to allow MPLS-TP to be carried within MPLS LSP that make use of multipath. Limitations of this approach in the absence of protocol extensions is discussed.

[I-D.villamizar-mpls-multipath-extn] provides protocol extensions needed to overcome the limitations in the absence of protocol extensions is discussed in [I-D.ietf-mpls-multipath-use].

#### 7. Required Protocol Extensions and Mechanisms

Prior sections have reviewed key characteristics, architecture tradeoffs, new challenges, existing mechanisms, and relevant mechanisms proposed in existing new documents.

This section first summarizes and groups requirements specified in [I-D.ietf-rtgwg-cl-requirement] (see Section 7.1). A set of documents coverage groupings are proposed with existing works-in-progress noted where applicable (see Section 7.2). The set of extensions are then grouped by protocol affected as a convenience to implementors (see (see Section 7.3).

##### 7.1. Brief Review of Requirements

The following list provides a categorization of requirements specified in [I-D.ietf-rtgwg-cl-requirement] along with a short phrase indication what topic the requirement covers.

routing information aggregation

FR#1 (routing summarization), FR#20 (Advanced Multipath may be a component of another Advanced Multipath)

restoration speed

FR#2 (restoration speed meeting performance objectives), FR#12 (minimally disruptive load rebalance), DR#6 (fast convergence), DR#7 (fast worst case failure convergence)

load distribution, stability, minimal disruption

FR#3 (automatic load distribution), FR#5 (must not oscillate), FR#11 (dynamic placement of flows), FR#12 (minimally disruptive load rebalance), FR#13 (bounded rearrangement frequency), FR#18 (flow placement must satisfy performance objectives), FR#19 (flow identification finer than per top level LSP), MR#6 (operator initiated flow rebalance)

backward compatibility and migration

FR#4 (smooth incremental deployment), FR#6 (management and diagnostics must continue to function), DR#1 (extend existing protocols), DR#2 (extend LDP, no LDP TE)

delay and delay variation

FR#7 (expose lower layer measured delay), FR#8 (precision of latency reporting), FR#9 (limit latency on per LSP basis), FR#15 (minimum delay path), FR#16 (bounded delay path), FR#17 (bounded jitter path)

admission control, preemption, traffic engineering

FR#10 (admission control, preemption), FR#14 (packet ordering), FR#21 (ingress specification of path), FR#22 (path symmetry), DR#3 (IP and LDP traffic), MR#3 (management specification of path)

single vs multiple domain

DR#4 (IGP extensions allowed within single domain), DR#5 (IGP extensions disallowed in multiple domain case)

general network management

MR#1 (polling, configuration, and notification), MR#2 (activation and de-activation)

path determination, connectivity verification

MR#4 (path trace), MR#5 (connectivity verification)

The above list is not intended as a substitute for

[I-D.ietf-rtgwg-cl-requirement], but rather as a concise grouping and reminder or requirements to serve as a means of more easily determining requirements coverage of a set of protocol documents.

## 7.2. Proposed Document Coverage

The primary areas where additional protocol extensions and mechanisms are required include the topics described in the following subsections.

There are candidate documents for a subset of the topics below. This

grouping of topics does not require that each topic be addressed by a separate document. In some cases, a document may cover multiple topics, or a specific topic may be addressed as applicable in multiple documents.

#### 7.2.1. Component Link Grouping

An extension to link bundling is needed to specify a group of components with common attributes. This can be a TLV defined within the link bundle that carries the same encapsulations as the link bundle. Two interface indices would be needed for each group.

- a. An index is needed that if included in an ERO would indicate the need to place the LSP on any one component within the group.
- b. A second index is needed that if included in an ERO would indicate the need to balance flows within the LSP across all components of the group. This is equivalent to the "all-ones" component for the entire bundle.

[I-D.ospf-cc-stlv] can be extended to include multipath treatment capabilities. An ISIS solution is also needed. An extension of RSVP-TE signaling is needed to indicate multipath treatment preferences.

If a component group is allowed to support all of the parameters of a link bundle, then a group TE metric would be accommodated. This can be supported with the component TLV (C-TLV) defined in [I-D.ospf-cc-stlv].

The primary focus of this document, among the sets of requirements listed in Section 7.1 is the "routing information aggregation" set of requirements. The "restoration speed", "backward compatibility and migration", and "general network management" requirements must also be considered.

#### 7.2.2. Delay and Jitter Extensions

A extension is needed in the IGP-TE advertisement to support delay and delay variation for links, link bundles, and forwarding adjacencies. Whatever mechanism is described must take precautions that insure that route oscillations cannot occur. The following set of drafts address this.

1. [I-D.ietf-ospf-te-metric-extensions]
2. [I-D.previdi-isis-te-metric-extensions]

### 3. [I-D.atlas-mpis-te-express-path]

The primary focus of this document, among the sets of requirements listed in Section 7.1 is the "delay and delay variation" set of requirements. The "restoration speed", "backward compatibility and migration", and "general network management" requirements must also be considered.

#### 7.2.3. Path Selection and Admission Control

Path selection and admission control changes must be documented in each document that proposes a protocol extension that advertises a new capability or parameter that must be supported by changes in path selection and admission control.

It would also be helpful to have an informational document which covers path selection and admission control issues in detail and briefly summarizes and references the set of documents which propose extensions. This document could be advanced in parallel with the protocol extensions.

The primary focus of this document, among the sets of requirements listed in Section 7.1 are the "load distribution, stability, minimal disruption" and "admission control, preemption, traffic engineering" sets of requirements. The "restoration speed" and "path determination, connectivity verification" requirements must also be considered. The "backward compatibility and migration", and "general network management" requirements must also be considered.

#### 7.2.4. Dynamic Multipath Balance

FR#11 explicitly calls for dynamic placement of flows. Load balancing similar to existing dynamic multipath would satisfy this requirement. In implementations where flow identification uses a coarse granularity, the adjustments would have to be equally coarse, in the worst case moving entire LSP. The impact of flow identification granularity and potential dynamic multipath approaches may need to be documented in greater detail than provided here.

The primary focus of this document, among the sets of requirements listed in Section 7.1 are the "restoration speed" and the "load distribution, stability, minimal disruption" sets of requirements. The "path determination, connectivity verification" requirements must also be considered. The "backward compatibility and migration", and "general network management" requirements must also be considered.



#### 7.2.5. Frequency of Load Balance

IGP-TE and RSVP-TE extensions are needed to support frequency of load balancing rearrangement called for in FR#13, and FR#15-FR#17. Constraints are not defined in RSVP-TE, but could be modeled after administrative attribute affinities in RFC3209 and elsewhere.

The primary focus of this document, among the sets of requirements listed in Section 7.1 is the "load distribution, stability, minimal disruption" set of requirements. The "path determination, connectivity verification" must also be considered. The "backward compatibility and migration" and "general network management" requirements must also be considered.

#### 7.2.6. Inter-Layer Communication

Lower layer to upper layer communication called for in FR#7 and FR#20. Specific parameters, specifically delay and delay variation, need to be addressed. Passing information from a lower non-MPLS layer to an MPLS layer needs to be addressed, though this may largely be generic advice encouraging a coupling of MPLS to lower layer management plane or control plane interfaces. This topic can be addressed in each document proposing a protocol extension, where applicable.

The primary focus of this document, among the sets of requirements listed in Section 7.1 is the "restoration speed" set of requirements. The "backward compatibility and migration" and "general network management" requirements must also be considered.

#### 7.2.7. Packet Ordering Requirements

A document is needed to define extensions supporting various packet ordering requirements, ranging from requirements to preserve microflow ordering only, to requirements to preserve full LSP ordering (as in MPLS-TP). This is covered by [I-D.ietf-mpls-multipath-use] and [I-D.villamizar-mpls-multipath-extn].

The primary focus of this document, among the sets of requirements listed in Section 7.1 are the "admission control, preemption, traffic engineering" and the "path determination, connectivity verification" sets of requirements. The "backward compatibility and migration" and "general network management" requirements must also be considered.

#### 7.2.8. Minimally Disruption Load Balance

The behavior of hash methods used in classic multipath needs to be described in terms of FR#12 which calls for minimally disruptive load adjustments. For example, reseeding the hash violates FR#12. Using modulo operations is significantly disruptive if a link comes or goes down, as pointed out in [RFC2992]. In addition, backwards compatibility with older hardware needs to be accommodated.

The primary focus of this document, among the sets of requirements listed in Section 7.1 is the "load distribution, stability, minimal disruption" set of requirements.

#### 7.2.9. Path Symmetry

Protocol extensions are needed to support dynamic load balance as called for to meet FR#22 (path symmetry) and to meet FR#11 (dynamic placement of flows).

Currently path symmetry can only be supported in link bundling if the path is pinned. When a flow is moved both ingress and egress must make the move as close to simultaneously as possible to satisfy FR#22 and FR#12 (minimally disruptive load rebalance). There is currently no protocol to coordinate this move.

If a group of flows are identified using a hash, then the hash must be identical on the pair of LSR at the endpoint, using the same hash seed and with one side swapping source and destination. If the label stack is used, then either the entire label stack must be a special case flow identification, since the set of labels in either direction are not correlated, or the two LSR must conspire to use the same flow identifier. For example, using a common entropy label value, and using only the entropy label in the flow identification would satisfy the forwarding requirement. There is no protocol to indicate special treatment of a label stack within a hierarchical LSP. Adding such an extension may add significant complexity and ultimately may prove unscalable.

The primary focus of this document, among the sets of requirements listed in Section 7.1 are the "load distribution, stability, minimal disruption" and the "admission control, preemption, traffic engineering" sets of requirements. The "backward compatibility and migration" and "general network management" requirements must also be considered. Path symmetry simplifies support for the "path determination, connectivity verification" set of requirements, but with significant complexity added elsewhere.

#### 7.2.10. Performance, Scalability, and Stability

A separate document providing analysis of performance, scalability, and stability impacts of changes may be needed. The topic of traffic adjustment oscillation must also be covered. If sufficient coverage is provided in each document covering a protocol extension, a separate document would not be needed.

The primary focus of this document, among the sets of requirements listed in Section 7.1 is the "restoration speed" set of requirements. This is not a simple topic and not a topic that is well served by scattering it over multiple documents, therefore it may be best to put this in a separate document and put citations in documents called for in Section 7.2.1, Section 7.2.2, Section 7.2.3, Section 7.2.9, Section 7.2.11, Section 7.2.12, Section 7.2.13, and Section 7.2.14. Citation may also be helpful in Section 7.2.4, and Section 7.2.5.

#### 7.2.11. IP and LDP Traffic

A document is needed to define the use of measurements of native IP and native LDP traffic levels which are then used to reduce link advertised bandwidth amounts.

The primary focus of this document, among the sets of requirements listed in Section 7.1 are the "load distribution, stability, minimal disruption" and the "admission control, preemption, traffic engineering" set of requirements. The "path determination, connectivity verification" must also be considered. The "backward compatibility and migration" and "general network management" requirements must also be considered.

#### 7.2.12. LDP Extensions

Extending LDP is called for in DR#2. LDP can be extended to couple FEC admission control to local resource availability without providing LDP traffic engineering capability. Other LDP extensions such as signaling a bound on microflow size and LDP LSP requirements would provide useful information without providing LDP traffic engineering capability.

The primary focus of this document, among the sets of requirements listed in Section 7.1 is the "admission control, preemption, traffic engineering" set of requirements. The "backward compatibility and migration" and "general network management" requirements must also be considered.

#### 7.2.13. Pseudowire Extensions

Pseudowire (PW) extensions such as signaling a bound on microflow size and signaling requirements specific to PW would provide useful information. This information can be carried in the PW LDP signaling [RFC3985] and the the PW requirements could then be used in a containing LSP.

The primary focus of this document, among the sets of requirements listed in Section 7.1 is the "admission control, preemption, traffic engineering" set of requirements. The "backward compatibility and migration" and "general network management" requirements must also be considered.

#### 7.2.14. Multi-Domain Advanced Multipath

DR#5 calls for Advanced Multipath to span multiple network topologies. Component LSP may already span multiple network topologies, though most often in practice these are LDP signaled. Component LSP which are RSVP-TE signaled may also span multiple network topologies using at least three existing methods (per domain [RFC5152], BRPC [RFC5441], PCE [RFC4655]). When such component links are combined in an Advanced Multipath, the Advanced Multipath spans multiple network topologies. It is not clear in which document this needs to be described or whether this description in the framework is sufficient. The authors and/or the WG may need to discuss this. DR#5 mandates that IGP-TE extension cannot be used. This would disallow the use of [RFC5316] or [RFC5392] in conjunction with [RFC5151].

The primary focus of this document, among the sets of requirements listed in Section 7.1 are "single vs multiple domain" and "admission control, preemption, traffic engineering". The "routing information aggregation" and "load distribution, stability, minimal disruption" requirements need attention due to their use of the IGP in single domain Advanced Multipath. Other requirements such as "delay and delay variation", can more easily be accommodated by carrying metrics within BGP. The "path determination, connectivity verification" requirements need attention due to requirements to restrict disclosure of topology information across domains in multi-domain deployments. The "backward compatibility and migration" and "general network management" requirements must also be considered.

#### 7.3. Framework Requirement Coverage by Protocol

As an aid to implementors, this section summarizes requirement coverage listed in Section 7.2 by protocol or LSR functionality affected.

Some documentation may be purely informational, proposing no changes and proposing usage at most. This includes Section 7.2.3, Section 7.2.8, Section 7.2.10, and Section 7.2.14.

Section 7.2.9 may require a new protocol.

#### 7.3.1. OSPF-TE and ISIS-TE Protocol Extensions

Many of the changes listed in Section 7.2 require IGP-TE changes, though most are small extensions to provide additional information. This set includes Section 7.2.1, Section 7.2.2, Section 7.2.5, Section 7.2.6, and Section 7.2.7. An adjustment to existing advertised parameters is suggested in Section 7.2.11.

#### 7.3.2. PW Protocol Extensions

The only suggestion of pseudowire (PW) extensions is in Section 7.2.13.

#### 7.3.3. LDP Protocol Extensions

Potential LDP extensions are described in Section 7.2.12.

#### 7.3.4. RSVP-TE Protocol Extensions

RSVP-TE protocol extensions are called for in Section 7.2.1, Section 7.2.5, Section 7.2.7, and Section 7.2.9.

#### 7.3.5. RSVP-TE Path Selection Changes

Section 7.2.3 calls for path selection to be addressed in individual documents that require change. These changes would include those proposed in Section 7.2.1, Section 7.2.2, Section 7.2.5, and Section 7.2.7.

#### 7.3.6. RSVP-TE Admission Control and Preemption

When a change is needed to path selection, a corresponding change is needed in admission control. The same set of sections applies: Section 7.2.1, Section 7.2.2, Section 7.2.5, and Section 7.2.7. Some resource changes such as a link delay change might trigger preemption. The rules of preemption remain unchanged, still based on holding priority.

#### 7.3.7. Flow Identification and Traffic Balance

The following describe either the state of the art in flow identification and traffic balance or propose changes: Section 7.2.4,

Section 7.2.5, Section 7.2.7, and Section 7.2.8.

## 8. IANA Considerations

This is a framework document and therefore does not specify protocol extensions. This memo includes no request to IANA.

## 9. Security Considerations

The security considerations for MPLS/GMPLS and for MPLS-TP are documented in [RFC5920] and [RFC6941].

The types protocol extensions proposed in this framework document provide additional information about links, forwarding adjacencies, and LSP requirements. The protocol semantics changes described in this framework document propose additional LSP constraints applied at path computation time and at LSP admission at midpoints LSR. The additional information and constraints provide no additional security considerations beyond the security considerations already documented in [RFC5920] and [RFC6941].

## 10. Acknowledgments

Authors would like to thank Adrian Farrel, Fred Jounay, Yuji Kamite for his extensive comments and suggestions regarding early versions of this document, Ron Bonica, Nabil Bitar, Eric Gray, Lou Berger, and Kireeti Kompella for their reviews of early versions and great suggestions.

Authors would like to thank Iftexhar Hussain for review and suggestions regarding recent versions of this document.

In the interest of full disclosure of affiliation and in the interest of acknowledging sponsorship, past affiliations of authors are noted. Much of the work done by Ning So occurred while Ning was at Verizon. Much of the work done by Curtis Villamizar occurred while at Infinera. Infinera continues to sponsor this work on a consulting basis.

## 11. References

## 11.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, December 2001.
- [RFC3630] Katz, D., Kompella, K., and D. Yeung, "Traffic Engineering (TE) Extensions to OSPF Version 2", RFC 3630, September 2003.
- [RFC4201] Kompella, K., Rekhter, Y., and L. Berger, "Link Bundling in MPLS Traffic Engineering (TE)", RFC 4201, October 2005.
- [RFC4206] Kompella, K. and Y. Rekhter, "Label Switched Paths (LSP) Hierarchy with Generalized Multi-Protocol Label Switching (GMPLS) Traffic Engineering (TE)", RFC 4206, October 2005.
- [RFC5036] Andersson, L., Minei, I., and B. Thomas, "LDP Specification", RFC 5036, October 2007.
- [RFC5305] Li, T. and H. Smit, "IS-IS Extensions for Traffic Engineering", RFC 5305, October 2008.
- [RFC5712] Meyer, M. and JP. Vasseur, "MPLS Traffic Engineering Soft Preemption", RFC 5712, January 2010.
- [RFC6107] Shiimoto, K. and A. Farrel, "Procedures for Dynamically Signaled Hierarchical Label Switched Paths", RFC 6107, February 2011.
- [RFC6374] Frost, D. and S. Bryant, "Packet Loss and Delay Measurement for MPLS Networks", RFC 6374, September 2011.
- [RFC6391] Bryant, S., Filsfils, C., Drafz, U., Kompella, V., Regan, J., and S. Amante, "Flow-Aware Transport of Pseudowires over an MPLS Packet Switched Network", RFC 6391, November 2011.

## 11.2. Informative References

- [DBP] Bertsekas, D., "Dynamic Behavior of Shortest Path Routing Algorithms for Communication Networks", IEEE Trans. Auto. Control 1982.
- [I-D.atlas-mpls-te-express-path]

Atlas, A., Drake, J., Giacalone, S., Ward, D., Previdi, S., and C. Filsfils, "Performance-based Path Selection for Explicitly Routed LSPs", draft-atlas-mppls-te-express-path-02 (work in progress), February 2013.

[I-D.ietf-mppls-multipath-use]

Villamizar, C., "Use of Multipath with MPLS-TP and MPLS", draft-ietf-mppls-multipath-use-00 (work in progress), February 2013.

[I-D.ietf-ospf-te-metric-extensions]

Giacalone, S., Ward, D., Drake, J., Atlas, A., and S. Previdi, "OSPF Traffic Engineering (TE) Metric Extensions", draft-ietf-ospf-te-metric-extensions-04 (work in progress), June 2013.

[I-D.ietf-rtgwg-cl-requirement]

Villamizar, C., McDysan, D., Ning, S., Malis, A., and L. Yong, "Requirements for Advanced Multipath in MPLS Networks", draft-ietf-rtgwg-cl-requirement-11 (work in progress), July 2013.

[I-D.ietf-rtgwg-cl-use-cases]

Ning, S., Malis, A., McDysan, D., Yong, L., and C. Villamizar, "Advanced Multipath Use Cases and Design Considerations", draft-ietf-rtgwg-cl-use-cases-04 (work in progress), July 2013.

[I-D.ospf-cc-stlv]

Osborne, E., "Component and Composite Link Membership in OSPF", draft-ospf-cc-stlv-00 (work in progress), August 2011.

[I-D.previdi-isis-te-metric-extensions]

Previdi, S., Giacalone, S., Ward, D., Drake, J., Atlas, A., and C. Filsfils, "IS-IS Traffic Engineering (TE) Metric Extensions", draft-previdi-isis-te-metric-extensions-03 (work in progress), February 2013.

[I-D.villamizar-mppls-multipath-extn]

Villamizar, C., "Multipath Extensions for MPLS Traffic Engineering", draft-villamizar-mppls-multipath-extn-00 (work in progress), November 2012.

[RFC2475] Blake, S., Black, D., Carlson, M., Davies, E., Wang, Z., and W. Weiss, "An Architecture for Differentiated



Services", RFC 2475, December 1998.

- [RFC2991] Thaler, D. and C. Hopps, "Multipath Issues in Unicast and Multicast Next-Hop Selection", RFC 2991, November 2000.
- [RFC2992] Hopps, C., "Analysis of an Equal-Cost Multi-Path Algorithm", RFC 2992, November 2000.
- [RFC3260] Grossman, D., "New Terminology and Clarifications for Diffserv", RFC 3260, April 2002.
- [RFC3468] Andersson, L. and G. Swallow, "The Multiprotocol Label Switching (MPLS) Working Group decision on MPLS signaling protocols", RFC 3468, February 2003.
- [RFC3945] Mannie, E., "Generalized Multi-Protocol Label Switching (GMPLS) Architecture", RFC 3945, October 2004.
- [RFC3985] Bryant, S. and P. Pate, "Pseudo Wire Emulation Edge-to-Edge (PWE3) Architecture", RFC 3985, March 2005.
- [RFC4385] Bryant, S., Swallow, G., Martini, L., and D. McPherson, "Pseudowire Emulation Edge-to-Edge (PWE3) Control Word for Use over an MPLS PSN", RFC 4385, February 2006.
- [RFC4448] Martini, L., Rosen, E., El-Aawar, N., and G. Heron, "Encapsulation Methods for Transport of Ethernet over MPLS Networks", RFC 4448, April 2006.
- [RFC4655] Farrel, A., Vasseur, J., and J. Ash, "A Path Computation Element (PCE)-Based Architecture", RFC 4655, August 2006.
- [RFC4928] Swallow, G., Bryant, S., and L. Andersson, "Avoiding Equal Cost Multipath Treatment in MPLS Networks", BCP 128, RFC 4928, June 2007.
- [RFC5151] Farrel, A., Ayyangar, A., and JP. Vasseur, "Inter-Domain MPLS and GMPLS Traffic Engineering -- Resource Reservation Protocol-Traffic Engineering (RSVP-TE) Extensions", RFC 5151, February 2008.
- [RFC5152] Vasseur, JP., Ayyangar, A., and R. Zhang, "A Per-Domain Path Computation Method for Establishing Inter-Domain Traffic Engineering (TE) Label Switched Paths (LSPs)", RFC 5152, February 2008.
- [RFC5316] Chen, M., Zhang, R., and X. Duan, "ISIS Extensions in Support of Inter-Autonomous System (AS) MPLS and GMPLS

Traffic Engineering", RFC 5316, December 2008.

- [RFC5392] Chen, M., Zhang, R., and X. Duan, "OSPF Extensions in Support of Inter-Autonomous System (AS) MPLS and GMPLS Traffic Engineering", RFC 5392, January 2009.
- [RFC5441] Vasseur, JP., Zhang, R., Bitar, N., and JL. Le Roux, "A Backward-Recursive PCE-Based Computation (BRPC) Procedure to Compute Shortest Constrained Inter-Domain Traffic Engineering Label Switched Paths", RFC 5441, April 2009.
- [RFC5920] Fang, L., "Security Framework for MPLS and GMPLS Networks", RFC 5920, July 2010.
- [RFC5921] Bocci, M., Bryant, S., Frost, D., Levrau, L., and L. Berger, "A Framework for MPLS in Transport Networks", RFC 5921, July 2010.
- [RFC6790] Kompella, K., Drake, J., Amante, S., Henderickx, W., and L. Yong, "The Use of Entropy Labels in MPLS Forwarding", RFC 6790, November 2012.
- [RFC6941] Fang, L., Niven-Jenkins, B., Mansfield, S., and R. Graveman, "MPLS Transport Profile (MPLS-TP) Security Framework", RFC 6941, April 2013.

#### Authors' Addresses

So Ning  
Tata Communications  
  
Email: ning.so@tatacommunications.com

Dave McDysan  
Verizon  
22001 Loudoun County PKWY  
Ashburn, VA 20147  
USA  
  
Email: dave.mcdysan@verizon.com

Eric Osborne  
Cisco

Email: eosborne@cisco.com

Lucy Yong  
Huawei USA  
5340 Legacy Dr.  
Plano, TX 75025  
USA

Phone: +1 469-277-5837  
Email: lucy.yong@huawei.com

Curtis Villamizar  
Outer Cape Cod Network Consulting

Email: curtis@occnc.com



Network Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: August 29, 2013

Z. Li  
N. Wu  
Q. Zhao  
Huawei Technologies  
February 25, 2013

Routing Extension for Fast-Reroute Using Maximally Redundant Trees  
draft-li-rtgwg-igp-ext-mrt-frr-01

Abstract

The document proposes the routing protocol extension and procedures to support fast reroute using maximally redundant trees.

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 RFC 2119 [RFC2119].

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 August 29, 2013.

Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect

to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

## Table of Contents

1. Introduction . . . . .	3
2. Terminology . . . . .	3
3. MRT-FRR TLV Format . . . . .	3
3.1. IS-IS MRT-FRR Sub-TLV Format . . . . .	3
3.2. OSPF MRT FRR TLV Format . . . . .	5
4. Elements of Procedure . . . . .	7
4.1. IS-IS . . . . .	7
4.1.1. Sending . . . . .	7
4.1.2. Receiving . . . . .	8
4.2. OSPF . . . . .	9
4.2.1. Sending . . . . .	9
4.2.2. Receiving . . . . .	10
5. Backward Compatibility . . . . .	10
6. IANA Considerations . . . . .	11
6.1. IS-IS . . . . .	11
6.2. OSPF . . . . .	11
7. Security Considerations . . . . .	11
8. Normative References . . . . .	11
Authors' Addresses . . . . .	12

## 1. Introduction

[I-D.ietf-rtgwg-mrt-frr-architecture] describes the architecture based on Maximally Redundant Trees (MRT) to provide 100% coverage for FRR of unicast traffic. Protocol extensions and considerations are also been proposed. The document defines the extension of routing protocols for fast reroute using maximally redundant trees. The detailed procedures are defined based the extension.

## 2. Terminology

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

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

## 3. MRT-FRR TLV Format

### 3.1. IS-IS MRT-FRR Sub-TLV Format

IS-IS MRT FRR sub-TLV is defined to advertise necessary information related with MRT FRR. It is an optional sub-TLV which can be advertised in the router capability TLV([RFC4971]) . The information has only level-wide scope. If there is no MRT FRR sub-TLV advertised, that router should be seen as that it does not support MRT FRR.

The IS-IS MRT FRR sub-TLV is composed of 1 octet for the type, 1 octet specifying the TLV length and a value field. The IS-IS MRT FRR sub-TLV format (Figure 1) is as follows:

TYPE: TBD

LENGTH: 12

	No. of Octets
+-----+  R R R R  Primary MT ID	2
+-----+  R R R R  Blue MRT MT ID	2
+-----+  R R R R  Red MRT MT ID	2
+-----+   MRT Capabilities Available	2
+-----+  MRT Algorithm ID	1
+-----+  MRT Fd Mechanism	1
+-----+   GADAG Root Election Priority	2
+-----+	

Figure 1 IS-IS MRT FRR Sub-TLV Format

The IS-IS MRT FRR sub-TLV is made of the following fields:

-- Primary MT-ID: This specifies the MT-ID of the primary topology, 12 bits.

-- Blue MRT MT-ID: This specifies the MT-ID to be associated with the Blue MRT forwarding topology. It is needed for use in signaling. All routers in the MRT Island MUST agree on a value, 12 bits.

-- Red MRT MT-ID: This specifies the MT-ID to be associated with the Red MRT forwarding topology. It is needed for use in signaling. All routers in the MRT Island MUST agree on a value, 12 bits.

-- MRT Capabilities Available: This specifies the set of capabilities that the router can support.

```

+---+---+---+---+---+---+---+---+---+---+
|0|1|2|3|4|5|6|*| Reserved |
+---+---+---+---+---+---+---+---+

```

```

Bit0 - MRT-BIT
Bit1 - IP-BIT
Bit2 - LDP-BIT
Bit3 - PIM-BIT
Bit4 - PIMG-BIT
Bit5 - mLDP-BIT
Bit6 - mLDPG-BIT

```

-- MRT Algorithm ID: This identifies the particular MRT algorithm used by the router. By having an Algorithm ID, it is possible to



change the algorithm used or use different ones in different networks. It may be desirable to advertise a list ordered by preference to allow transitions.

```

+---+---+---+---+
|0|1|2|*|*|*|*|
+---+---+---+---+

```

Bit0 - LP-BIT  
 Bit1 - SPF-BIT  
 Bit2 - HYBRID-BIT

-- MRT Fd Mechanism: This specifies which forwarding mechanisms the router supports. If IP-in-IPv4 or IP-in-IPv6 is used as forwarding mechanisms for IP, Red MRT Loopback Address and Blue MRT Loopback Address should be advertised by the Multi-Topology Reachable IPv4/IPv6 Prefixes TLV ([RFC5120]).

```

+---+---+---+---+
|0|*|2|3|4|*|*|*|
+---+---+---+---+

```

Bit0: LDP Destination-Topology Label  
 Bit2: IP-in-IPv4  
 Bit3: IP-in-IPv6  
 Bit4: Encode MT-ID in Labels

-- GADAG Root Election Priority: This specifies the priority of the router for being used as the GADAG root of its island. A GADAG root is elected from the set of routers with the highest priority; ties are broken based upon highest System ID. The sensitivity of the MRT Algorithms to GADAG root selection is still being evaluated. This provides the network operator with a knob to force particular GADAG root selection.

### 3.2. OSPF MRT FRR TLV Format

OSPF MRT FRR TLV is defined to advertise necessary information related with MRT FRR. It is an optional TLV which can be advertised in the OSPF router information LSA([RFC4970]) . The information has only area-wide scope. If there is no MRT FRR TLV advertised, that router should be seen as that it does not support MRT FRR.

The OSPF MRT FRR TLV has the following format:

TYPE: TBD

LENGTH: 12

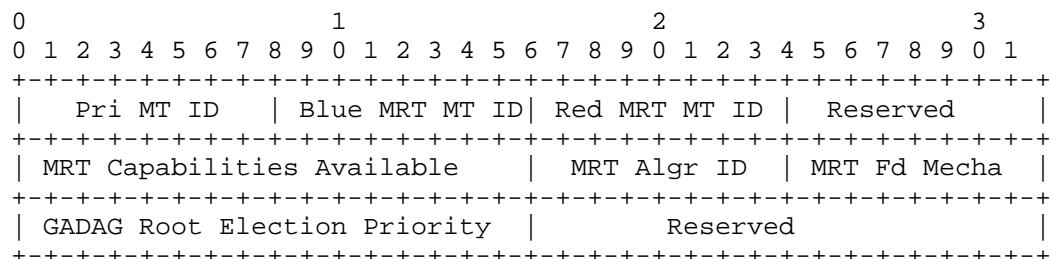


Figure 2 OSPF MRT FRR TLV Format

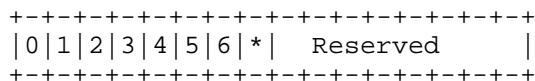
The OSPF MRT FRR TLV is made of the following fields:

-- Pri MT-ID: This specifies the MT-ID of the primary topology. It is a 7-bit-field since AS-External-LSAs use the high-order bit in the MT-ID field (E-bit) for the external metric-type.

-- Blue MRT MT-ID: This specifies the MT-ID to be associated with the Blue MRT forwarding topology. It is needed for use in signaling. All routers in the MRT Island MUST agree on a value, 7 bits.

-- Red MRT MT-ID: This specifies the MT-ID to be associated with the Red MRT forwarding topology. It is needed for use in signaling. All routers in the MRT Island MUST agree on a value, 7 bits.

-- MRT Capabilities Available: This specifies the set of capabilities that the router can support.



Bit0 - MRT-BIT  
 Bit1 - IP-BIT  
 Bit2 - LDP-BIT  
 Bit3 - PIM-BIT  
 Bit4 - PIMG-BIT  
 Bit5 - mLDP-BIT  
 Bit6 - mLDPG-BIT

-- MRT Algr ID: This identifies the particular MRT algorithm used by the router. By having an Algorithm ID, it is possible to change the algorithm used or use different ones in different networks. It may be desirable to advertise a list ordered by preference to allow transitions.

```

+--+--+--+--+--+--+--+
|0|1|2|*|*|*|*|*|
+--+--+--+--+--+--+--+

```

Bit0 - LP-BIT  
 Bit1 - SPF-BIT  
 Bit2 - HYBRID-BIT

-- MRT Fd Mecha: This specifies which forwarding mechanisms the router supports. If IP-in-IPv4 or IP-in-IPv6 are used as forwarding mechanisms for IP, Red MRT Loopback Address and Blue MRT Loopback Address should be advertised by the Multi-Topology Reachable IPv4/IPv6 Prefixes TLV ([RFC4915]).

```

+--+--+--+--+--+--+--+
|0|*|2|3|4|*|*|*|
+--+--+--+--+--+--+--+

```

Bit0: LDP Destination-Topology Label  
 Bit2: IP-in-IPv4  
 Bit3: IP-in-IPv6  
 Bit4: Encode MT-ID in Labels

-- GADAG Root Election Priority: This specifies the priority of the router for being used as the GADAG root of its island. A GADAG root is elected from the set of routers with the highest priority; ties are broken based upon highest Router ID. The sensitivity of the MRT Algorithms to GADAG root selection is still being evaluated. This provides the network operator with a knob to force particular GADAG root selection.

## 4. Elements of Procedure

### 4.1. IS-IS

#### 4.1.1. Sending

MRT FRR sub-TLV is encapsulated in the Router Capability TLV and advertised through LSP PDU in the level-wide. MRT FRR sub-TLV is an optional TLV. If the router cannot support MRT FRR, it MUST NOT send the sub-TLV. Since the advertisement scope of the MRT sub-TLV is level-wide, when it is advertised the D-Bit and S-Bit of the Router Capability TLV MUST be set as 0. If other sub-TLVs in the Router Capability TLV need different values for the two bits, there MUST be an independent Router Capability TLV for the MRT FRR sub-TLV.

If the router can support multiple MRT FRR instances, there can be multiple MRT FRR sub-TLVs to be advertised. In these sub-TLVs and

there are different primary MT-IDs and associated Red/Blue MT-IDs. MT-IDs advertised by the MRT FRR sub-TLV MUST NOT be the reserved values for MT-ID([RFC5120]). In one MRT FRR sub-TLV, the Blue MT-ID MUST be different from the Red MT-ID.

MRT FRR sub-TLV MUST advertise the actual MRT capabilities, MRT algorithms and MRT forwarding mechanisms that the router can support. The corresponding bit MUST be set as 1 if it is supported. Otherwise, it MUST be set as 0.

GADAG Root Election Priority is a 16-bit unsigned integer which default value is set to 0x8000. The higher numerical value means the higher priority. GADAG Root Election is ordered with the priority and the IS-IS system ID is used as the tiebreaker. That is, if the priority is the same, the router with higher IS-IS system ID will be chosen. When a new MRT-capable router is added and its priority is higher than the current root, the MRT island will recalculate GADAG and new Blue/Red next hops for each prefix in the primary topology. If the current root fails, the new root will be re-elected and MRT calculation will be done according to the new root. Routers that are marked as overloaded([RFC3787]) is not qualified as candidate for root selection.

When MRT related information is changed in the router or existing IS-IS LSP mechanisms are triggered for refresh or update, MRT sub-TLV MUST be advertised with LSP.

#### 4.1.2. Receiving

MRT capability SHOULD NOT affect the peer setup and the routing calculation of the default SPF topology.

MRT FRR sub-TLV should be validated like other sub-TLVs when received. MRT FRR sub-TLV is also taken for the checksum calculation and authentication.

If MRT FRR sub-TLV is received and the D-Bit and S-Bit of the router capability TLV may be not 0, MRT FRR sub-TLV MUST NOT leak to other levels. If the receiver router can not support MRT, it MUST ignore MRT FRR sub-TLV.

If multiple MRT FRR sub-TLVs are received in one LSP, it means multiple MRT instances are supported by the sender router. If the MT-ID conflict is found in the multiple MRT FRR sub-TLVs, the associated sub-TLVs MUST be ignored.

The MRT capability field, the MRT algorithm field and the MRT forwarding mechanism field MUST NOT be 0. If one of these field is

0, the sub-TLV MUST be ignored.

According to the group {primary MT-ID, the Red MRT MT-ID and the Blue MRT MT-ID} identified by the received MRT FRR sub-TLV, the receiver router will take the MRT capability for intersection. If there is no intersection, the router will stop processing for the group. For the MRT algorithm, the receiver router will also take it for intersection. If there is no intersection, the router will stop processing. For the MRT forwarding mechanism, the receiver router not only check if there is intersection, but also check if the intersection found can satisfy the requirement of the MRT capability intersection.

After the intersection is found for the group {primary MT-ID, the Red MRT MT-ID and the Blue MRT MT-ID}, the receiver router will elect the root node according to the GADAG Root Election Priority. If there are updates about the priority for existing MRT FRR sub-TLV or the new MRT FRR sub-TLV is received or the current root node fails, the receiver will recalculate GADAG and new Blue/Red next hops for each prefix in the primary topology.

## 4.2. OSPF

### 4.2.1. Sending

MRT FRR TLV is encapsulated in the router information LSA whose opaque type is 10 advertised in the area-wide. MRT FRR TLV is an optional TLV. If the router can not support MRT FRR, it MUST NOT send the TLV.

If the router can support multiple MRT FRR instances, there can be multiple MRT FRR TLVs to be advertised. In these TLVs and there are different primary MT-IDs and associated Red/Blue MT-IDs. MT-IDs advertised by the MRT FRR TLV MUST NOT be the reserved values for MT-ID([RFC4915]). In one MRT FRR TLV, the Blue MT-ID MUST be different from the Red MT-ID.

MRT FRR TLV MUST advertise the actual MRT capabilities, MRT algorithms and MRT forwarding mechanisms that the router can support. The corresponding bit MUST be set as 1 if it is supported. Otherwise, it MUST be set 0.

GADAG Root Election Priority is a 16-bit unsigned integer which default value is set to 0x8000. The higher numerical value means the higher priority. GADAG Root Election is ordered with the priority and the OSPF Router ID is used as the tiebreaker. That is, if the priority is the same, the router with higher OSPF Router ID will be chosen. When a new MRT-capable router is added and its priority is

higher than the current root, the MRT island will recalculate GADAG and new Blue/Red next hops for each prefix in the primary topology. If the current root fails, the new root will be re-elected and MRT calculation will be done according to the new root. Routers that are marked as stub router([RFC3137]) are not qualified as candidate for root selection.

When MRT related information is changed in the router or existing OSPF LSA mechanisms are triggered for refresh or update, MRT TLV MUST be advertised with LSA.

#### 4.2.2. Receiving

MRT capability SHOULD NOT affect the neighbor setup and the routing calculation of the default SPF topology.

MRT FRR TLV should be validated like other TLVs when received. MRT FRR TLV is also taken for the checksum calculation and authentication.

The MRT capability field, the MRT algorithm field and the MRT forwarding mechanism field MUST NOT be 0. If one of these field is 0, the TLV MUST be ignored.

According to the group {primary MT-ID, the Red MRT MT-ID and the Blue MRT MT-ID} identified by the received MRT FRR TLV, the receiver router will take the MRT capability for intersection. If there is no intersection, the router will stop processing for the group. For the MRT algorithm, the receiver router will also take it for intersection. If there is no intersection, the router will stop processing. For the MRT forwarding mechanism, the receiver router not only check if there is intersection, but also check if the intersection found can satisfy the requirement of the MRT capability intersection.

After the intersection is found for the group {primary MT-ID, the Red MRT MT-ID and the Blue MRT MT-ID}, the receiver router will elect the root node according to the GADAG Root Election Priority. If there are updates about the priority for existing MRT FRR TLV or the new MRT FRR TLV is received or the current root node fails, the receiver will recalculate GADAG and new Blue/Red next hops for each prefix in the primary topology.

#### 5. Backward Compatibility

The MRT FRR TLVs defined in this document do not introduce any interoperability issue. For OSPF, a router not supporting the MRT

FRR TLV SHOULD just silently ignore the TLV as specified in [RFC2370]. For an IS-IS, a router not supporting the MRT FRR sub-TLV SHOULD just silently ignore the sub-TLV.

## 6. IANA Considerations

### 6.1. IS-IS

This document introduces a new sub-TLV for IS-IS. The type is to be determined by IANA.

### 6.2. OSPF

This document introduces a new TLV for OSPF. The type is to be determined by IANA.

## 7. Security Considerations

This routing extension is not currently believed to introduce new security concerns.

## 8. Normative References

- [I-D.enyedi-rtgwg-mrt-frr-algorithm]  
Atlas, A., Envedi, G., Csaszar, A., and A. Gopalan,  
"Algorithms for computing Maximally Redundant Trees for  
IP/LDP Fast- Reroute",  
draft-enyedi-rtgwg-mrt-frr-algorithm-02 (work in  
progress), October 2012.
- [I-D.ietf-rtgwg-mrt-frr-architecture]  
Atlas, A., Kebler, R., Envedi, G., Csaszar, A.,  
Konstantynowicz, M., White, R., and M. Shand, "An  
Architecture for IP/LDP Fast-Reroute Using Maximally  
Redundant Trees", draft-ietf-rtgwg-mrt-frr-architecture-01  
(work in progress), March 2012.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate  
Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC3137] Retana, A., Nguyen, L., White, R., Zinin, A., and D.  
McPherson, "OSPF Stub Router Advertisement", RFC 3137,  
June 2001.
- [RFC3787] Parker, J., "Recommendations for Interoperable IP Networks

using Intermediate System to Intermediate System (IS-IS)", RFC 3787, May 2004.

- [RFC4915] Psenak, P., Mirtorabi, S., Roy, A., Nguyen, L., and P. Pillay-Esnault, "Multi-Topology (MT) Routing in OSPF", RFC 4915, June 2007.
- [RFC4970] Lindem, A., Shen, N., Vasseur, JP., Aggarwal, R., and S. Shaffer, "Extensions to OSPF for Advertising Optional Router Capabilities", RFC 4970, July 2007.
- [RFC4971] Vasseur, JP., Shen, N., and R. Aggarwal, "Intermediate System to Intermediate System (IS-IS) Extensions for Advertising Router Information", RFC 4971, July 2007.
- [RFC5120] Przygienda, T., Shen, N., and N. Sheth, "M-ISIS: Multi Topology (MT) Routing in Intermediate System to Intermediate Systems (IS-ISs)", RFC 5120, February 2008.

#### Authors' Addresses

Zhenbin Li  
Huawei Technologies  
Huawei Bld., No.156 Beiqing Rd.  
Beijing 100095  
China

Email: lizhenbin@huawei.com

Nan Wu  
Huawei Technologies  
Huawei Bld., No.156 Beiqing Rd.  
Beijing 100095  
China

Email: eric.wu@huawei.com

Quintin Zhao  
Huawei Technologies  
125 Nagog Technology Park  
Acton, MA 01719  
US

Email: quintin.zhao@huawei.com





Network Working Group  
Internet-Draft  
Intended status: Informational  
Expires: October 28, 2013

Zhenbin Li  
Tao Zhou  
Quintin Zhao  
Huawei Technologies  
Tianle Yang  
China Mobile  
April 26, 2013

Applicability of LDP Multi-Topology for Unicast Fast-reroute Using  
Maximally Redundant Trees  
draft-li-rtgwg-ldp-mt-mrt-frr-02

Abstract

In this document, procedures' considerations on the applicability of LDP MT for unicast fast-reroute using MRT are provided. The behaviors of label allocation and label forwarding entry setup with LDP Multi-Topology and MRT fast-reroute are described in details. Different application scenarios of the combining of the LDP multi-topology(MT) and unicast fast-reroute using Maximally Redundant Trees(MRT) are also analyzed.

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 RFC 2119 [RFC2119].

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 October 28, 2013.

Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

## Table of Contents

1. Introduction . . . . .	3
2. Terminology . . . . .	3
3. Operation Procedures . . . . .	4
3.1. Routing Calculation . . . . .	4
3.2. Label Distribution . . . . .	5
3.3. Forwarding Entry Creation . . . . .	5
3.4. Switchover and Re-Convergence . . . . .	5
3.5. Switchback . . . . .	6
4. Application Scenario Analysis . . . . .	6
4.1. 2-Connected Network . . . . .	6
4.2. Non-2-Connected Network . . . . .	10
4.3. Proxy Node . . . . .	11
4.4. Inter-Area and Inter-AS . . . . .	11
4.5. Partial Deployment . . . . .	14
4.6. LDP over TE . . . . .	15
4.7. IP-Only Network . . . . .	16
5. Deployment Considerations . . . . .	16
5.1. IGP MT and LDP MT . . . . .	16
5.2. Simplified Provision . . . . .	17
5.3. IGP Multi-process . . . . .	18
5.4. Multiple IGP . . . . .	21
5.5. Label Space . . . . .	22
5.6. Proxy Egress . . . . .	22
5.7. Policy Control . . . . .	22
5.8. Resource Allocations . . . . .	23
5.9. LDP DOD . . . . .	23
6. IANA Considerations . . . . .	23
7. Security Considerations . . . . .	23
8. Acknowledgements . . . . .	23
9. Normative References . . . . .	23
Authors' Addresses . . . . .	24

## 1. Introduction

[I-D.ietf-rtgwg-mrt-frr-architecture] describes the architecture based on Maximally Redundant Trees (MRT) to provide 100% coverage for fast-reroute of unicast traffic. LDP multi-topology [I-D.ietf-mpls-ldp-multi-topology] has been proposed to provide multi-topology-based unicast forwarding in the architecture. Guidance is provided for different application scenarios to improve the applicability. The analysis of the applicability of LDP MT for unicast fast-reroute using MRT is provided. The procedures are described and typical examples are provided based on LDP MT and MRT unicast FRR architecture.

When LDP MT is combined with MRT FRR, follow advantages can be achieved:

- o Provide 100% coverage for unicast traffic.
- o The complexity of the algorithm is moderate in  $O(e)$  or  $o(e + n \log n)$ .
- o Co-deployment with LFA to provide better protection coverage.
- o Simplify operation and management with few additional configurations and states introduced.
- o Inherit procedures of LDP to achieve high scalability.
- o Propose no additional change on label forwarding behavior in the forwarding plane to facilitate incremental deployment.

## 2. Terminology

Some of terminologies defined in the [I-D.ietf-rtgwg-mrt-frr-architecture] are repeated here for the clarity of this document.

- o 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.
- o 2-connected cluster: A maximal set of nodes that are 2-connected.
- o 2-edge-connected: A network graph where at least two links must be removed to partition the network.
- o 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.

- o cut-vertex: A vertex whose removal partitions the network.
- o ECMP Equal cost multi-path: Where, for a particular destination D, multiple primary next-hops are used to forward traffic because there exist multiple shortest paths from S via different output layer-3 interfaces.
- o FIB Forwarding Information Base. The database used by the packet forwarder to determine what actions to perform on a packet.
- o LFA Loop-Free Alternate. A neighbor N, that is not a primary neighbor E, whose shortest path to the destination D does not go back through the router S. The neighbor N must meet the following condition:  $\text{Distance\_opt}(N, D) < \text{Distance\_opt}(N, S) + \text{Distance\_opt}(S, D)$
- o 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.
- o Redundant Trees (RT): A pair of trees where the path from any node X to the root R along the first tree is node-disjoint with the path from the same node X to the root along the second tree. These can be computed in 2-connected graphs.
- o SPF Shortest Path First, e.g., Dijkstra's algorithm.
- o SPT Shortest path tree

### 3. Operation Procedures

#### 3.1. Routing Calculation

IGP will flood information related with MRT FRR of each router and calculate routes for all destinations based on MRT. The details for the algorithm can refer to [I-D.enyedi-rtgwg-mrt-frr-algorithm]. For each destination, there are at least three routes that are associated with default topology, red topology and blue topology. The route of red topology or blue topology will be chosen as the secondary route. During the routing calculation, consistency of all nodes in the network must be kept. In order to achieve the consistence, following rules should be specified for the MRT calculation:

-- rules for choosing the root node;

-- rules for choosing the next-hop in the blue topology and the red topology.

Rules for choosing the secondary route can be determined locally if multiple routes exist. According to [I-D.ietf-rtgwg-mrt-frr-architecture], the secondary route derived from LFA calculation is preferred since it exists in the default topology. If there is no secondary route of LFA, the secondary route will be chosen in the blue topology or the red topology.

### 3.2. Label Distribution

When LDP MT is used for MRT fast-reroute, LDP will negotiate the MT Capability with LDP peer. Once IGP calculates routes for destinations based on MRT, LDP will advertise label mapping message with corresponding MT-ID for the specific FEC. There are at least three label bindings for each FEC that are associated with default topology, red topology and blue topology. We use `L_primary` for the label binding of the default topology, `L_blue` for the label binding of the blue topology and `L_red` for the label binding of the red topology.

MT-IDs, used in IGP and LDP, for the blue topology and the red topology can be specified administratively. In order to avoid inconsistency and simplify operation and management, we suggest MT-IDs should be reserved for MRT fast-reroute usage and the MT-IDs used in IP and LDP should be consistent.

### 3.3. Forwarding Entry Creation

LDP creates label forwarding entry for each FEC in different topologies. The route calculated based on MRT determines which label binding should be chosen for each FEC in a specific topology. For the default topology, the secondary label forwarding entry is determined by the secondary route in the blue topology or the red topology. Though the forwarding entry need be chosen according to MT information, there is not any MT information which should be processed in the forwarding plane so that the existing label forwarding mechanism can be reused well for MRT fast-reroute.

### 3.4. Switchover and Re-Convergence

When failure happens, the traffic can be switched to the secondary forwarding entry by forwarding plane within 50ms. The control plane will do the re-convergence process according to the link state change. During the course of re-convergence, the micro-loop may be

produced. The methods proposed by [RFC5714] can be used to prevent such micro-loop.

MRT fast-reroute calculation should be delayed. Thus the MRT topologies are carrying traffic and should not be disrupted until the new SPF routes are installed everywhere so that traffic is moved off of the MRT topologies. This way can prevent traffic loss to some extent. On the other hand, if failure happens again in the delayed period of MRT FRR calculation, it may cause traffic loss since the secondary route entry can not be guaranteed.

### 3.5. Switchback

When link failure or node failure recovers, IGP-LDP synchronization should be used for the default topology to prevent traffic loss. During the period of IGP-LDP synchronization, MRT FRR calculation can also be done. It will provide a best-effort protection if failure happens in the period.

## 4. Application Scenario Analysis

### 4.1. 2-Connected Network

In order to explain how LDP MT works for MRT FRR, we choose the following typical topology as an example. In the figure, (a) is the original topology, (b) is the blue topology calculated by MRT and (c) is the red topology calculated by MRT.

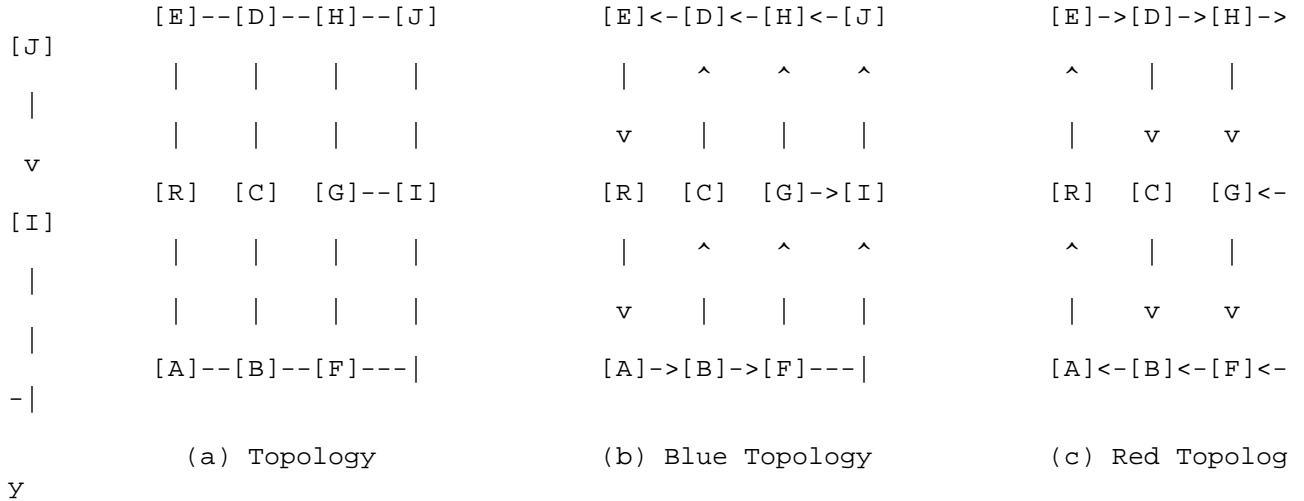


Figure 1: 2-Connected Network

According to the MRT calculation, for a specific destination H, there are following paths in different topologies for other nodes,

	Default Topology	Blue Topology	Red Topology
R	R->A->B->F->G->H	R->A->B->F->G->H	R->E->D->H
A	A->B->F->G->H	A->B->F->G->H	A->R->E->D->H
B	B->F->G->H	B->F->G->H	B->A->R->E->D->H

C	C->B->F->G->H	C->B->F->G->H	C->D->H
D	D->C->B->F->G->H	D->E->R->A->B->F	D->H
E	E->D->C->B->F->G->H	E->R->A->B->F->G->H	E->D->H
F	F->G->H	F->G->H	F->B->A->R->E->D->H
G	G->H	G->H	G->F->B->A->R->E->D->H
I	I->G->H	I->J->H	I->G->F->B->A->R->E->D->H
H			
J	J->H	J->H	J->I->G->F->B->A->R->E->D->H

Figure 2: Paths in Different Topologies for H

Note:

1. Assume that the metric of {E,R}, {D,H}, {R,C}, {G,I} and {F,I} is extreme high so that the route of the default topology is reasonable.

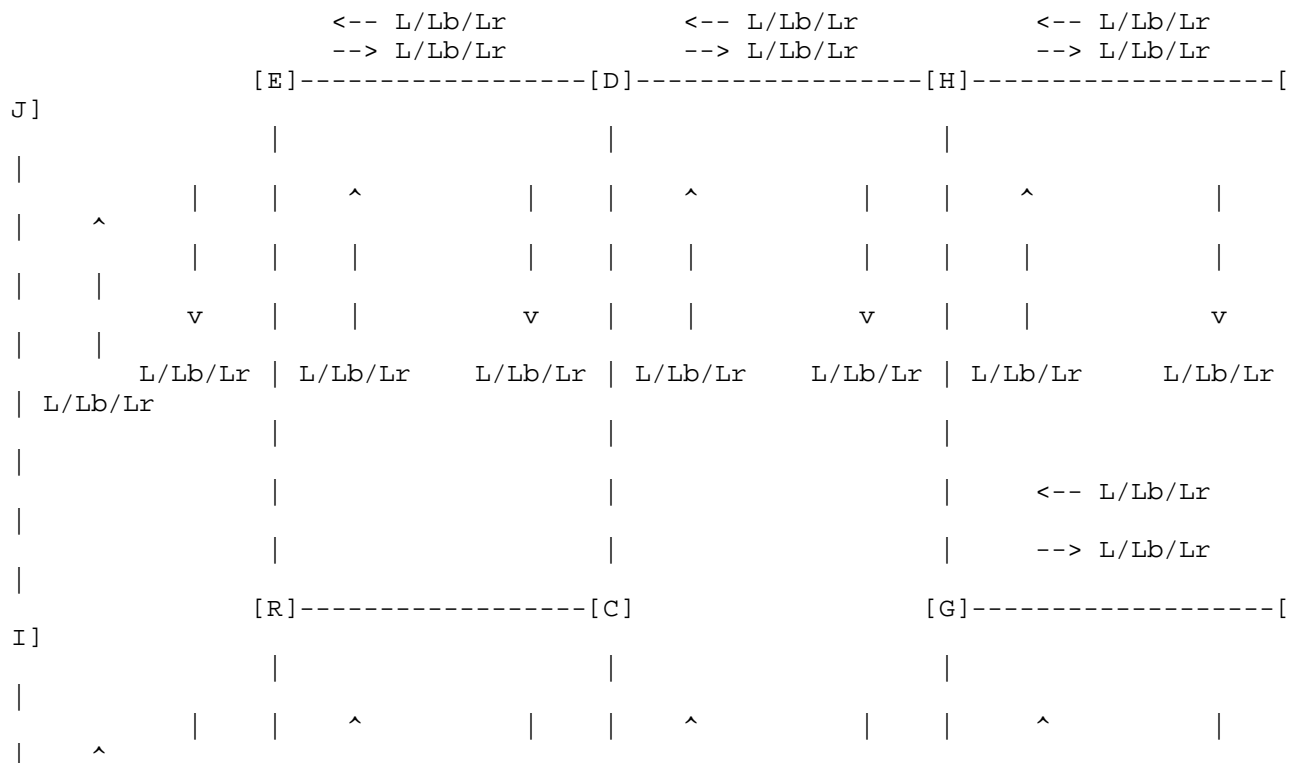
2. Assume tie-breaking rules determine that in blue topology the route from G to H chooses the path G->H instead of G->I->J->H.

3. Assume tie-breaking rules determine that in red topology the route from I to H chooses the path I->G->F->B->A->R->E->D->H instead of I->F->B->A->R->E->D->H.

4. For D node, both blue topology and red topology are available for the backup. The blue topology is preferred.

From the above calculation example, we can see that how the tie-breaking rule has to be applied when choose the nexthop in a specific topology and the topology which is used for the secondary route. For the reason of simplicity, there is no LFA calculation for the secondary route. If exists, it should be preferred.

We assume that labels are allocated in different topologies as the following figure.









## Note:

1. "<--" means the direction in which the label is distributed. For example, "<--" from D to E means that the label is distributed by D to E.

2. L means the label for H distributed in the default topology. Lb means the label for H distributed in the blue topology. Lr means the label for H distributed in the red topology.

3. L distributed by different nodes in the default topology does not mean they must be the same. This is also applied to Lb and Lr.

According to above MRT calculation result and label allocation for multi-topology, following forwarding entries will be installed for each node:

		Default Topology	Blue Topology	Red Topology
R	Ingress	--/L A /Lr E		
	Transit	L/L A /Lr E	Lb/Lb A /Lr E	Lr/Lr E
A	Ingress	--/L B /Lr R		
	Transit	L/L B /Lr R	Lb/Lb B /Lr R	Lr/Lr R
B	Ingress	--/L F /Lr A		
	Transit	L/L F /Lr A	Lb/Lb F /Lr A	Lr/Lr A
C	Ingress	--/L B /Lr D		
	Transit	L/L B /Lr D	Lb/Lb B /Lr D	Lr/Lr D
D	Ingress	--/L C /Lb E		

	Transit	L/L C	Lb/Lb E	Lr/Lr H
		/Lb E	/Lr H	
E	Ingress	--/L D		
		/Lb R		
	Transit	L/L D	Lb/Lb R	Lr/Lr D
		/Lb R	/Lr D	
F	Ingress	--/L G		
		/Lr B		
	Transit	L/L G	Lb/Lb G	Lr/Lr B
		/Lr B	/Lr B	
G	Ingress	--/L H		
		/Lr F		
	Transit	L/L H	Lb/Lb H	Lr/Lr F
		/Lr F	/Lr F	
I	Ingress	--/L G		
		/Lb J		
	Transit	L/L G	Lb/Lb J	Lr/Lr G
		/Lb J	/Lr G	
J	Ingress	--/L H		
		/Lr I		
	Transit	L/L H	Lb/Lb H	Lr/Lr I
		/Lr I	/Lr I	

Figure 4: Label Forwarding Entries Installed in Each Node for FEC H

Note:

1. For an ingress label forwarding entry as follows, when forward, L will be pushed and sent to the next hop A. If failure happens, Lr will be pushed and sent to the next hop E.

```
Ingress    --/L  A
            /Lr  E
```

2. For a transit label forwarding entry as follows, when packet with the incoming label L arrives, L will be swapped to L and sent to the next hop A. If failure happens, L will be swapped to Lr and sent to the next hop E.

```
Transit    L/L  A
            /Lr  E
```

Above forwarding entries construct the label switch path used for fast-reroute in the forwarding plane. We can see that the existing MPLS label forwarding can be used without any extension.

#### 4.2. Non-2-Connected Network

[I-D.ietf-rtgwg-mrt-frr-architecture] proposes following non-2-connected network.

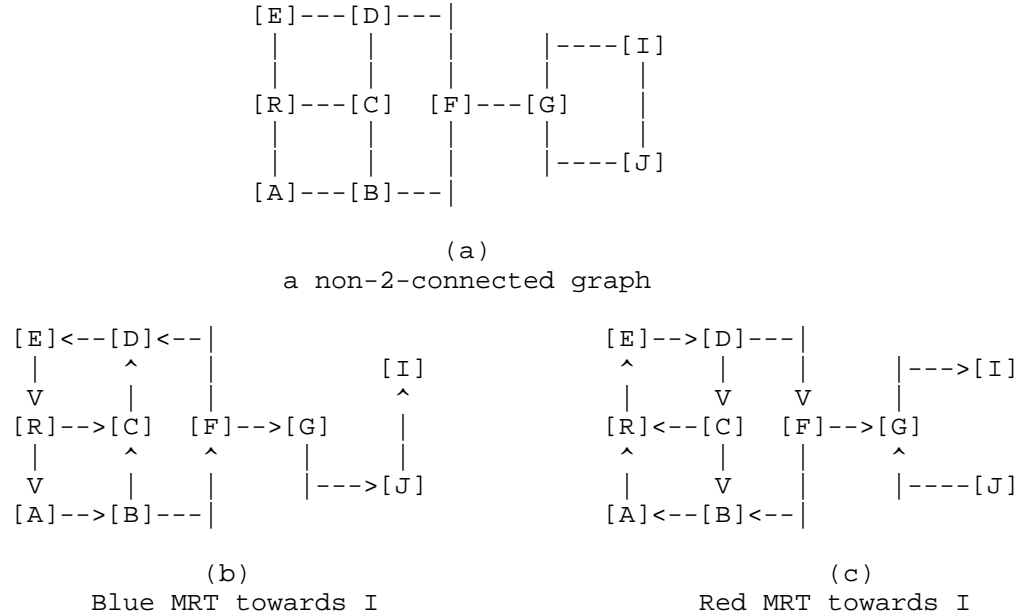


Figure 5: A non-2-connected network

We will not explain how LDP MT is applied for the MRT FRR in detail. We choose the node I as the destination and choose R and F to observe how MRT and LDP MT work for fast-reroute.

According to MRT calculation, the path from R to I in the blue topology is R->A->B->F->G->J->I and the path from R to I in the red topology is R->E->D->F->G->I. We assume in the default topology the path from R to I is R->C->D->F->G->I. Then following forwarding entry will be created in the node R for the destination I.

		Default Topology	Blue Topology	Red Topology
R	Ingress	--/L C /Lb A		
	Transit	L/L C /Lb A	Lb/Lb A /Lr E	Lr/Lr E

Figure 6: Label Forwarding Entry of Node R for FEC I

For the node F, the path from F to I in the blue topology is F->G->J->I and the path in the red is F->G->I. We assume in the default topology the path from F to I is F->G->I. The following forwarding entry will be created in the node F for the destination I.

		Default Topology	Blue Topology	Red Topology
F	Ingress	--/L G		
	Transit	L/L G	Lb/Lb G	Lr/Lr G

Figure 7: Label Forwarding Entry of Node F for FEC I

We can see that there is no secondary route in the node F for the destination I and correspondingly there is no LDP FRR forwarding entry.

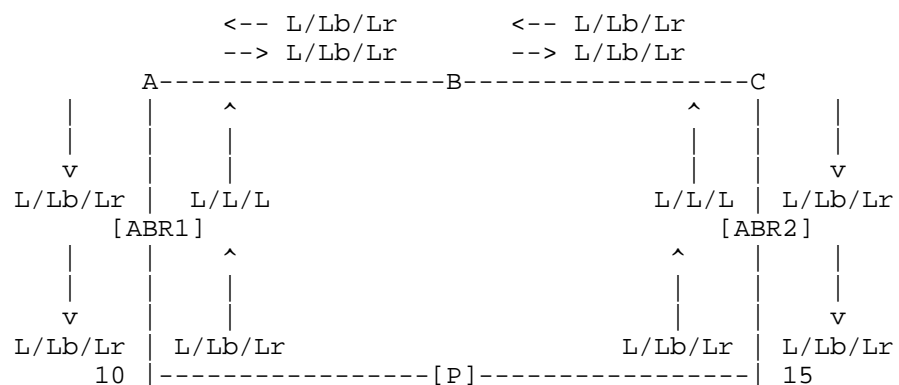
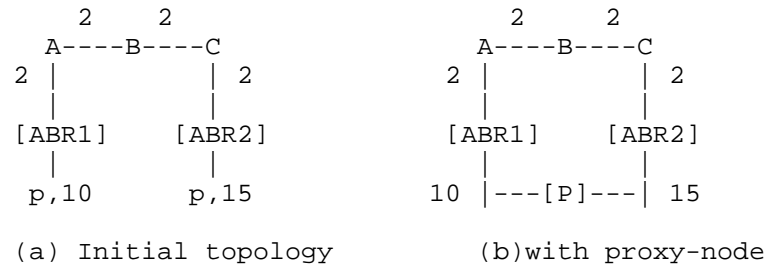
#### 4.3. Proxy Node

There are several application scenarios proposed by [I-D.ietf-rtgwg-mrt-frr-architecture] which will use proxy node for MRT. That is, if a set of prefixes are advertised by border routers of an MRT island, a single proxy node can be used to represent the set and the proxy node and associated links are added to the network topology for MRT calculation. The application scenarios include inter-area, inter-AS and partial deployment of compatible MRT FRR routers.

#### 4.4. Inter-Area and Inter-AS

For Inter-area scenarios, it is desirable to go back to the default forwarding topology when leaving an area/level. There are two mechanisms proposed by [I-D.ietf-rtgwg-mrt-frr-architecture]. The first one is that ABR will advertise different labels for one specific FEC in different areas. The second one is that penultimate hop pop is done through additional computation by the penultimate router along the in-local-area MRT immediately before the ARB/LBR is reached. The first one need change of the traditional label allocation method for LDP which always distributes the same label for one FEC to all peers. When the second one used, it must be

guaranteed that the IP forwarding should be done by ABR. If there is an inner label, it will cause wrong forwarding behavior. Since it is difficult to determine the type of the packet, the second mechanism must be used carefully. In order to optimize the second mechanism, when the penultimate router receive a packet with MRT label, it can swap the label to corresponding FEC's default topology label instead of penultimate hop pop.



## (d) Label Distribution Change

Note: In (b),(c) and (d), label distributed by proxy nodes is actually distributed for proxy nodes by nodes in different areas from A/B/C nodes.

Figure 8: Inter-area Network and LDP MT Label Distribution for MRT FRR

According to the label distribution and MRT computation as shown in (c) of the above figure, following forwarding entries can be created in the node ABR1 and ABR2:

		Default Topology	Blue Topology	Red Topology
ABR1	Ingress	--/L P /Lr A		
	Transit	L/L P /Lr A	Lb/Lb P /Lr A	Lr/Lr A
ABR2	Ingress	--/L P /Lb C		
	Transit	L/L P /Lb C	Lb/Lb C /Lr P	Lr/Lr P

Figure 9: Label Forwarding Entry of Node ABR1 and ABR2 for Proxy Node

If the first method on change of label allocation as shown in (d) of the above figure, following forwarding entry will be created in the node A and C:

		Default Topology	Blue Topology	Red Topology
A	Ingress	--/L ABR1 /Lr B		
	Transit	L/L ABR1 /Lr B	Lb/L ABR1 /Lr B	Lr/Lr B
C	Ingress	--/L ABR2 /Lb B		
	Transit	L/L ABR2 /Lb B	Lb/Lb B /L ABR2	Lr/L ABR2

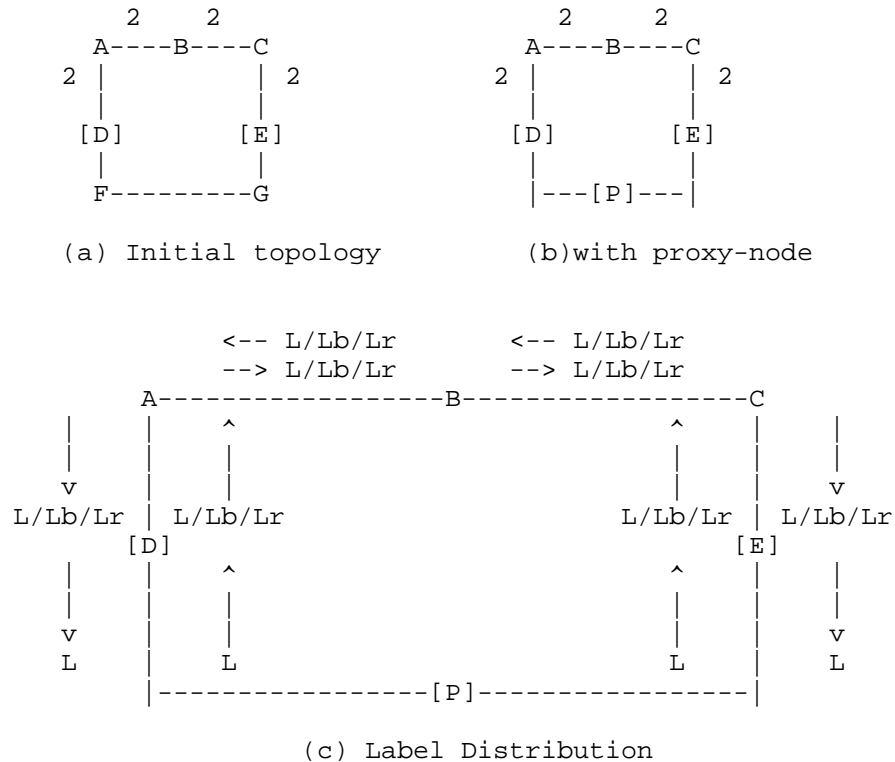
Figure 10: Label Forwarding Entry of Node A and C for Proxy Node

For inter-AS scenarios, prefixes advertised by ASBRs will set up LSP in the default topology as proxy egress. The number of prefixes will have great effect on the label allocation of LDP. When MRT fast-reroute deploys, it should be confirmed firstly that labels are enough. Or else, MRT will have negative effect on the deployment of normal service. Besides this, the complexities for ASBR protection

has been proposed by [I-D.ietf-rtgwg-mrt-frr-architecture]. It need further study.

#### 4.5. Partial Deployment

For partial deployment and islands of compatible MRT FRR routers, proxy nodes and associated links are added to the network topology for MRT computation. The difference between partial deployment and inter-area is that in the partial deployment scenario the border nodes need proxy egress process for LDP in the blue topology and the red topology. That is, in the blue topology and red topology, the border node of the MRT network topology is not the actual egress for a prefix out of the MRT network. The border node has to advertise label for the prefix as the proxy egress.



Note: In (c), label distributed by proxy nodes is actually distributed for proxy nodes by nodes connected to [D] and [E].

Figure 11: Partial Deployment Network and LDP MT Label Distribution for MRT FRR



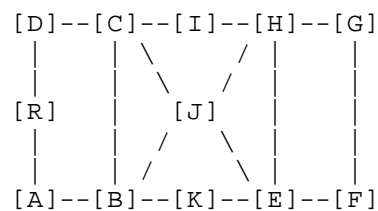
According to the label distribution and MRT computation as shown in (c) of the above figure, following forwarding entries can be created in the node D and E:

		Default Topology	Blue Topology	Red Topology
D	Ingress	--/L P		
		/Lr A		
	Transit	L/L P	Lb/L P	Lr/Lr A
E	Ingress	--/L P	/Lr A	
		/Lb C		
	Transit	L/L P	Lb/Lb C	Lr/L P
		/Lb C	/L P	

Figure 12: Label Forwarding Entry of Node D and E for Proxy Node

#### 4.6. LDP over TE

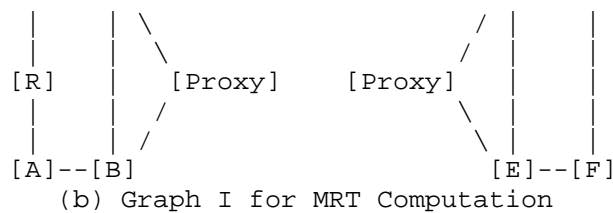
There is also additional complexity for LDP over TE scenario in which nodes in the LDP domain need to calculate two different MRT FRR for different nodes in the MPLS TE domain: edge nodes which can support both LDP and TE and internal nodes which only supports TE. Edges nodes combine with nodes in the LDP domain to form a complete topology for MRT FRR calculation. Internal nodes don't support LDP, but are not hidden from IGP topology. Some IP traffic to internal nodes which do not support LDP maybe need MRT FRR provided by nodes in the LDP domain. Nodes in the LDP domain calculate MRT FRR for these internal nodes like partial deployment. An example deployment is shown in the following figure. LDP over TE is deployed in edge nodes B, C, E and H. Internal nodes I, J and K do not support LDP. For MRT FRR, the deployment can be seen as two independent topologies. For internal nodes I, J and K, as shown in the figure (b) it is similar as the process of partial deployment. For other nodes, as shown in the figure (c) it is similar as the process of 2-connected network and the bidirectional MPLS TE paths can be used as the virtual links in MRT computation.



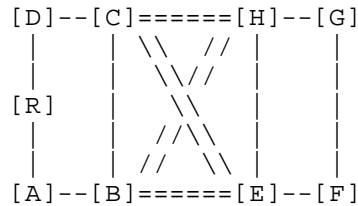
(a) Default Topology

[D]--[C]

[H]--[G]



(b) Graph I for MRT Computation



(c) Graph II for MRT Computation

Figure 13: LDP over TE Network and LDP MT Label Distribution for MRT FRR

#### 4.7. IP-Only Network

In the IP-only network IP-in-IP has to be used. This means additional loopback addresses have to be introduced. And they are announced with associated MRT color. It will propose complexities for operation and management of the network. We recommend LDP MT should be deployed in the network for the fast-reroute usage to reduce the complexities. It also will not introduce any complexity of IP MT forwarding in the ingress node since the multi-topology only takes effect for protection. Comparing with tunnel IP packet in IP, LDP MT is an easy way to provision fast-reroute.

### 5. Deployment Considerations

#### 5.1. IGP MT and LDP MT

MRT computation can be seen as a local process for IGP if only the computation is consistent for all nodes of the network. That is, multi-topology is not necessary for IGP to advertise link states with MT-ID. MT-ID is only advertised in LDP for LDP's FEC usage. That is, for MRT fast-reroute, IGP MT-ID can be independent of LDP MT-ID. But this proposes complexity for operation and management. It seems desirable to keep the consistency of MT-IDs for both IGP and LDP.

There exists another issue regarding the relationship of IGP and LDP. IGP does not support IPv4 and IPv6 in one topology. When multi-topology is used for MRT fast-reroute, the blue topology and the red topology of IPv4 should be different from those of IPv6. However,

for LDP, the address family is adopted for FEC in one multi-topology. Label distribution should be done for both IPv4 and IPv6 in one multi-topology. If the MT-ID is consistent for IGP and LDP, there should be four MT-IDs for IPv4 and IPv6 in one MRT network. For protocol extensions of MRT fast-reroute, both IPv4 and IPv6 should be taken into account for IGP to advertise information related with the blue topology and the red topology.

Besides the inconsistency of IGP MT and LDP MT, there exists the inconsistency between the OSPF MT[RFC4915] and the IS-IS MT[RFC5120]. Different MT-ID ranges for OSPF and IS-IS which cause the difficulty in reserving the same MRT MT-IDs for OSPF and IS-IS.

When multi-topology is used for MRT fast-reroute, it is error-prone for MT-ID configuration for the blue topology and the red topology on all nodes of the MRT network. In order to simplify operation and management, it is recommended that MT-IDs could be reserved for the MRT fast-reroute usage. Owing to the inconsistency of OSPF MT and IS-IS MT and the inconsistency of IGP MT and LDP MT, it seems a little challenge to reserve these possible values.

## 5.2. Simplified Provision

It is necessary to configure many parameters related with MRT FRR and advertise these capabilities and information by IGP[I-D.li-rtgwg-igp-ext-mrt-frr]. It is concerned that the provision complexity will have negative effect on the utility of MRT FRR.

There are two different things for the MRT FRR provision:

The first thing is the capability information which can be supported by a MRT-FRR-enabled node. The information can be directly derived without configuration.

The second thing is what parameters should be agreed on by all nodes to compute an MRT island.

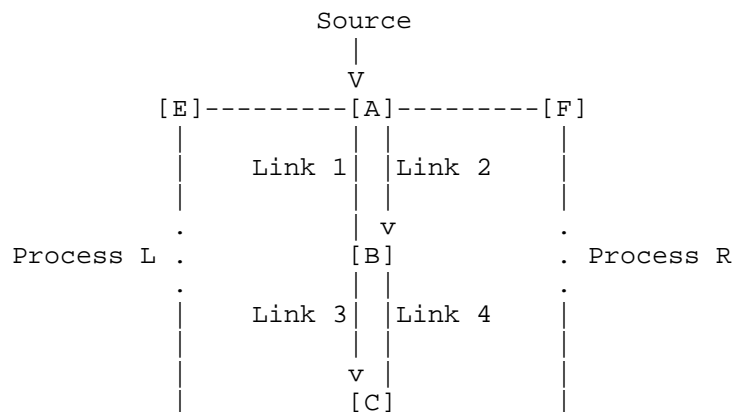
For example, as to [I-D.li-rtgwg-igp-ext-mrt-frr], a node can advertise the supported different algorithms through IGP. The supported algorithms are internal capability which is not necessary to configure. After all nodes advertise the information, they should choose one specific algorithm to compute MRT FRR. This has to be configured or all nodes should agree on a default value in advance.

The second thing should be considered more in order to simplify MRT FRR provision. In fact, LDP MT is just the method to simplify the MRT FRR provision comparing with the method that tunnels IP packet in

IP. If the latter method is preferred, it will be more difficult to design IP address carefully for each node than that only blue and red MT IDs is chosen for all nodes in the former method. There are few parameters for LDP-MT-based MRT FRR to be provisioned. The key parameters is just MT IDs and the algorithm's related parameters. As in the section 3.6, It is strongly recommended that the IGP and LDP MT IDs used for MRT FRR should be reserved. It is also the preferred default value for MRT algorithms should be defined in the appropriate documents. By this way a default profile for MRT FRR provision is determined which is composed of a set of default values. This can simplify the MRT FRR provision greatly. If it is not available, all nodes can agree on a internal default profile which is determined by the implementation and save configuration work for MRT FRR. If new nodes add to the network which use different default MT ID values and algorithm-related parameters, it can be changed administratively.

### 5.3. IGP Multi-process

IGP multi-process is used to isolate different areas. If an ip prefix is advertised in multiple processes, each process will calculate routes for the prefix and the shortest one will be chosen to install forwarding entry. Each IGP process calculates routes of MRT FRR independently and has its own pair of MRT topology (blue MRT topology and red MRT topology. Since the MRT paths maybe different in different process, one process' MRT next hop can not be used in another process for a specific prefix to avoid loop. So the primary route and its MRT next hop must be chosen in the same process. In order to achieve this object, there should be different blue MRT MT-IDs and red MRT MT-IDs for these processes. If there are only one pair of MRT topology for multiple processes (i.e. there is only one blue MRT MT-ID and one red MRT MT-ID), loop can happen for MRT FRR when each node chooses its primary next hop and the MRT routes in the same process for the multiple processes.



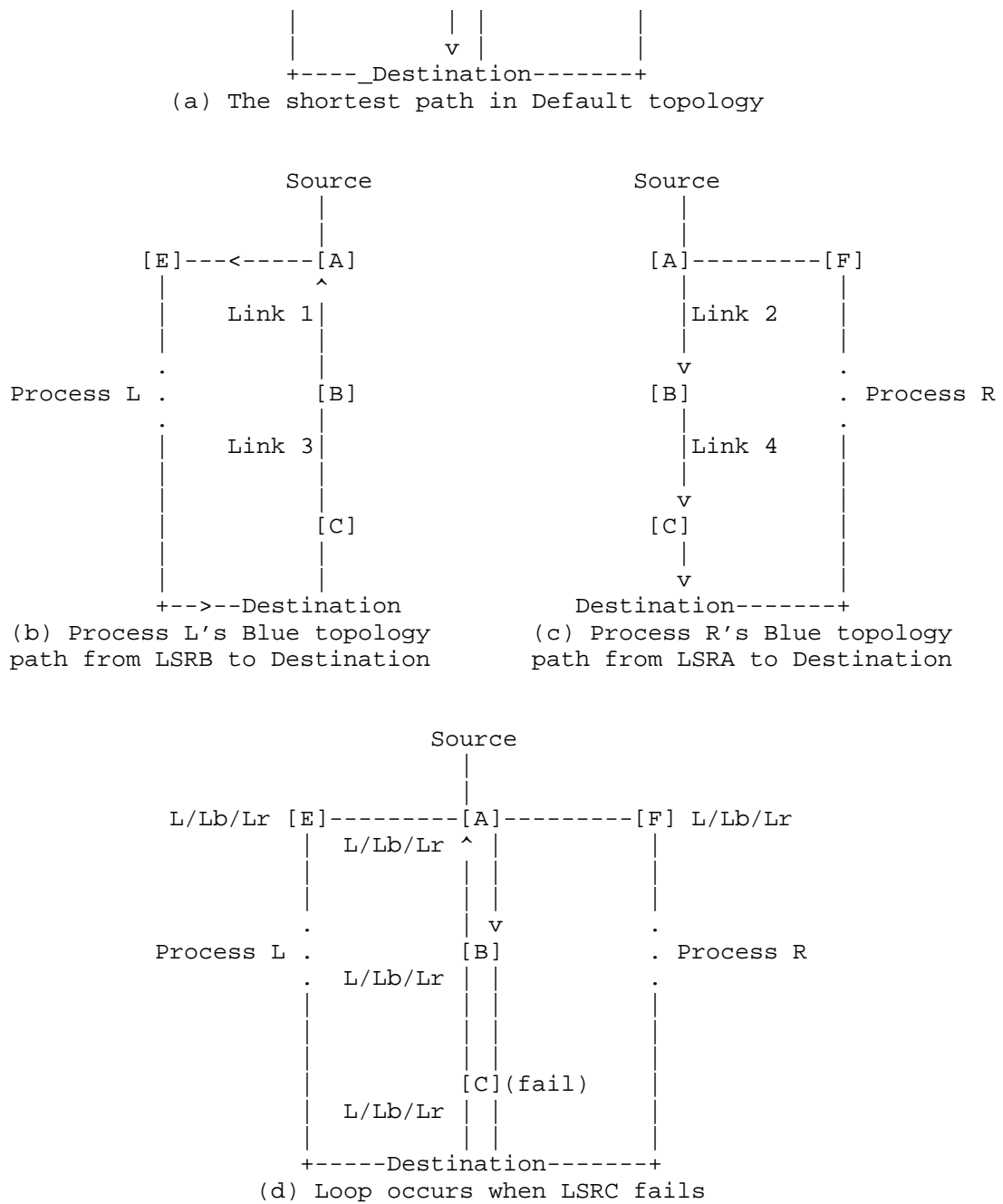


Figure 14: Loop Issue in IGP Mul-tiprocess

An IGP multi-process deployment is shown in the above figure. Node A, B and C are located in both processes: the left process L and the right process R. The process L is a ring topology, including link1 and link3. And the process R is also a ring topology, including link2 and link4. For the traffic from the Source to the Destination, assume that A chooses the shortest path determined by the process R and using link2 as the primary next hop and B chooses the shortest path determined by the process L and using link3 as the primary next hop. Process L and process R calculate MRT topologies independently, but there is only one pair of MRT MT-IDs and the label distribution is the same for different processes, this will cause the following forwarding entries are installed:

Node B: The shortest path is determined by process L. The MRT path is calculated in the same process. Assume that B calculates the blue MRT topology shown in the (b) and chooses link1 in the blue MRT topology as its secondary route. Then there is following forwarding entries for node B:

	Default Topology	Blue Topology	Red Topology
B Transit	L/L C /Lb A	Lb/Lb A /Lr C	Lr/Lr C

Node A: The shortest path is determined by process R. The MRT path is calculated in the same process. Assume that A calculates the blue MRT topology shown in the (c). Then there is following forwarding entries for node A:

	Default Topology	Blue Topology	Red Topology
A Transit	L/L B /Lr F	Lb/Lb B /Lr F	Lr/Lr F

According to above forwarding entries, if node C fails, the traffic will be sent by B to A with label Lb using the secondary route. When A receives the traffic with label Lb, it will send the traffic to B using the forwarding entry for the blue topology. Then loop happens for the traffic.

The solution of the issue is to use different MRT MTs for different processes. That is, different MRT topologies should be provisioned for different processes so that the different label distribution is done for the multiple processes. This will guarantee that when failure happens the switched traffic will be forwarded in the same process. The following figure shows the solution for as to the above example:

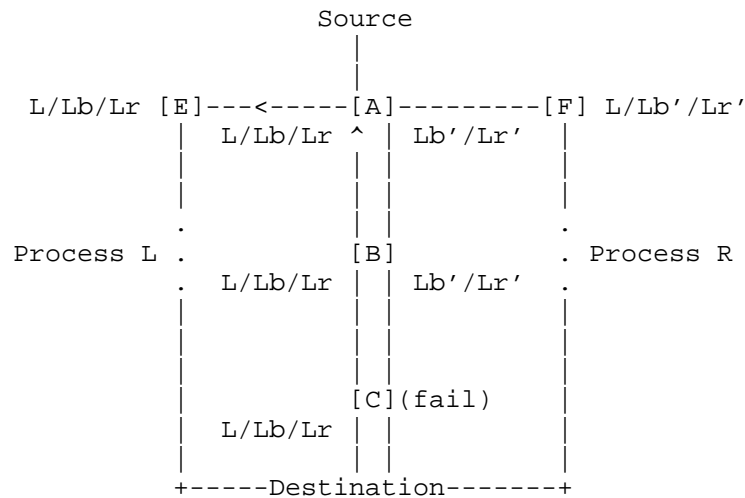


Figure 15: Separate MRT MT for Multi-process

Following forwarding entries will be created for A and B. We can see that if failure happens, the switched traffic is forwarded from A to B to E and the loop issue is avoided.

		Default Topology		Blue MT		Red MT		Blue MT'		Red MT'	
A	Transit	L/L	B	Lb/Lb	E	Lr/Lr	B	Lb'/Lb'	B	Lr'/Lr'	F
		/Lr'	F	/Lr	B			/Lr'	F		
B	Transit	L/L	C	Lb/Lb	A	Lr/Lr	C	Lb'/Lb'	C	Lr'/Lr'	A
		/Lb	A	/Lr	C			/L'	A		

Owing to the loop issue in the IGP multi-process scenario, it must be checked carefully for the reserved MT-IDs or the default profile described above for simplifying provision which will cause multiple processes share the same MRT MT-IDs. In order to prevent loop issue, separate MRT MTs for IGP multi-process have to be taken into account.

#### 5.4. Multiple IGP

If multiple IGPs deploy in one network, the best route will be determined according to priority of these IGPs. This might cause the inconsistency issue for MRT fast-reroute. For example, when IS-IS and OSPF are deployed in one network, some nodes will use the best reroute computed by IS-IS and some nodes will use the best route computed by OSPF. If the link state is not consistent in IS-IS and OSPF, the MRT fast-reroute cannot work well. It is highly desirable that in one network only one IGP protocol is deployed and link states should be guaranteed consistent if multiple IGPs deploys.

### 5.5. Label Space

Advantages of LDP MT in MRT fast-reroute are apparent for simplified operation and management comparing with using IP tunnel. The main issue of LDP MT for MRT fast-reroute is resource occupancy. MRT FRR need create two redundant topologies to provide backup path. The two topologies cover all links and nodes of the MRT network. It will impact on the system resource occupancy since it will also take more resource to install routes and label forwarding entries for different topologies. When deploying LDP MT for MRT FRR, especially in the scenario of upgrading, consideration should be taken so that there is enough system resource to accommodate more routes and forwarding entries. Besides the issue related with resource occupancy, label usage is also an important issue to be taken into account. For one FEC, there are at least three label bindings distributed by one router. The number of labels for MRT fast-route is triple of that of the network without MRT fast-reroute. When LDP MT for MRT FRR is deployed, it should be guaranteed that enough labels are available so that it will not have impact on normal services such as L2VPN, L3VPN, etc.

### 5.6. Proxy Egress

In several scenarios where MRT FRR is deployed, proxy egress LSPs have to be setup by LDP. The proxy egress LSP maybe not end-to-end to bear VPN service in the network. But it will deteriorate label usage if LDP MT is deployed for MRT FRR. It is highly desirable that such unnecessary LSPs should be prohibited to setup to facilitate MRT FRR deployment.

### 5.7. Policy Control

Policy can be used to reducing the effect of more labels for MRT FRR. It is important to control on the setup of LSP in the default topology. There are two basic scenarios. The first one is the IP-only network. It is difficult to control the number of LSPs for protection since LDP MT is an extension for IP to implement protection. The second one is the multi-service network based on VPN. Policy can be applied to permit only host addresses to setup LSPs.

Policy is not recommended to control on LSP in the blue topology and the red topology since it is easy to cause inconsistency of the protection. For example, if one node need to set up MRT backup LSP for one FEC but this FEC is not allowed creating LSP by the policy in the MRT topologies, then the node cannot create the MRT backup LSP.



## 5.8. Resource Allocations

During the deployment of this solution, more system resource and extra label occupancy must be taken into account to avoid the possible resource exhausting.

## 5.9. LDP DoD

LDP DoD is used in some scenarios such as Seamless MPLS[I-D.ietf-mpls-seamless-mpls]. When MRT fast-reroute is deployed, label request will be sent according to the path calculated for different topology. The label forwarding entry will be created as the method above. Comparing with LDP DU, there are less label binding distribution for LDP DoD. In addition, LDP DoD is always used combining with conservative label retention mode. Thus there is no label binding distributed for the secondary route calculated in the default topology so that LFA cannot not be used easily. The label forwarding entry in the blue topology or the red topology will be used as the secondary one directly.

## 6. IANA Considerations

This document makes no request of IANA.

## 7. Security Considerations

There is no security issue introduced by this specification.

## 8. Acknowledgements

## 9. Normative References

[I-D.enyedi-rtgwg-mrt-frr-algorithm]

Atlas, A., Enyedi, G., Csaszar, A., and A. Gopalan,  
"Algorithms for computing Maximally Redundant Trees for IP  
/LDP Fast- Reroute", draft-enyedi-rtgwg-mrt-frr-  
algorithm-02 (work in progress), October 2012.

[I-D.ietf-mpls-ldp-multi-topology]

Zhao, Q., Fang, L., Zhou, C., Li, L., and K. Raza, "LDP  
Extensions for Multi Topology Routing", draft-ietf-mpls-  
ldp-multi-topology-06 (work in progress), December 2012.

[I-D.ietf-mpls-seamless-mpls]

Leymann, N., Decraene, B., Filsfils, C., Konstantynowicz,  
M., and D. Steinberg, "Seamless MPLS Architecture", draft-  
ietf-mpls-seamless-mpls-02 (work in progress), October  
2012.

- [I-D.ietf-rtgwg-mrt-frr-architecture]  
Atlas, A., Kebler, R., Envedi, G., Csaszar, A., Tantsura, J., Konstantynowicz, M., White, R., and M. Shand, "An Architecture for IP/LDP Fast-Reroute Using Maximally Redundant Trees", draft-ietf-rtgwg-mrt-frr-architecture-02 (work in progress), February 2013.
- [I-D.li-rtgwg-igp-ext-mrt-frr]  
Li, Z., Wu, N., and Q. Zhao, "Routing Extension for Fast-Reroute Using Maximally Redundant Trees", draft-li-rtgwg-igp-ext-mrt-frr-01 (work in progress), March 2013.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC4915] Psenak, P., Mirtorabi, S., Roy, A., Nguyen, L., and P. Pillay-Esnault, "Multi-Topology (MT) Routing in OSPF", RFC 4915, June 2007.
- [RFC5120] Przygienda, T., Shen, N., and N. Sheth, "M-ISIS: Multi Topology (MT) Routing in Intermediate System to Intermediate Systems (IS-ISs)", RFC 5120, February 2008.
- [RFC5714] Shand, M. and S. Bryant, "IP Fast Reroute Framework", RFC 5714, January 2010.

## Authors' Addresses

Zhenbin Li  
Huawei Technologies  
Huawei Bld., No.156 Beiqing Rd.  
Beijing 100095  
China

Email: lizhenbin@huawei.com

Tao Zhou  
Huawei Technologies  
Huawei Bld., No.156 Beiqing Rd.  
Beijing 100095  
China

Email: tao.chou@huawei.com

Quintin Zhao  
Huawei Technologies  
125 Nagog Technology Park  
Acton, MA 01719  
US

Email: quintin.zhao@huawei.com

Tianle Yang  
China Mobile  
32, Xuanwumenxi Ave.  
Beijing 01719  
China

Email: yangtianle@chinamobile.com

Routing Area Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: August 22, 2013

S. Litkowski  
B. Decraene  
Orange  
C. Filsfils  
K. Raza  
Cisco Systems  
February 18, 2013

Operational management of Loop Free Alternates  
draft-litkowski-rtgwg-lfa-manageability-01

## Abstract

Loop Free Alternates (LFA), as defined in RFC 5286 is an IP Fast ReRoute (IP FRR) mechanism enabling traffic protection for IP traffic (and MPLS LDP traffic by extension). Following first deployment experiences, this document provides operational feedback on LFA, highlights some limitations, and proposes a set of refinements to address those limitations. It also proposes required management specifications.

This proposal is also applicable to remote LFA solution.

## 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].

## 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 August 22, 2013.

## Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

## Table of Contents

1. Introduction . . . . .	3
2. Operational issues with default LFA tie breakers . . . . .	3
2.1. Case 1: Edge router protecting core failures . . . . .	4
2.2. Case 2: Edge router chosen to protect core failures while core LFA exists . . . . .	5
2.3. Case 3: suboptimal core alternate choice . . . . .	6
2.4. Case 4: ISIS overload bit on LFA computing node . . . . .	7
3. Configuration requirements . . . . .	7
3.1. LFA enabling/disabling scope . . . . .	7
3.2. Policy based LFA selection . . . . .	8
3.2.1. Mandatory criteria . . . . .	8
3.2.2. Enhanced criteria . . . . .	9
4. Operational aspects . . . . .	13
4.1. ISIS overload bit on LFA computing node . . . . .	13
4.2. Manual triggering of FRR . . . . .	14
4.3. Required local information . . . . .	14
4.4. Coverage monitoring . . . . .	15
5. Security Considerations . . . . .	15
6. Contributors . . . . .	15
7. Acknowledgements . . . . .	15
8. IANA Considerations . . . . .	15
9. References . . . . .	16
9.1. Normative References . . . . .	16
9.2. Informative References . . . . .	16
Authors' Addresses . . . . .	17

## 1. Introduction

Following the first deployments of Loop Free Alternates (LFA), this document provides feedback to the community about the management of LFA.

Section 2 provides real uses cases illustrating some limitations and suboptimal behavior.

Section 3 proposes requirements for activation granularity and policy based selection of the alternate.

Section 4 express requirements for the operational management of LFA.

## 2. Operational issues with default LFA tie breakers

[RFC5286] introduces the notion of tie breakers when selecting the LFA among multiple candidate alternate next-hops. When multiple LFA exist, RFC 5286 has favored the selection of the LFA providing the best coverage of the failure cases. While this is indeed a goal, this is one among multiple and in some deployment this lead to the selection of a suboptimal LFA. The following sections details real use cases of such limitations.

Note that the use case of per-prefix LFA is assumed throughout this analysis.

## 2.1. Case 1: Edge router protecting core failures

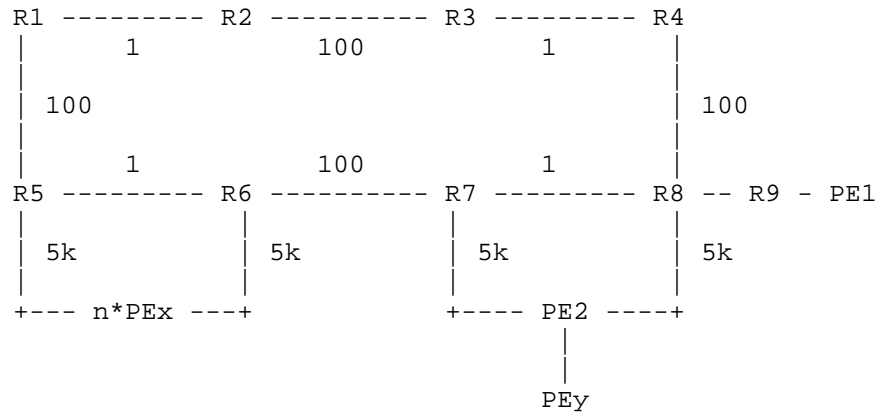


Figure 1

Rx routers are core routers using  $n \times 10G$  links. PEs are connected using links with lower bandwidth.

In figure 1, let us consider the traffic flowing from PE1 to PEx. The nominal path is R9-R8-R7-R6-PEx. Let us consider the failure of link R7-R8. For R8, R4 is not an LFA and the only available LFA is PE2.

When the core link R8-R7 fails, R8 switches all traffic destined to all the PEx towards the edge node PE2. Hence an edge node and edge links are used to protect the failure of a core link. Typically, edge links have less capacity than core links and congestion may occur on PE2 links. Note that although PE2 was not directly affected by the failure, its links become congested and its traffic will suffer from the congestion.

In summary, in case of failure, the impact on customer traffic is:

- o From PE2 point of view :
  - \* without LFA: no impact
  - \* with LFA: traffic is partially dropped (but possibly prioritized by a QoS mechanism). It must be highlighted that in such situation, traffic not affected by the failure may be affected by the congestion.





## 2.3. Case 3: suboptimal core alternate choice

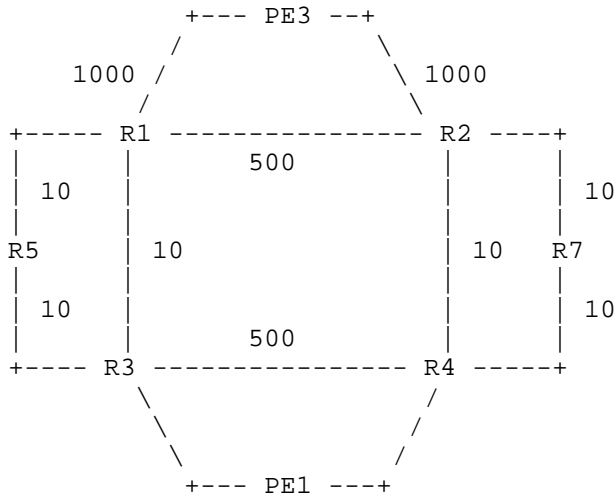


Figure 3

Rx routers are core routers. R1-R2 and R3-R4 links are 1G links. All others inter Rx links are 10G links.

In the figure above, let us consider the failure of link R1-R3. For destination PE3, R3 has two possible alternates:

- o R4, which is node-protecting
- o R5, which is link-protecting

R4 is chosen as best LFA due to its better protection type. However, it may not be desirable to use R4 for bandwidth capacity reason. A service provider may prefer to use high bandwidth links as preferred LFA. In this example, preferring shortest path over protection type may achieve the expected behavior, but in cases where metric are not reflecting bandwidth, it would not work and some other criteria would need to be involved when selecting the best LFA.

## 2.4. Case 4: ISIS overload bit on LFA computing node

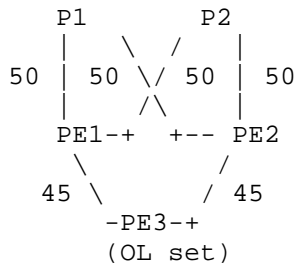


Figure 4

In the figure above, PE3 has its overload bit set (permanently, for design reason) and wants to protect traffic using LFA for destination PE2.

On PE3, the loopfree condition is not satisfied :  $100 \nless 45 + 45$ . PE1 is thus not considered as an LFA. However thanks to the overload bit set on PE3, we know that PE1 is loopfree so PE1 is an LFA to reach PE2.

In case of overload condition set on a node, LFA behavior must be clarified.

## 3. Configuration requirements

Controlling best alternate and LFA activation granularity is a requirement for Service Providers. This section defines configuration requirements for LFA.

### 3.1. LFA enabling/disabling scope

The granularity of LFA activation should be controlled (as alternate nexthop consume memory in forwarding plane).

An implementation of LFA SHOULD allow its activation with the following criteria:

- o Per address-family : ipv4 unicast, ipv6 unicast, LDP IPv4 unicast, LDP IPv6 unicast ...
- o Per routing context : VRF, virtual/logical router, global routing table, ...

- o Per interface
- o Per protocol instance, topology, area
- o Per prefixes: prefix protection SHOULD have a better priority compared to interface protection. This means that if a specific prefix must be protected due to a configuration request, LFA must be computed and installed for this prefix even if the primary outgoing interface is not configured for protection.

### 3.2. Policy based LFA selection

When multiple alternates exist, LFA selection algorithm is based on tie breakers. Current tie breakers do not provide sufficient control on how the best alternate is chosen. This document proposes an enhanced tie breaker allowing service providers to manage all specific cases:

1. An implementation of LFA SHOULD support policy-based decision for determining the best LFA.
2. Policy based decision SHOULD be based on multiple criterions, with each criteria having a level of preference.
3. If the defined policy does not permit to determine a unique best LFA, an implementation SHOULD pick only one based on its own decision, as a default behavior. An implementation SHOULD also support election of multiple LFAs, for loadbalancing purposes.
4. Policy SHOULD be applicable to a protected interface or to a specific set of destinations. In case of application on the protected interface, all destinations primarily routed on this interface SHOULD use the interface policy.
5. It is an implementation choice to reevaluate policy dynamically or not (in case of policy change). If a dynamic approach is chosen, the implementation SHOULD recompute the best LFAs and reinstall them in FIB, without service disruption. If a non-dynamic approach is chosen, the policy would be taken into account upon the next IGP event. In this case, the implementation SHOULD support a command to manually force the recomputation/reinstallation of LFAs.

#### 3.2.1. Mandatory criteria

An implementation of LFA MUST support the following criteria:

- o Non candidate link: A link marked as "non candidate" will never be used as LFA.
- o A primary nexthop being protected by another primary nexthop of the same prefix (ECMP case).
- o Type of protection provided by the alternate: link protection, node protection. In case of node protection preference, an implementation SHOULD support fallback to link protection if node protection is not available.
- o Shortest path: lowest IGP metric used to reach the destination.
- o SRLG (as defined in [RFC5286] Section 3).

### 3.2.2. Enhanced criteria

An implementation of LFA SHOULD support the following enhanced criteria:

- o Downstreamness of a neighbor : preference of a downstream path over a non downstream path SHOULD be configurable.
- o Link coloring with : include, exclude and preference based system.
- o Link Bandwidth.
- o Neighbor preference.
- o Neighbor type: link or tunnel alternate. This means that user may change preference between link alternate or tunnel alternate (link preferred over tunnel, or considered as equal).

#### 3.2.2.1. Link coloring

Link coloring is a powerful system to control the choice of alternates. Protecting interfaces are tagged with colors. Protected interfaces are configured to include some colors with a preference level, and exclude others.

Link color information SHOULD be signalled in the IGP. How signalling is done is out of scope of the document but it may be useful to reuse existing admin-groups from traffic-engineering extensions.

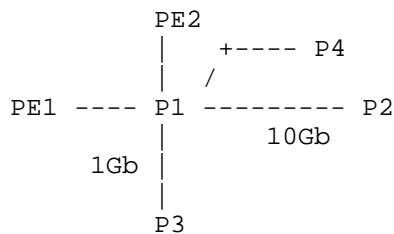


Figure 5

Example : P1 router is connected to three P routers and two PEs.

P1 is configured to protect the P1-P4 link. We assume that given the topology, all neighbors are candidate LFA. We would like to enforce a policy in the network where only a core router may protect against the failure of a core link, and where high capacity links are preferred.

In this example, we can use the proposed link coloring by:

- o Marking PEs links with color RED
- o Marking 10Gb CORE link with color BLUE
- o Marking 1Gb CORE link with color YELLOW
- o Configured the protected interface P1->P4 with :
  - \* Include BLUE, preference 200
  - \* Include YELLOW, preference 100
  - \* Exclude RED

Using this, PE links will never be used to protect against P1-P4 link failure and 10Gb link will be preferred.

The main advantage of this solution is that it can easily be duplicated on other interfaces and other nodes without change. A Service Provider has only to define the color system (associate color with a significance), as it is done already for TE affinities or BGP communities.

An implementation of link coloring:

- o SHOULD support multiple include and exclude colors on a single protected interface.

- o SHOULD provide a level of preference between included colors.
- o SHOULD support multiple colors configuration on a single protecting interface.

#### 3.2.2.2. Bandwidth

As mentionned in previous sections, not taking into account bandwidth of an alternate could lead to congestion during FRR activation. We propose to base the bandwidth criteria on the link speed information for the following reason :

- o if a router S has a set of X destinations primarily forwarded to N, using per prefix LFA may lead to have a subset of X protected by a neighbor N1, another subset by N2, another subset by Nx ...
- o S is not aware about traffic flows to each destination and is not able to evaluate how much traffic will be sent to N1,N2, ... Nx in case of FRR activation.

Based on this, it is not useful to gather available bandwidth on alternate paths, as the router does not know how much bandwidth it requires for protection. The proposed link speed approach provides a good approximation with a small cost as information is easily available.

The bandwidth criteria of the policy framework SHOULD work in two ways :

- o PRUNE : exclude a LFA if link speed to reach it is lower than the link speed of the primary nexthop interface.
- o PREFER : prefer a LFA based on his bandwidth to reach it compared to the link speed of the primary nexthop interface.

#### 3.2.2.3. Neighbor preference

Rather than tagging interface on each node (using link color) to identify neighbor node type (as example), it would be helpful if routers could be identified in the IGP. This would permit a grouped processing on multiple nodes. Some existing IGP extension like SUB-TLV 1 of TLV 135 may be useful for this purpose. As an implementation must be able to exclude some specific neighbors (see mandatory criterions), an implementation :

- o SHOULD be able to give a preference to specific neighbor.

- o SHOULD be able to give a preference to a group of neighbor.
- o SHOULD be able to exclude a group of neighbor.

A specific neighbor may be identified by its interface or IP address and group of neighbors may be identified by a marker like SUB-TLV1 in TLV135. As multiple prefixes may be present in TLVs 135, an heuristic is required to choose the appropriate one that will identify the neighbor and will transport the tag associated with the neighbor preference.

We propose the following algorithm to select the prefix :

1. Select the prefix in TLV#135 that is equal to TLV#134 value (Router ID) and prefix length is 32.
2. Select the prefix in TLV#135 that is equal to TLV#132 value (IP Addresses) and prefix length is 32, it must be noted that TLV#132 may transport multiple addresses and so multiple matches may happen.
3. If multiple prefixes are matching TLV#132 values, choose the highest one.

Consider the following network:

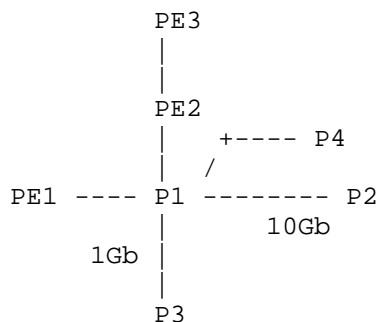


Figure 6

In the example above, each node is configured with a specific tag flooded through the IGP.

- o PE1,PE3: 200 (non candidate).

- o PE2: 100 (edge/core).
- o P1,P2,P3: 50 (core).

A simple policy could be configured on P1 to choose the best alternate for P1->P4 based on router function/role as follows :

- o criteria 1 -> neighbor preference: exclude tag 100 and 200.
- o criteria 2 -> bandwidth.

#### 3.2.2.4. Link vs remote alternate

In addition to LFA, tunnels (IP, LDP or RSVP-TE) to distant routers may be used to complement LFA coverage (tunnel tail used as virtual neighbor). When a router has multiple alternate candidates for a specific destination, it may have connected alternates (link alternates) and remote alternates reachable via a tunnel. Link alternates may not always provide an optimal routing path and it may be preferable to select a remote alternate over a link alternate. The usage of tunnels to extend LFA coverage is described in [I-D.ietf-rtgwg-remote-lfa] and [I-D.litkowski-rtgwg-lfa-rsvpte-cooperation].

In figure 1, there is no core alternate for R8 to reach PEs located behind R6, so R8 is using PE2 as alternate, which may generate congestion when FRR is activated. Instead, we could have a remote core alternate for R8 to protect PEs destinations. For example, a tunnel from R8 to R3 would ensure a LFA protection without any impact.

There is a requirement to be able to compare remote alternates (reachable through a tunnel) to link alternates (a remote alternate may provide a better protection than a link alternate based on service provider's criteria). Policy will associate a preference to each alternate whatever their type (link or remote) and will elect the best one.

## 4. Operational aspects

### 4.1. ISIS overload bit on LFA computing node

In [RFC5286], Section 3.5, the setting of the overload bit condition in LFA computation is only taken into account for the case where a neighbor has the overload bit set.

In addition to RFC 5286 inequality 1 Loop-Free Criterion



(Distance\_opt(N, D) < Distance\_opt(N, S) + Distance\_opt(S, D)), the IS-IS overload bit of the LFA calculating neighbor (S) SHOULD be taken into account. Indeed, if it has the overload bit set, no neighbor will loop back to traffic to itself.

#### 4.2. Manual triggering of FRR

Service providers often use using manual link shutdown (using router CLI) to perform some network changes/tests. Especially testing or troubleshooting FRR requires to perform the manual shutdown on the remote end of the link as generally a local shutdown would not trigger FRR. To enhance such situation, an implementation SHOULD support triggering/activating LFA Fast Reroute for a given link when a manual shutdown is done.

#### 4.3. Required local information

LFA introduction requires some enhancement in standard routing information provided by implementations. Moreover, due to the non 100% coverage, coverage informations is also required.

Hence an implementation :

- o MUST be able to display, for every prefixes, the primary nexthop as well as the alternate nexthop information.
- o MUST provide coverage information per activation domain of LFA (area, level, topology, instance, virtual router, address family ...).
- o MUST provide number of protected prefixes as well as non protected prefixes globally.
- o SHOULD provide number of protected prefixes as well as non protected prefixes per link.
- o MAY provide number of protected prefixes as well as non protected prefixes per priority if implementation supports prefix-priority insertion in RIB/FIB.
- o SHOULD provide a reason for choosing an alternate (policy and criteria) and for excluding an alternate.
- o SHOULD provide the list of non protected prefixes and the reason why they are not protected (no protection required or no alternate available).

#### 4.4. Coverage monitoring

It is pretty easy to evaluate the coverage of a network in a nominal situation, but topology changes may change the coverage. In some situations, the network may no longer be able to provide the required level of protection. Hence, it becomes very important for service providers to get alerted about changes of coverage.

An implementation SHOULD :

- o provide an alert system if total coverage (for a node) is below a defined threshold or comes back to a normal situation.
- o provide an alert system if coverage of a specific link is below a defined threshold or comes back to a normal situation.

An implementation MAY :

- o provide an alert system if a specific destination is not protected anymore or when protection comes back up for this destination

Although the procedures for providing alerts are beyond the scope of this document, we recommend that implementations consider standard and well used mechanisms like syslog or SNMP traps.

#### 5. Security Considerations

This document does not introduce any change in security consideration compared to [RFC5286].

#### 6. Contributors

Significant contributions were made by Pierre Francois, Hannes Gredler and Mustapha Aissaoui which the authors would like to acknowledge.

#### 7. Acknowledgements

#### 8. IANA Considerations

This document has no action for IANA.

#### 9. References

### 9.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC5286] Atlas, A. and A. Zinin, "Basic Specification for IP Fast Reroute: Loop-Free Alternates", RFC 5286, September 2008.

### 9.2. Informative References

- [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-01 (work in progress), December 2012.
- [I-D.litkowski-rtgwg-lfa-rsvpte-cooperation]  
Litkowski, S., Decraene, B., Filsfils, C., and K. Raza, "Interactions between LFA and RSVP-TE", draft-litkowski-rtgwg-lfa-rsvpte-cooperation-01 (work in progress), February 2013.
- [RFC3630] Katz, D., Kompella, K., and D. Yeung, "Traffic Engineering (TE) Extensions to OSPF Version 2", RFC 3630, September 2003.
- [RFC3906] Shen, N. and H. Smit, "Calculating Interior Gateway Protocol (IGP) Routes Over Traffic Engineering Tunnels", RFC 3906, October 2004.
- [RFC4090] Pan, P., Swallow, G., and A. Atlas, "Fast Reroute Extensions to RSVP-TE for LSP Tunnels", RFC 4090, May 2005.
- [RFC5305] Li, T. and H. Smit, "IS-IS Extensions for Traffic Engineering", RFC 5305, October 2008.
- [RFC5714] Shand, M. and S. Bryant, "IP Fast Reroute Framework", RFC 5714, January 2010.
- [RFC5715] Shand, M. and S. Bryant, "A Framework for Loop-Free Convergence", RFC 5715, January 2010.
- [RFC6571] Filsfils, C., Francois, P., Shand, M., Decraene, B., Uttaro, J., Leymann, N., and M. Horneffer, "Loop-Free Alternate (LFA) Applicability in Service Provider (SP) Networks", RFC 6571, June 2012.

Authors' Addresses

Stephane Litkowski  
Orange

Email: [stephane.litkowski@orange.com](mailto:stephane.litkowski@orange.com)

Bruno Decraene  
Orange

Email: [bruno.decraene@orange.com](mailto:bruno.decraene@orange.com)

Clarence Filsfils  
Cisco Systems

Email: [cfilsfil@cisco.com](mailto:cfilsfil@cisco.com)

Kamran Raza  
Cisco Systems

Email: [skraza@cisco.com](mailto:skraza@cisco.com)



Network Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: January 22, 2015

J. Jeganathan  
H. Gredler  
Juniper Networks  
B. Decraene  
France Telecom - Orange  
July 21, 2014

2547 egress PE Fast Failure Protection  
draft-minto-2547-egress-node-fast-protection-03

Abstract

This document specifies a fast-protection mechanism for protecting [RFC2547] based VPN service against egress node failure. This mechanism enables local repair to be performed immediately upon a egress node failure. In particular, the routers upstream to egress node could redirect VPN traffic to a protector (a new role) to repair in the order of tens of milliseconds, achieving fast protection that is comparable to MPLS fast reroute.

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 22, 2015.

Copyright Notice

Copyright (c) 2014 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect

to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

## Table of Contents

1. Introduction . . . . .	2
2. Specification of Requirements . . . . .	3
3. Terminology . . . . .	3
4. Reference topology . . . . .	3
5. Theory of Operation . . . . .	5
5.1. Protector and Protection Models . . . . .	6
5.1.1. Co-located protector . . . . .	6
5.1.2. Centralized protector . . . . .	7
5.1.3. Hybrid protector . . . . .	7
5.2. Context Identifier and VPN prefixes. . . . .	7
5.3. MPLS egress Fast reroute . . . . .	8
5.3.1. RSVP . . . . .	8
5.3.2. LDP . . . . .	8
5.4. Forwarding State on Protector PE . . . . .	9
5.4.1. Alternate egress PE for protected prefix. . . . .	9
6. Egress node Failure . . . . .	10
7. Deployment Considerations . . . . .	10
7.1. Discussion on deployment models. . . . .	10
7.2. Simple deployment model. . . . .	11
7.3. Deployment requirements. . . . .	12
8. Security Considerations . . . . .	12
9. Acknowledgements . . . . .	12
10. References . . . . .	12
10.1. Normative References . . . . .	12
10.2. Informative References . . . . .	13
Authors' Addresses . . . . .	14

## 1. Introduction

This document specifies a fast-protection mechanism for protecting RFC 2547 based VPN against egress PE failure. The procedures in this document are relevant only when a VPN site is multi-homed to two or more PEs. This is mainly designed based on MPLS context specific label switching[RFC5331]. This fast-protection refers to the ability to provide local repair upon a failure in the order of tens of milliseconds, which is comparable to MPLS fast-reroute [RFC4090]. This fast-protection is achieved by establishing local protection as close to a failure as possible. Compared with the existing global repair mechanisms that rely on control plane convergence, these procedures could provide faster and more deterministic restoration

for VPN traffic. However, this is intended to complement the global repair mechanisms, rather than replacing them in any way.

## 2. Specification of Requirements

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119.

## 3. Terminology

Protected PE: A PE which request fast-protection for set of VPN-IP prefixes.

Protected VPN-IP prefix: A multi-homed VPN-IP prefix that required protection in event of protected node goes down.

Protector: A router which protect one or more Protected VPN-IP prefix when a Protected node goes down.

BGP nexthop: A nexthop advertised in the BGP-Update for the VPN-IP prefix by a BGP speaker.

VPN label: A label advertised by a BGP speaker for set of VPN-IP prefixes. This label could be per-VRF label or per-nexthop label or per-prefix label.

Transport LSP: A MPLS LSP setup to BGP nexthop either by LDP or RSVP.

Alternative egress PE: A PE originates VPN-IP prefix with same IP prefix of the protected VPN-IP prefix in a same VPN.

Context MPLS table: A context-specific label space FIB. This table is populated with VPN labels advertised by the protected-PE for the protected VPN-IP prefix.

Context label: A label from protector provides context for context-specific label forwarding.

Context VRF: A IP FIB with alternate nexthop per context per site.

PLR: Point of Local Repair.

## 4. Reference topology

This document refers to the following topologies to describe various roles, procedures and solution.



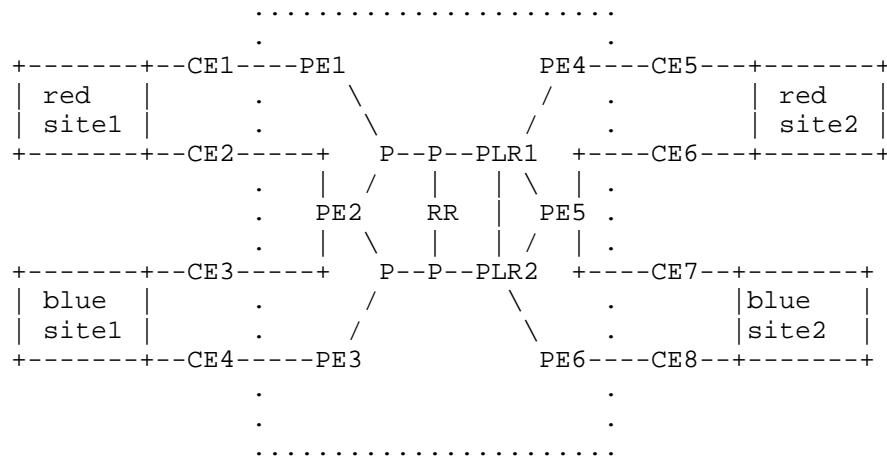


Figure 1

In Figure 1 there are two VPNs red and blue with two multi-homed sites connecting to their PEs. Assume blue VPN site2 and red VPN site2 required egress protection in case of PE5 goes down. Then PE5 is protected PE for red VPN site2 for and blue VPN site2. VPN-IP prefixes originated by PE5 associated with red site2 and blue site2 are protected VPN prefixes. The MPLS label associated with VPN-IP prefix is VPN Label. The PE4 is an alternative egress PE for red site2 and PE6 is an alternative egress PE for blue site2. The protector role could be delegated to any existing router in the network. For example PE4 could act as protector for red VPN site2 and PE6 could acts as protector for blue VPN site2. This protector model is co-located model. Alternatively, RR or any other router participates in VPN-IP control plane and not connected to VPN sites could also act as protector for both red and blue VPN site2. This model is centralized model.

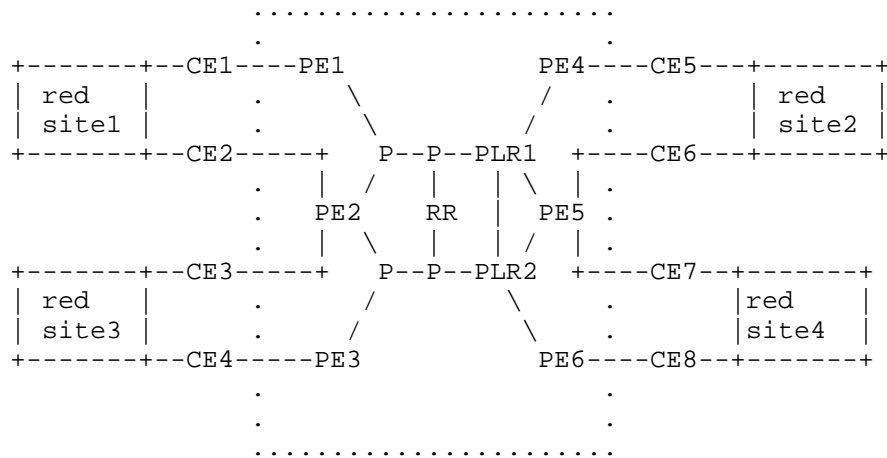


Figure 2

In Figure 2 there is a VPN red with four sites and all sites are multi homed to their PEs. Assume site2 and site4 require egress protection in case PE5 goes down. Then PE5 is the protected PE for site2 and site4. PE4 and PE6 are alternate PEs for site2 and site4 respectively. Here also the protector role could be delegated to any existing router in the network. For example PE4 could act as a protector for site2 and PE6 could act as a protector for site4. This is called the 'co-located model'. Also PE4 or PE6 could act as protector for both sites. This is called the 'hybrid model'.

The various protector models and deployment guidance are spelled-out in Section 5.1 and Section 7.

## 5. Theory of Operation

Each (egress) PE attached to a given multi-homed site originates VPN-IP route(s) associated with the destination(s) within that site. Each such route should have its own Route Distinguisher, and its own next-hop, although all these routes have the same Route Target(s). Each (ingress) PE attached to other sites within the same VPN, import these route(s) into VRF creating more than one possible path to multi-homed sites. When an egress PE goes down, all VPN traffic destined to the multi homed sites attached to the downed egress PE gets rerouted to alternate egress PE(s) attached to same multi-homed site by ingress PE(s) after it detects the egress PE down. Until ingress PE(s) reroute the VPN traffic, the traffic that used to go through the failed PE get dropped in penultimate hop router. Even though connectivity of multi-homed site is not bound to an egress PE, the VPN traffic gets dropped in the P router as a result of the

downed transport LSP that binds to that egress PE. This document specifies a mechanism that repairs VPN traffic at the point of failure (typically a P router which is penultimate hop of the transport LSP) and still keep P router unaware of the VPN information with the help of a protector. Section 5.1 explain the details. The penultimate hop router(s) of the transport LSP to egress PE(PLR) reroutes VPN traffic to protector through a bypass LSP in the event of egress PE failure. Protector forwards VPN traffic received from PLR in the bypass LSP to the alternative egress PE until the ingress PE reroute traffic to alternate egress PE.

### 5.1. Protector and Protection Models

Protector, a new role, could be delegated to a router which participates in VPN-IP control plane for VPN-IP prefixes that requires egress node protection. In a network, protector could be the alternate egress PE of a egress protected multi homed site (precisely: the egress protected VPN-IP prefixes), or any other PE or stand-alone router for egress protection.

This specification defines three types of protector:

- o co-located
- o centralized
- o hybrid

Its designation is dependent on the protector having direct links to the alternate site for a given VPN. A network MAY use either protection model or a combination depending on the requirements and actual network topology.

#### 5.1.1. Co-located protector

In this model, the protector role is delegated to the alternate egress PE for a protected VPN site. Protector is co-located with the alternate PE for the protected VPN site, and it has a direct connection to the multi-homed site that originates the protected VPN-IP prefix. In the event of an egress node failure, the protector receives traffic from the PLR, and forwards VPN traffic to the multi-homed site. In the Figure 1 co-located protector could be PE4 red VPN site2 and PE6 could be the co-located protector for blue VPN site2.

### 5.1.2. Centralized protector

In this model, the protector serves as a centralized protector and does not have a direct connection to egress protected multi-homed sites. This model can be played by existing PEs or a dedicated protector. In the event of an egress PE failure, protector MUST forwards the traffic to an alternate egress PE with the VPN label advertised by the alternate egress PE for the VPN-IP prefix, which in turn forwards the traffic to the multi-homed site. In the Figure 1RR could act as protector for red's site2 and blue's site2 or PE6 could act as protector for red's site2 and PE4 acts as protector for VPN blue's site2. This is centralized protector model (A PE protecting VPN(s) and not connected to any protected VPN site).

### 5.1.3. Hybrid protector

In this model, the protector is co-located for some egress protected sites and centralized for other egress protected sites. These protected egress sites could be in the same VPN or in different VPN. In the Figure 2either PE4 or PE6 could act as hybrid protector. Figure 1PE6 could act as hybrid protector for VPNs red site2 and blue site2.

## 5.2. Context Identifier and VPN prefixes.

Context-identifier is an IP address that is either globally unique or unique in the private address space of the routing domain. A context-identifier is shared between protected PE and protector(s) and It provides forwarding context for protected PE and protector. In the Protected PE each VPN-IP prefix is assigned to a context-identifier. The granularity of a context identifier is {Egress PE, VPN-IP prefix} tuple. However, a given context identifier MAY be assigned to one or multiple VPN-IP prefixes. A given context identifier MUST NOT be used by more than one protected PE and should never used for setting up BGP sessions or any control plane sessions.

The egress PE that requires protection for a VPN-IP prefix MUST set context-identifier as the BGP nexthop for VPN-IPv4 and IPv4-Mapped context-identifier for VPN-IPv6. This context-identifier as nexthop indicates to the protector that a particular VPN-IP prefix need protection. For example in Figure 1 PE5 (protected PE) advertises VPN-IP prefixes with context-identifier as BGP nexthop. The context identifier MUST also be advertised in the IGP and in LDP if LDP is used to establish transport LSP.

Possible context identifier assignments are

- o Unique context-identifier for all VPN-IP prefixes, both VPN-IPv4 and VPN-IPv6. Here all the VRFs on a PE share same context-identifier.
- o Unique context-identifier per address family. Here all the VRFs on the PE share the same context-identifier for given address family.
- o Unique context-identifier per site for all VPN-IP prefixes, both VPN-IPv4 and VPN-IPv6. Here every VRFs has different context-identifier.
- o Unique context-identifier per site per address family. Here every VRFs has different context-identifiers for a given address family.
- o Unique context-identifier per CE address (nexthop). Here every CE in a VRF has a different context-identifier.
- o Unique context identifier for each VPN-IP prefix. Here every VPN-IP has a different context-identifier.

The first one is coarsest granularity of a context identifier and the last one is finest granularity of a context identifier. While all of the above options are possible in principle, their practical usage is likely to vary, as not all of them may be of practical usage.

### 5.3. MPLS egress Fast reroute

A Protector should be able to receive the traffic from PLR in the event of an egress PE failure with forwarding context that enables protector to repair VPN traffic.

#### 5.3.1. RSVP

If RSVP LSP is used for transport then protector and primary MUST follow procedures specified in [rsvp-egress-frr]. The context-identifier will be used as destination address of the protected LSP and the protector will be backup egress node of the protected LSP. PLR MUST follow [rsvp-egress-frr] procedure if alias method is used.

#### 5.3.2. LDP

If LDP is used for transport then LDP FEC MUST be the context identifier. The protector for the context identifier and context label could be learned through IGP which is beyond the scope of the document. The node protecting bypass path could be computed either by remote LFA or LFA for the context identifier to protector. This

bypass LSP to protector with context label, learned through IGP,  
provide forwarding context to protector.

#### 5.4. Forwarding State on Protector PE

A Protector MUST maintain multiple forwarding tables. Protector maintains the forwarding state in context-specific label space on per context-Identifier basis. It also maintains context specific IP forwarding table, context VRF, populated by extracting IP from VPN-IP prefix with nexthop to alternate egress PE for egress protected prefixes. In particular, the protector MUST learn VPN labels associated with VPN-IP prefixes by participating in VPN routing and MUST keep routes and labels associated with VPN(s) site(s) that required protection. For each VPN label with an associated context-identifier, the protector MUST map the context identifier to a context-specific label space [RFC5331], and programs the VPN label in that label space into its forwarding plane. The VPN label in the context-specific label space identifies the IP forwarding table, that need to be looked up to send it alternate egress PE.

The protector MAY maintain only VPN-IP prefix originated with-in the multi-homed site for given {egress PE, VPN} tuple. These VPN labels in context table and context VRF will not be used in forwarding after the ingress PE reroutes the traffic to the new best PE. Protector MUST delete VPN label and the VPN context table after ingress reroute the traffic. This SHOULD be achieved with a timer. This timer default value is 180 seconds, allowing to be able to sustain large reroute events.

Note that if the protected PE does advertise a distinct label per VPN-IP prefix, as an optimization, the protector PE does not need to create an context VRF as the MPLS lookup on the VPN label is enough to identify the outgoing PE and label.

##### 5.4.1. Alternate egress PE for protected prefix.

Any route with BGP nexthop which has the following properties

- Exact matching route-target set

- Exact matching Prefix part (excluding the RD)

will be eligible as alternate egress PE for prefix.

## 6. Egress node Failure

This section summarizes the procedure for egress protection as described in the above section for completeness. A Egress PE, Protector, PLR follows the methods described in Section 5.3. The protector programs forwarding state in such a way that packets received on the bypass LSP will be forwarded based on VPN label in the context table, and prefix lookup in context VPN table. The context table is identified by the UHP label of the bypass LSP, i.e. the context identifier.

When the penultimate Hop router receives a VPN packet from the MPLS network, if the egress PE is down, the PLR tunnels the packet through the bypass LSP to the protector. The protector PE identifies the forwarding context of the egress PE based on the top label of the packet which is the UHP label of the bypass LSP. The protector further performs a second label lookup in the protected PE's context label space followed by layer-3 lookup in the VPN context table. These UHP label, context table label and layer-3 lookup results in forwarding the packet to the site or send it to alternate egress PE based on protector model.

For example in Figure 1 RR acts as Protector and PE5 requires protection for red, blue site2 VPN-IP prefixes. As red site2 and blue site2 VPN-IP prefixes are advertised with context-identifier, the protector sets up the forwarding table for VPN-IP prefixes from site2 with alternative egress PE as nexthop. When PLR detects PE5 failure it sends all the traffic that PLR used to forward directly to PE5 to protector through bypass LSP. In the protector the top label identifies the context specific table. The VPN label in the context table identifies the VPN layer-3 forwarding table which contains site2 VPN-IP prefixes with alternate PE as nexthop. A Layer-3 lookup gives mpls path to alternate egress PE and protector will forward the packet to alternate egress PE and reach to the site2.

## 7. Deployment Considerations

### 7.1. Discussion on deployment models.

As the context-identifiers are advertised in the IGP, they introduce additional states in the network and the forwarding tables. As such, in general, it's desirable to keep their number limited. The granularity of context-identifier is also related to the protector model used. If a centralized or hybrid protector model is used, a unique context-identifier per egress PE is enough. If a co-located protector model is used, a context identifier per VPN or per CE may be needed.

The centralized protector model, using a single context identifier per protected PE, limits the number of additional states in the network (IGP, forwarding tables) but may add extra latency during the protection time. It also minimizes the configuration effort as zero configuration is achievable. On the contrary the co-located mode, having a more granular context identifier, will minimize the latency during the protection time at the cost of adding more states in the network. It requires more configuration as the service provider will need to define the PE pairs (protected, protector). The hybrid model is expected to offer the best trade-off as the number of IGP states in the network can be minimized by using a single context identifier per protected PE, while the additional latency can be limited by geographically distributing the protector PE in the network.

## 7.2. Simple deployment model.

We propose the following simple deployment model:

- o a single centralized Protector PE.
- o a single context-identifier per protected PE, with all VPN routes advertised with this context-identifier as BGP next-hop.

It provides the following benefits:

- o minimize the number of IGP states in the network.
- o minimize the configuration required: no per VPN configuration on the protector PE.

Regarding the IGP states, no additional states are required if the PEs use secondary loopback address as BGP nexthop for VPN-IP address family. Otherwise, one additional IP address per PE is needed. However, the number of IP addresses used as BGP next-hop for the customer traffic is not increased, hence if the routers allow the prioritization of the prefix during FIB update, there is no impact on the IGP convergence time.

Regarding the configuration required on the network:

- o The protected PE is configured once with an additional IP address which serves as a context identifier. The BGP Next-Hop of the BGP routes are set to this context-identifier.
- o The centralized protector PE does not require per VPN configuration. But it should allow setting of context-identifiers to control VPN or PE it needs to protect. This will be useful in



multiple protectors in the network and set of PEs are protected by a given protector. The configured context-identifiers in protector protects subset of sites or PEs.

If one want to limit the protection to only a subset of VPN or a subset of PE (for lower VPN-SLA reasons, FIB capacities reasons on the protector, forwarding capacity reason during the protection time, for the hybrid model), one may not set context-identifier as a nexthop to the VPN-IP routes that required protection. VPN per protected PE configuration is required if user wants to limit egress protection for subset of sites. In this case protected be should allow user to not set the context-identifier as BGP nexthop for advertised VPN-IP prefixes.

### 7.3. Deployment requirements.

This solution does not mandate any protocol extension on any router. It does not mandate any additional feature on any routers except the new protector PE. In particular, it does not mandate implementation change on ingress nor egress PE, hence could works with legacy PE. In most topology, when LDP is used, the PLR will need to support the use of a LDP LSP as a targeted LFA. This is similar to R-LFA but the ability to configure a specific LSP to reach the protector PE may be specific.

## 8. Security Considerations

The security considerations discussed in RFC 5036, RFC 5331, RFC 3209, and RFC 4090 apply to this document.

## 9. Acknowledgements

This draft is based on the ideas originally developed by JL Le Roux, Bruno Decraene and Zubair Ahmad. This document leverages work done by Yakov Rekhter and several others on LSP tail-end protection. Thanks to Nischal Sheth, Nitin Bahadur, Yimin shen, Kaliraj Vairavakkalai and Maciek Konstantynowicz for their valuable contribution.

## 10. References

### 10.1. Normative References

[RFC5331] Aggarwal, R., Rekhter, Y., and E. Rosen, "MPLS Upstream Label Assignment and Context-Specific Label Space", RFC 5331, August 2008.

- [RFC2547] Rosen, E. and Y. Rekhter, "BGP/MPLS VPNs", RFC 2547, March 1999.
- [RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, February 2006.
- [RFC5036] Andersson, L., Minei, I., and B. Thomas, "LDP Specification", RFC 5036, October 2007.
- [RFC2205] Braden, B., Zhang, L., Berson, S., Herzog, S., and S. Jamin, "Resource ReSerVation Protocol (RSVP) -- Version 1 Functional Specification", RFC 2205, September 1997.
- [RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, December 2001.
- [RFC4090] Pan, P., Swallow, G., and A. Atlas, "Fast Reroute Extensions to RSVP-TE for LSP Tunnels", RFC 4090, May 2005.
- [RFC3471] Berger, L., "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Functional Description", RFC 3471, January 2003.
- [RFC3031] Rosen, E., Viswanathan, A., and R. Callon, "Multiprotocol Label Switching Architecture", RFC 3031, January 2001.
- [LDP-UPSTREAM] Aggarwal, R. and J. Roux, "MPLS Upstream Label Assignment for LDP", draft-ietf-mpls-ldp-upstream (work in progress), 2011.
- [RSVP-NON-PHP-OOB] Ali, A., Swallow, Z., and R. Aggarwal, "Non PHP Behavior and out-of-band mapping for RSVP-TE LSPs", draft-ietf-mpls-rsvp-te-no-php-oob-mapping (work in progress), 2011.

## 10.2. Informative References

- [RFC5920] Fang, L., "Security Framework for MPLS and GMPLS Networks", RFC 5920, July 2010.
- [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.

[rsvp-egress-frr]

Jeganathan, J., Gredler, H., and Y. Shen, "IP Fast Reroute Framework", draft-minto-rsvp-lsp-egress-fast-protection-01 (work in progress), Oct 2012, <rsvp egress frr>.

Authors' Addresses

Jeyananth Minto Jeganathan  
Juniper Networks  
1194 N Mathilda Avenue  
Sunnyvale, CA 94089  
USA

Email: minto@juniper.net

Hannes Gredler  
Juniper Networks  
1194 N Mathilda Avenue  
Sunnyvale, CA 94089  
USA

Email: hannes@juniper.net

Bruno Decraene  
France Telecom - Orange  
38 rue du General Leclerc  
Issy Moulineaux cedex 9 92794  
France

Email: bruno.decraene@orange.com

Network Working Group  
Internet Draft  
Intended status: Informational  
Expires: April 2014

A. Bashandy, Ed.  
C. Filsfils  
Cisco Systems  
P. Mohapatra  
Cumulus Networks  
October 21, 2013

BGP Prefix Independent Convergence  
draft-rtgwg-bgp-pic-02.txt

## Abstract

In the network comprising thousands of iBGP peers exchanging millions of routes, many routes are reachable via more than one path. Given the large scaling targets, it is desirable to restore traffic after failure in a time period that does not depend on the number of BGP prefixes. In this document we proposed a technique by which traffic can be re-routed to ECMP or pre-calculated backup paths in a timeframe that does not depend on the number of BGP prefixes. The objective is achieved through organizing the forwarding chains in a hierarchical manner and sharing forwarding elements among the maximum possible number of routes. The proposed technique achieves prefix independent convergence while ensuring incremental deployment, complete transparency and automation, and zero management and provisioning effort

## Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

This document may contain material from IETF Documents or IETF Contributions published or made publicly available before November 10, 2008. The person(s) controlling the copyright in some of this material may not have granted the IETF Trust the right to allow modifications of such material outside the IETF Standards Process. Without obtaining an adequate license from the person(s) controlling the copyright in such materials, this document may not be modified outside the IETF Standards Process, and derivative works of it may not be created outside the IETF Standards Process, except to format it for publication as an RFC or to translate it into languages other than English.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

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."

The list of current Internet-Drafts can be accessed at  
<http://www.ietf.org/ietf/lid-abstracts.txt>

The list of Internet-Draft Shadow Directories can be accessed at  
<http://www.ietf.org/shadow.html>

This Internet-Draft will expire on September 21, 2013.

#### Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

#### Table of Contents

1. Introduction.....	3
1.1. Conventions used in this document.....	3
1.2. Terminology.....	4
2. Constructing the Shared Hierarchical Forwarding Chain.....	5
2.1. Databases.....	5
2.2. Constructing the forwarding chain from a downloaded route.....	6
2.3. Examples.....	7
2.3.1. Example 1: Forwarding Chain for iBGP ECMP.....	7
2.3.2. Example 2: Primary Backup Paths.....	9
3. Forwarding Behavior.....	10
4. Forwarding Chain Adjustment at a Failure.....	10
4.1. BGP-PIC core.....	11
4.2. BGP-PIC edge.....	12
4.2.1. Adjusting forwarding Chain in egress node failure....	12
4.2.2. Adjusting Forwarding Chain on PE-CE link Failure....	12
4.2.3. Loop Avoidance using Special Label (backup/repair label).....	13
5. Properties.....	14
6. Dependency.....	17

7. Security Considerations.....	18
8. IANA Considerations.....	18
9. Conclusions.....	18
10. References.....	18
10.1. Normative References.....	18
10.2. Informative References.....	18
11. Acknowledgments.....	19
Appendix A. Modification History.....	20
A.1.1. Changes from Version 01.....	20
A.1.2. Changes from Version 00.....	20

## 1. Introduction

As a path vector protocol, BGP is inherently slow due to the serial nature of reachability propagation. BGP speakers exchange reachability information about prefixes[2][3] and, for labeled address families, namely AFI/SAFI 1/4, 2/4, 1/128, and 2/128, an edge router assigns local labels to prefixes and associates the local label with each advertised prefix such as L3VPN [6], 6PE [7], and Softwire [5]. A BGP speaker then applies the path selection steps to choose the best path. In modern networks, it is not uncommon to have a prefix reachable via multiple edge routers. In addition to proprietary techniques, multiple techniques have been proposed to allow for more than one path for a given prefix [4][9][10], whether in the form of equal cost multipath or primary-backup. Another more common and widely deployed scenario is L3VPN with multi-homed VPN sites.

This document proposes a hierarchical and shared forwarding chain organization that allows traffic to be restored to pre-calculated alternative equal cost primary path or backup path in a time period that does not depend on the number of BGP prefixes. The technique relies on internal router behavior that is completely transparent to the operator and can be incrementally deployed and enabled with zero operator intervention.

### 1.1. Conventions used in this document

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC-2119 [1].

In this document, these words will appear with that interpretation only when in ALL CAPS. Lower case uses of these words are not to be interpreted as carrying RFC-2119 significance.

## 1.2. Terminology

This section defines the terms used in this document. For ease of use, we will use terms similar to those used by L3VPN [6]

- o BGP prefix: It is a prefix P/m (of any AFI/SAFI) that a BGP speaker has a path for.
- o IGP prefix: It is a prefix P/m (of any AFI/SAFI) that is learnt via an Interior Gateway Protocol, such as OSPF and ISIS, has a path for. The prefix may be learnt directly through the IGP redistributed from other protocol(s)
- o CE: It is an external router through which an egress PE can reach a prefix P/m.
- o Ingress PE, "iPE": It is a BGP speaker that learns about a prefix through another IBGP peer and chooses that IBGP peer as the next-hop for the prefix.
- o Path: It is the next-hop in a sequence of unique connected nodes starting from the current node and ending with the destination node or network identified by the prefix.
- o Recursive path: It is a path consisting only of the IP address of the next-hop without the outgoing interface. Subsequent lookups are needed to determine the outgoing interface.
- o Non-recursive path: It is a path consisting of the IP address of the next-hop and one outgoing interface
- o Primary path: It is a recursive or non-recursive path that can be used all the time. A prefix can have more than one primary path
- o Backup path: It is a recursive or non-recursive path that can be used only after some or all primary paths become unreachable
- o Leaf: A leaf is container data structure for a prefix or local label. Alternatively, it is the data structure that contains prefix specific information.
- o IP leaf: Is the leaf corresponding to an IPv4 or IPv6 prefix
- o Label leaf. It is the leaf corresponding to a locally allocated label such as the VPN label on an egress PE [6].

- o Pathlist: It is an array of paths used by one or more prefix to forward traffic to destination(s) covered by a IP prefix. Each path in the pathlist carries its "path-index" that identifies its position in the array of paths. A pathlist may contain a mix of primary and backup paths
- o OutLabel-Array: Each labeled prefix is associated with an OutLabel-Array. The OutLabel-Array is a list of one or more outgoing labels and/or label actions where each label or label action has 1-to-1 correspondence to a path in the pathlist. The number of entries in the OutLabel-array is identical to the number of paths in the pathlist and the ith outlabel entry is associated with the path whose path-index is "i". Label actions are: push the label, pop the label, or swap the incoming label with the outlabel. The prefix may be an IGP or BGP prefix
- o Adjacency: It is the layer 2 encapsulation leading to the layer 3 directly connected next-hop
- o Dependency: An object X is said to be a dependent or Child of object Y if Object Y cannot be deleted unless object X is no longer a dependent/child of object Y
- o Route: It is a prefix with one or more paths associated with it. Hence the minimum set of objects needed to construct a route is a leaf and a pathlist.

## 2. Constructing the Shared Hierarchical Forwarding Chain

### 2.1. Databases

The Forwarding Information Base (FIB) on a router maintains 3 basic databases

- o Pathlist-DB: A pathlist is uniquely identified by the list of paths. The Pathlist DB contains the set of all shared pathlists
- o Leaf-DB: A leaf is uniquely identified by the prefix or the label
- o Adjacency-DB: An adjacency is uniquely identified by the outgoing layer 3 interface and the IP address of the next-hop directly connected to the layer 3 interface. Adjacency DB contains the list of all adjacencies



## 2.2. Constructing the forwarding chain from a downloaded route

1. A prefix with a list of paths is downloaded to FIB from BGP. For labeled prefixes, an OutLabel-Array and possibly a local label (e.g. for a VPN [6] prefix on an egress PE) are also downloaded
2. If the prefix does not exist, construct a new IP leaf from the downloaded prefix. If a local label is allocated, construct a label leaf from the local label
3. Construct an OutLabel-Array and attach the Outlabel array to the IP and label leaf
4. The list of paths attached to the route is looked up in the pathlist-DB
5. If a pathlist PL is found
  - a. Retrieve the pathlist
6. Else
  - a. Construct a new pathlist
  - b. Insert the new pathlist in the pathlist-DB
  - c. Resolve the paths of the pathlist as follows
  - d. Recursive path:
    - i. Lookup the next-hop in the leaf-DB
    - ii. If a leaf with at least one reachable path is found, add the path to the dependency list of the leaf
    - iii. Otherwise the path remains unresolved and cannot be used for forwarding
  - e. Non-recursive path
    - i. Lookup the next-hop and outgoing interface in the adjacency-DB
    - ii. If an adjacency is found, add the path to the dependency list of adjacency
    - iii. Otherwise, create a new adjacency and add the path to its dependency list
7. Attach the leaf(s) as (a) dependent(s) of the pathlist

As a result of the above steps, a forwarding chain starting with a leaf and ending with one or more adjacency is constructed. It is noteworthy to mention that the forwarding chain is constructed without any operator intervention at all.

### 2.3. Examples

This section outlines two examples that we will use for illustration for the rest of the document. The examples use a standard multihomed VPN [6] prefix in a BGP-free core running LDP. The topology is depicted in Figure 1.

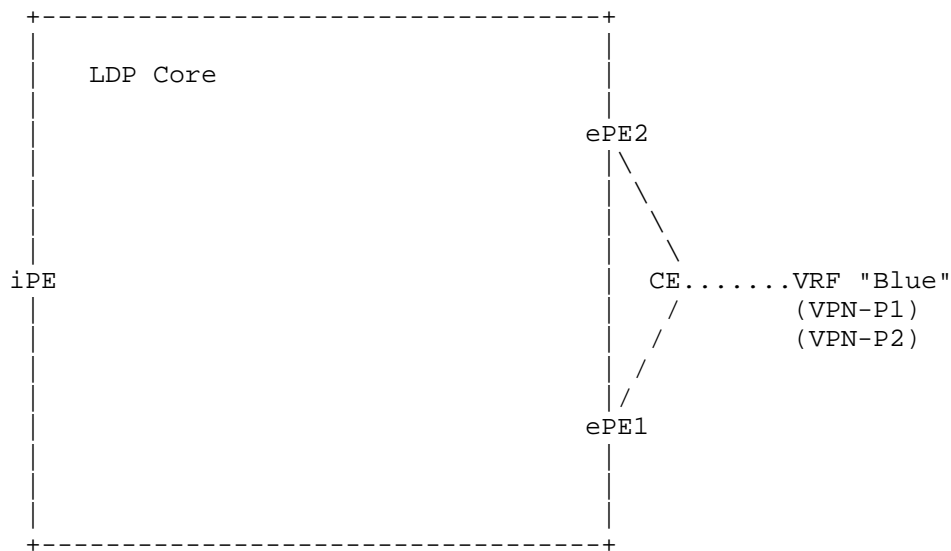


Figure 1 VPN prefix reachable via multiple PEs

The first example is an illustration of ECMP while the second example is an illustration of primary-backup paths

#### 2.3.1. Example 1: Forwarding Chain for iBGP ECMP

Consider the case of the ingress PE (iPE) in the multi-homed VPN prefixes depicted in Figure 1. Suppose the iPE receives route advertisements for the VPN prefixes VPN-P1 and VPN-P2 from two egress PEs, ePE1 and ePE2 with next-hop BGP-NH1 and BGP-NH2, respectively. Assume that ePE1 advertise the VPN labels VPN-L11 and VPN-L12 while ePE2 advertise the VPN labels VPN-L21 and VPN-L22 for VPN-P1 and VPN-P2, respectively. Suppose that BGP-NH1 and BGP-NH2 are resolved via the IGP prefixes IGP-P1 and IGP-P2, which also happen to have 2 ECMP paths with IGP-NH1 and IGP-NH2 reachable via the interfaces I1 and I2. Suppose that LDP on the downstream LSRs for IGP-P1 and IGP-P2 are assign the LDP labels LDP-L1 and LDP-L2 to

the prefixes IGP-P1 and IGP-P2. The forwarding chain on the ingress PE "iPE" for the VPN prefixes is depicted in Figure 2.

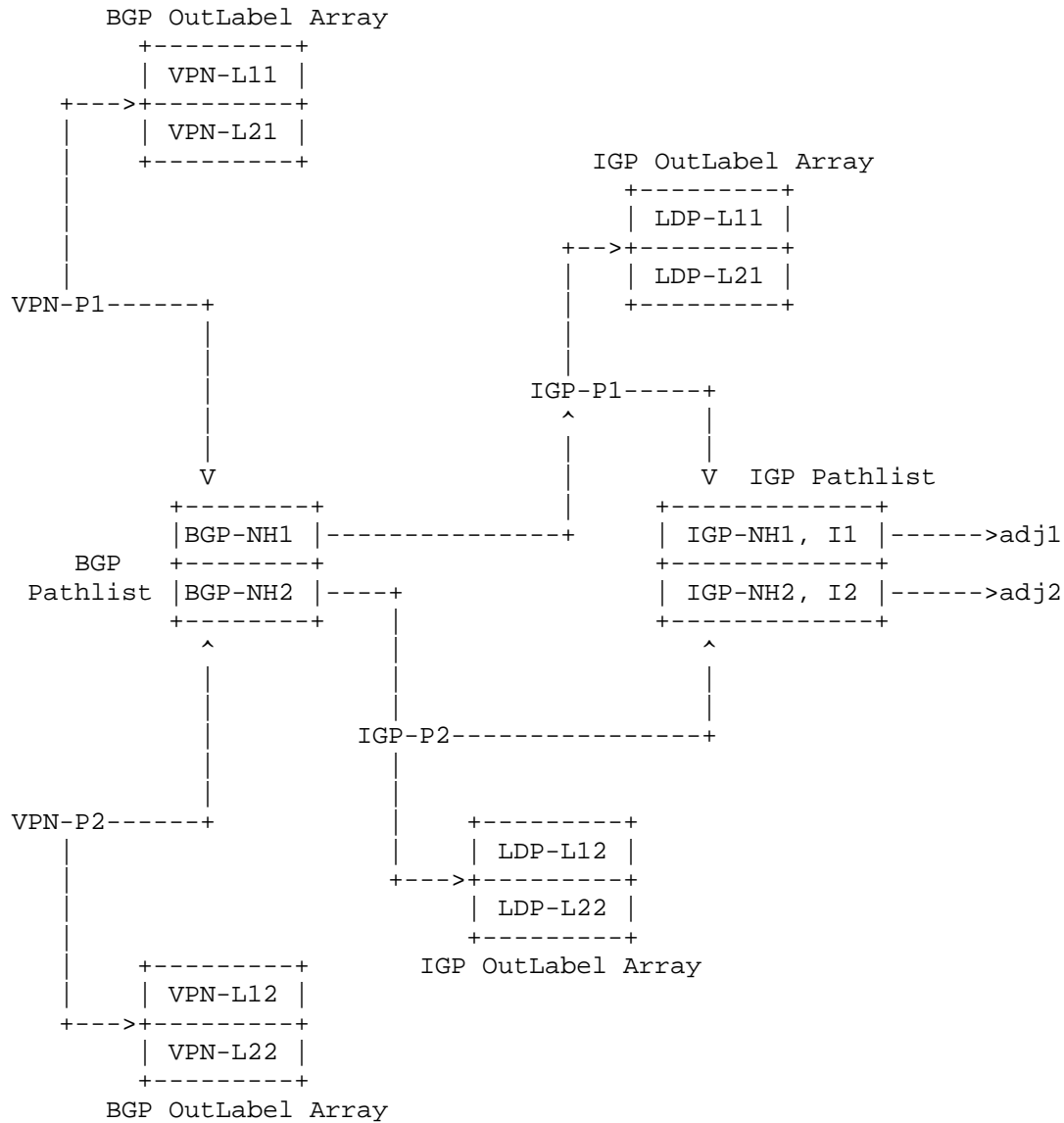


Figure 2 Forwarding Chain for VPN Prefixes with iBGP ECMP

The structure depicted in Figure 2 illustrates the two important properties discussed in this memo: sharing and hierarchy. We can

see that the both the BGP and IGP pathlists are shared among multiple BGP and IGP prefixes, respectively. At the same time, the forwarding chain objects depend on each other in a child-parent relation instead of being collapsed into a single level.

### 2.3.2. Example 2: Primary Backup Paths

Consider the egress PE ePE1 in the case of the multi-homed VPN prefixes in the BGP-free LDP core depicted in Figure 1. Suppose ePE1 determines that the primary path is the external path but the backup path is the iBGP path to the other PE ePE2 with next-hop BGP-NH2. ePE2 constructs the forwarding chain depicted in Figure 1. We are only showing a single VPN prefix for simplicity. But all prefixes that are multihomed to ePE1 and ePE2 share the BGP pathlist

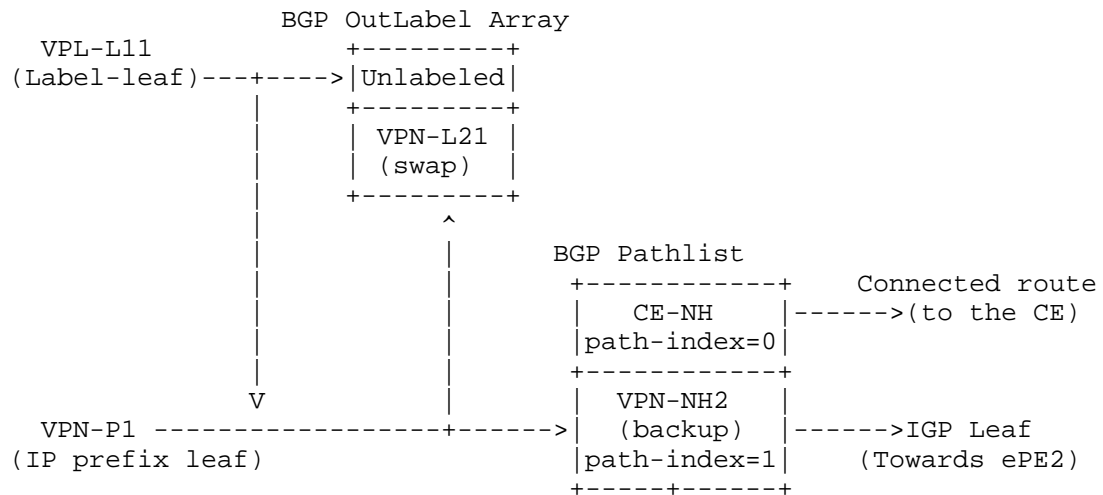


Figure 3 VPN Prefix Forwarding Chain with eiBGP paths on egress PE

The example depicted in Figure 3 differs from the example in Figure 2 in two main aspects. First as long as the primary path towards the CE (external path) is useable, it will be the only path used for forwarding while the OutLabel-Array contains both the unlabeled label (primary path) and the VPN label (backup path) advertised by the backup path ePE2. The second aspect is presence of the label leaf corresponding to the VPN prefix. This label leaf is used to match VPN traffic arriving from the core. Note that the label leaf shares the OutLabel-Array and the pathlist with the IP prefix.

### 3. Forwarding Behavior

When a packet arrives, it matches a leaf. A labeled packet matches a label leaf while an IP packet matches an IP prefix leaf. The forwarding engine walks the forwarding chain starting from the leaf until the walk terminates on an adjacency. Thus when a packet arrives, the chain is walked as follows:

1. Lookup the leaf based on the destination address or the label at the top of the packet
2. Retrieve the parent pathlist of the leaf
3. Pick the outgoing path from the list of resolved paths in the pathlist. The method by which the outgoing path is picked is beyond the scope of this document (i.e. flow-preserving hash exploiting entropy within the MPLS stack and IP header). Let the "path-index" of the outgoing path be "i".
4. If the prefix is labeled, use the "path-index" "i" to retrieve the ith label "Li" stored the ith entry in the OutLabel-Array and apply the label action of the label on the packet (e.g. for VPN label on the ingress PE, the label action is "push").
5. Move to the parent of the chosen path "i"
6. If the chosen path "i" is recursive, move to its parent prefix and go to step 2
7. If the chosen path "i" is non-recursive move to its parent adjacency
8. Encapsulate the packet in the L2 string specified by the adjacency and send the packet out.

Let's apply the above forwarding steps to the example described in Figure 1 Section 2.3.1. Suppose a packet arrives at ingress PE iPE from an external neighbor. Assume the packet matches the VPN prefix VPN-P1. While walking the forwarding chain, the forwarding engine applies hashing algorithm to choose the path and the hashing at the BGP level yields path 0 while the hashing at the IGP level yields path 1. In that case, the packet will be sent out of interface I1 with the label stack "LDP-L12,VPN-L21".

### 4. Forwarding Chain Adjustment at a Failure

The hierarchical and shared structure of the forwarding chain explained in Section 2. allows modifying a small number of forwarding chain objects to re-route traffic to a pre-calculated equal-cost or backup path without the need to modify the possibly

very large number of BGP prefixes. In this section, we go over various core and edge failure scenarios to illustrate how FIB manager can utilize the forwarding chain structure to achieve prefix independent convergence.

#### 4.1. BGP-PIC core

This section describes the adjustments to the forwarding chain when a core link or node fails but the BGP next-hop remains reachable.

There are two case: remote link failure and attached link failure. Node failures are treated as link failures.

When a remote link or node fails, IGP receives advertisement indicating a topology change so IGP re-converges to either find a new next-hop and outgoing interface or remove the path completely from the IGP prefix used to resolve BGP next-hops. IGP and LDP download the modified IGP leaves with modified outgoing labels for labeled core. FIB manager modifies the existing IGP leaf by executing the steps outlined in Section 2.2.

When a local link fails, FIB manager detects the failure almost immediately. The FIB manager marks the impacted path(s) as unuseable so that only useable paths are used to forward packets. Note that in this particular case there is actually no need even to backwalk to IGP leaves to adjust the OutLabel-Arrays because FIB can rely on the path-index stored in the useable paths in the loadinfo to pick the right label.

It is noteworthy to mention that because FIB manager modifies the forwarding chain starting from the IGP leaves only, BGP pathlists and leaves are not modified. Hence traffic restoration occurs within the time frame of IGP convergence, and, for local link failure, within the timeframe of local detection. Thus it is possible to achieve sub-50 msec convergence as described in [8] for local link failure

Let's apply the procedure to the forwarding chain depicted in Figure 2 Section 2.3.1. Suppose a remote link failure occurs and impacts the first ECMP IGP path to the remote BGP nhop. Upon IGP convergence, the IGP pathlist of the BGP nhop is updated to reflect the new topology (one path instead of two). As soon as the IGP convergence is effective for the BGP nhop entry, the new forwarding state is immediately available to all dependent BGP prefixes. The same behavior would occur if the failure was local such as an interface going down. As soon as the IGP convergence is complete for the BGP nhop IGP route, all its BGP depending routes benefit from the new path. In fact, upon local failure, if LFA protection is

enabled for the IGP route to the BGP nhop and a backup path was pre-computed and installed in the pathlist, upon the local interface failure, the LFA backup path is immediately activated (sub-50msec) and thus protection benefits all the depending BGP traffic through the hierarchical forwarding dependency between the routes.

#### 4.2. BGP-PIC edge

This section describes the adjustments to the forwarding chains as a result of edge node or edge link failure

##### 4.2.1. Adjusting forwarding Chain in egress node failure

When an edge node fails, IGP on neighboring core nodes send route updates indicating that the edge node is no longer reachable. IGP running on the iBGP peers instructs FIB to remove the IP and label leaves corresponding to the failed edge node from FIB. So FIB manager performs the following steps:

- o FIB manager deletes the IGP leaf corresponding to the failed edge node
- o FIB manager backwalks to all dependent BGP pathlists and marks that path using the deleted IGP leaf as unresolved
- o Note that there is no need to modify BGP leaves because each path in the pathlist carries its path index and hence the correct outgoing label will be picked. So for example the forwarding chain depicted in Figure 2, if the 1st path becomes unresolved, then the forwarding engine will only use the second path path for forwarding. Yet the pathindex of that single resolved path will still be 1 and hence the label VPN-L21 or VPN-L22 will be pushed

##### 4.2.2. Adjusting Forwarding Chain on PE-CE link Failure

Suppose the link between an edge router and its external peer fails. There are two scenarios (1) the edge node attached to the failed link performs next-hop self and (2) the edge node attached to the failure advertises the IP address of the failed link as the next-hop attribute to its iBGP peers.

In the first case, the rest of iBGP peers will remain unaware of the link failure and will continue to forward traffic to the edge node until the edge node attached to the failed link withdraws the BGP prefixes. If the destination prefixes are multi-homed to another iBGP peer, say ePE2, then FIB manager on the edge router detecting the link failure performs the following tasks

- o FIB manager backwalks to the BGP pathlists marks the path through the failed link to the external peer as unresolved
- o Hence traffic will be forwarded used the backup path towards ePE2
- o For labeled traffic
  - o The Outlabel-Array attached to the BGP leaves already contains an entry corresponding to the path towards ePE2.
  - o The label entry in OutLabel-Arrays corresponding to the internal path to ePE2 has swap action and the label advertised by ePE2
  - o For an arriving label packet (e.g. VPN), the top label is swapped with the label advertised by ePE2
- o For unlabeled traffic, packets are simply redirected towards ePE2

In the second case where the edge router uses the IP address of the failed link as the BGP next-hop, the edge router will still perform the previous steps. But, unlike the case of next-hop self, IGP on failed edge node informs the rest of the iBGP peers that IP address of the failed link is no longer reachable. Hence the FIB manager on iBGP peers will delete the IGP leaf corresponding to the IP prefix of the failed link. The behavior of the iBGP peers will be identical to the case of edge node failure outlined in Section 4.2.1.

It is noteworthy to mention that because the edge link failure is local to the edge router, sub-50 msec convergence can be achieved as described in [8].

Let's try to apply the case of next-hop self to the forwarding chain depicted in Figure 3. After failure of the link between ePE1 and CE, the forwarding engine will route traffic arriving from the core towards VPN-NH2 with path-index=1. A packet arriving from the core will contain the label VPN-L11 at top. The label VPN-L11 is swapped with the label VPN-L21 and the packet is forwarded towards ePE2

#### 4.2.3. Loop Avoidance using Special Label (backup/repair label)

The adjustment of the forwarding chain for edge link failure as specified in Section 4.2.2. can lead to loops in the following scenarios:

- o Unlabeled traffic when the iBGP and eBGP paths are treated as ECMP



- o Unlabeled traffic if there is an AS-wide single best path such as the case where the MED or LOCAL\_PREF [2] is used to determine the best path
- o Labeled and unlabeled traffic if the edge link failure was due to an external peer failure and the external peer is common to both edge nodes. This scenario results in edge link failure on both iBGP peers and may result in a mutual loop.

This section proposes advertising a special label as a path attribute to avoid the possibility of looping. When an edge router has an external path, whether this path is the BGP best path [2] or not [4][10][9], the edge router associates a non-transitive path attribute containing a backup/repair label. The semantics of the backup/repair label is as follows: A packet arriving with the backup/repair label at the top MUST either be sent outside the AS or dropped. Details for backup/repair label can be found in [TBD]

## 5. Properties

### 5.1 Coverage

All the possible failures are covered, whether they impact a local or remote IGP path or a local or remote BGP nhop as described in Section 4. This section provides details for each failure and how the hierarchical and shared FIB structure proposed in this document allows recovery that does not depend on number of BGP prefixes

#### 5.1.1 A remote failure on the path to a BGP nhop

Upon IGP convergence, the IGP leaf for the BGP nhop is updated upon IGP convergence and all the BGP depending routes leverage the new IGP forwarding state immediately.

This BGP resiliency property only depends on IGP convergence and is independent of the number of BGP prefixes impacted.

#### 5.1.2 A local failure on the path to a BGP nhop

Upon LFA protection, the IGP leaf for the BGP nhop is updated to use the precomputed LFA backup path and all the BGP depending routes leverage this LFA protection.

This BGP resiliency property only depends on LFA protection and is independent of the number of BGP prefixes impacted.

### 5.1.3 A remote iBGP nhop fails

Upon IGP convergence, the IGP leaf for the BGP nhop is deleted and all the depending BGP Path-Lists are updated to either use the remaining ECMP BGP best-paths or if none remains available to activate precomputed backups.

This BGP resiliency property only depends on IGP convergence and is independent of the number of BGP prefixes impacted.

### 5.1.4 A local eBGP nhop fails

Upon local link failure detection, the adjacency to the BGP nhop is deleted and all the depending BGP Path-Lists are updated to either use the remaining ECMP BGP best-paths or if none remains available to activate precomputed backups.

This BGP resiliency property only depends on local link failure detection and is independent of the number of BGP prefixes impacted.

## 5.2 Performance

When the failure is local (a local IGP nhop failure or a local eBGP nhop failure), a pre-computed and pre-installed backup is activated by a local-protection mechanism that does not depend on the number of BGP destinations impacted by the failure. Sub-50msec is thus possible even if millions of BGP routes are impacted.

When the failure is remote (a remote IGP failure not impacting the BGP nhop or a remote BGP nhop failure), an alternate path is activated upon IGP convergence. All the impacted BGP destinations benefit from a working alternate path as soon as the IGP convergence occurs for their impacted BGP nhop even if millions of BGP routes are impacted.

### 5.2.1 Perspective

The following table puts the BGP PIC benefits in perspective assuming

- o 1M impacted BGP prefixes
- o IGP convergence ~ 500 msec
- o local protection ~ 50msec
- o FIB Update per BGP destination ~ 100usec conservative,  
~ 10usec optimistic

- o BGP Convergence per BGP destination ~ 200usec conservative,  
~ 100usec optimistic

	Without PIC	With PIC
Local IGP Failure	10 to 100sec	50msec
Local BGP Failure	100 to 200sec	50msec
Remote IGP Failure	10 to 100sec	500msec
Local BGP Failure	100 to 200sec	500msec

Upon local IGP nhop failure or remote IGP nhop failure, the existing primary BGP nhop is intact and usable hence the resiliency only depends on the ability of the FIB mechanism to reflect the new path to the BGP nhop to the depending BGP destinations. Without BGP PIC, a conservative back-of-the-envelope estimation for this FIB update is 100usec per BGP destination. An optimistic estimation is 10usec per entry.

Upon local BGP nhop failure or remote BGP nhop failure, without the BGP PIC mechanism, a new BGP Best-Path needs to be recomputed and new updates need to be sent to peers. This depends on BGP processing time that will be shared between best-path computation, RIB update and peer update. A conservative back-of-the-envelope estimation for this is 200usec per BGP destination. An optimistic estimation is 100usec per entry.

### 5.3 Automated

The BGP PIC solution does not require any operator involvement. The process is entirely automated as part of the FIB implementation.

The salient points enabling this automation are:

- o Extension of the BGP Best Path to compute a backup BGP nhop [11]
- o Sharing of BGP Path-list across BGP destinations with same primary and backup BGP nhop
- o Hierarchical indirection and dependency between BGP Path-List and IGP-Path-List

### 5.4 Incremental Deployment

As soon as one router supports BGP PIC solution, it benefits from all its benefits without any requirement for other routers to support BGP PIC.

## 6. Dependency

This section describes the required functionality in the forwarding and control planes to support BGP-PIC described in this document

### 6.1 Hierarchical Hardware FIB

BGP PIC requires a hierarchical hardware FIB support: for each BGP forwarded packet, a BGP leaf is looked up, then a BGP Path-List is consulted, then an IGP Path-List then an Adjacency.

An alternative method consists in "flattening" the dependencies when programming the BGP destinations into HW FIB resulting in potentially eliminating both the BGP Path-List and IGP Path-List consultation. Such an approach decreases the number of memory lookup's per forwarding operation at the expense of HW FIB memory increase (flattening means less sharing hence duplication), loss of ECMP properties (flattening means less path-list entropy) and loss of BGP PIC properties.

### 6.2 Availability of a secondary BGP next-hop

When the primary BGP nhop fails, BGP PIC depends on the availability of a pre-computed and pre-installed secondary BGP nhop in the BGP Path-List.

The existence of a secondary next-hop is clear for the following reason: a service caring for network availability will require two disjoint network connections hence two BGP nhops.

The BGP distribution of the secondary next-hop is simple thanks to the following BGP mechanisms: Add-Path [9], BGP Best-External [4], diverse path [10], and the frequent use in VPN deployments of different VPN RD's per PE.

### 6.3 Pre-Computation of a secondary BGP nhop

[11] describes how a secondary BGP nhop can be precomputed on a per BGP destination basis.

## 7. Security Considerations

No additional security risk is introduced by using the mechanisms proposed in this document

## 8. IANA Considerations

No requirements for IANA

## 9. Conclusions

This document proposes a hierarchical and shared forwarding chain structure that allows achieving prefix independent convergence, and in the case of locally detected failures, sub-50 msec convergence. A router can construct the forwarding chains in a completely transparent manner with zero operator intervention. It supports incremental deployment.

## 10. References

### 10.1. Normative References

- [1] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [2] Rekhter, Y., Li, T., and S. Hares, "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 2006
- [3] Bates, T., Chandra, R., Katz, D., and Rekhter Y., "Multiprotocol Extensions for BGP", RFC 4760, January 2007

### 10.2. Informative References

- [4] Marques, P., Fernando, R., Chen, E., Mohapatra, P., Gredler, H., "Advertisement of the best external route in BGP", draft-ietf-idr-best-external-04.txt, April 2011.
- [5] Wu, J., Cui, Y., Metz, C., and E. Rosen, "Software Mesh Framework", RFC 5565, June 2009.
- [6] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, February 2006.
- [7] De Clercq, J., Ooms, D., Prevost, S., Le Faucheur, F., "Connecting IPv6 Islands over IPv4 MPLS Using IPv6 Provider Edge Routers (6PE)", RFC 4798, February 2007
- [8] O. Bonaventure, C. Filsfils, and P. Francois. "Achieving sub-50 milliseconds recovery upon bgp peering link failures, " IEEE/ACM Transactions on Networking, 15(5):1123-1135, 2007

- [9] D. Walton, E. Chen, A. Retana, J. Scudder, "Advertisement of Multiple Paths in BGP", draft-ietf-idr-add-paths-07.txt, June 2012
- [10] R. Raszuk, R. Fernando, K. Patel, D. McPherson, K. Kumaki, "Distribution of diverse BGP paths", draft-ietf-grow-diverse-bgp-path-dist-08.txt, July 2012
- [11] P. Mohapatra, R. Fernando, C. Filsfils, and R. Raszuk, "Fast Connectivity Restoration Using BGP Add-path", draft-pmohapat-idr-fast-conn-restore-02, October 2011

## 11. Acknowledgments

Special thanks to Neeraj Malhotra and Yuri Tsier for the valuable help

This document was prepared using 2-Word-v2.0.template.dot.

Appendix A.                      Modification History

A.1.1. Changes from Version 01

Some editorial corrections

A.1.2. Changes from Version 00

There were few editorial corrections.

Authors' Addresses

Ahmed Bashandy  
Cisco Systems  
170 West Tasman Dr, San Jose, CA 95134  
Email: bashandy@cisco.com

Clarence Filsfils  
Cisco Systems  
Brussels, Belgium  
Email: cfilsfil@cisco.com

Prodosh Mohapatra  
Cumulus Networks  
Email: pmohapat@cumulusnetworks.com





RTGWG  
Internet-Draft  
Intended status: Standards Track  
Expires: July 26, 2015

P. Thubert, Ed.  
Cisco  
P. Bellagamba  
Cisco Systems  
January 22, 2015

Available Routing Constructs  
draft-thubert-rtgwg-arc-03

Abstract

This draft introduces the concept of ARC, a two-edged routing construct that forms its own fault isolation and recovery domain. The new paradigm can be leveraged to improve the network utilization and resiliency for unicast and multicast traffic in multiple environments, and is optimized to compute short reroute paths in case of breakages.

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

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 July 26, 2015.

Copyright Notice

Copyright (c) 2015 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

## Table of Contents

1. Introduction . . . . .	3
2. Terminology . . . . .	4
3. ARC Set representations . . . . .	7
4. Lowest ARC First . . . . .	9
4.1. Init . . . . .	10
4.2. Growing Trees . . . . .	10
4.3. Being Safe . . . . .	11
4.4. Bending An ARC . . . . .	11
4.5. Orienting Links . . . . .	12
4.6. Looping or recursing . . . . .	13
5. Forwarding Along An ARC Set . . . . .	13
5.1. Control Plane Recovery . . . . .	14
5.2. Data Plane Recovery . . . . .	14
5.2.1. Label Switched ARCs . . . . .	15
5.2.2. Segment Routed ARCs . . . . .	15
5.3. Flooding . . . . .	16
6. ARC Signaling . . . . .	16
6.1. Serial ARC Representation . . . . .	16
6.2. Centralized vs. Distributed computation . . . . .	16
6.3. ARC Topology Injection . . . . .	17
6.4. ARC Operations, Administration, and Maintenance . . . . .	17
7. Other ARC Operations . . . . .	17
7.1. Node-Local vs. ARC-Wide reaction . . . . .	17
7.2. Load Balancing . . . . .	17
7.3. Shared Risk Link Group . . . . .	18
7.4. Olympic Rings . . . . .	19
7.5. Routing Hierarchies . . . . .	19
8. Manageability . . . . .	20
9. IANA Considerations . . . . .	20
10. Security Considerations . . . . .	20
11. Acknowledgments . . . . .	20
12. References . . . . .	20
12.1. Normative References . . . . .	20
12.2. Informative References . . . . .	20
Authors' Addresses . . . . .	21

## 1. Introduction

Traditional routing and forwarding uses the concept of path as the basic routing paradigm to get a packet from a source to a destination by following an ordered sequence of arrows between intermediate nodes. In this serial design, a path is broken as soon as a single arrow is, and getting around a breakage can require path re-computation, network re-convergence, and incur delays to till service is restored.

Multiple paths can be bound together for instance to form a Directed Acyclic Graph (DAG) to a destination, but that technique can be difficult to balance and cannot provide a full path redundancy even in the case of a biconnected graph. For instance, if the node that is closest to the DAG destination has only one link to that destination, then it does not have a alternate path to get to that destination.

It is also possible to compute an alternate routing topology for fast rerouting to a given destination, in which case some signaling, tagging or labeling can be put in place to indicate whether a packet follows the normal path or was rerouted over an alternate topology. Once a packet is rerouted, it is bound to the alternate topology so only one breakage can be handled with loop-free guarantees in most practical situations.

This draft introduces the concept of an Available Routing Construct (ARC) as a routing construct made of a bidirectional sequence of nodes and links with 2 outgoing edges, so that, upon a single breakage, each lively node in along ARC can still reach one of the outgoing edges.

The routing graph to reach a certain destination is expressed as a cascade of ARCs, each ARC providing its own independent domain of fault isolation and recovery. Unicast traffic may enter an ARC via any node but it may only leave the ARC through one of its two edges. One node along the ARC is designated as the Cursor. In normal unicast operations, the traffic inside an ARC flows away from the Cursor towards an edge. Upon a failure, packets may bounce on the breakage point and flow the other way along the ARC to take the other exit.

As a result an ARC is resilient to any single failure, and the recovery can be driven either from the data plane or the control plane. A second failure occurring within a same ARC will isolate an ARC segment. This can be further corrected from the control plane by reversing all the incoming Edges in a process that might recurse till

an exit is found. When ARC reversal is applied, an ARC topology is resilient to some cases of Shared Risk Link Group (SRLG) failures.

Properties of the Maximally Redundant Tree (MRT) and ARC are compared in [I-D.thubert-rtgwg-arc-vs-mrt]. The study shows that the reroute path that ARC derives is generally shorter than the alternate path that MRT computes. This property is largely due to the concept of cursor that delineates the shortest path on both sides of an ARC. Once a rerouted packet passes the cursor of the ARC in which it is rerouted, it should not cross a cursor again unless there is a second breakage later. It results that the packet follows the shortest path for the rest of the way, staying on the right side of each downstream ARC, when MRT would be following all subsequent eyes in the same direction.

This draft presents the concept and provides an intuition of how ARCs can simplify the operation and improve the network utilization and resiliency for all sorts of traffic in multiple environments, but defers to further documents to elaborate on the algorithms and optimizations in the different application domains. For instance, ARCs can also be used in datacenters for the purpose of fast-reroute, or within a service provider network to simplify load balancing operations or leverage optimally the ring topologies [RFC5921].

## 2. Terminology

The definition of the constituent parts of the "OAM" term is found in [RFC6291].

The draft uses the following terminology:

**ARC:** Available Routing Construct. An ARC is a loopless ordered set of nodes and links whereby traffic may enter via any node in the ARC but may only leave the ARC through either one of the ARC edges.

**Comb:** An ARC generalization: a Comb is a n-edged loopless set of nodes and links with  $n \geq 2$ ; traffic may enter via any node in the Comb but may only exit the Comb through one of its n edges. A Comb comes with a walk operation that enables to attempt to exit via every edge and to discover when all have been tried.

**Cursor:** A virtual point along an ARC that can be located on a node or on a link between 2 nodes. In normal operations, the traffic along the ARC flows away from its Cursor. If the Cursor is a node, then traffic can be distributed on both sides. The Cursor may be moved to change the way traffic is load balanced along an

ARC. It may also be placed at the location of a failure to direct traffic away from that point.

ARC Node: A Node that belongs to an ARC.

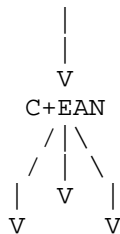
Edge ARC Node: An ARC Node at an edge of its ARC. An Edge ARC Node is a node via which traffic can exit the ARC.

Edge Link: A directed link outgoing from an Edge ARC Node. Traffic can only exit from an ARC via an Edge Link. An Edge Link does not accept traffic into an ARC.

Intermediate ARC Node: A node that is not at an edge of an ARC. A Intermediate ARC Node node that can receive traffic and forward traffic between its adjacent nodes.

Intermediate Link: A link between two Intermediate ARC Nodes. An Intermediate Link is reversible, meaning that traffic is allowed in both directions though an individual packet is constrained in the way its direction is reversed. For stable links such as wired links, the typical constraint is that the direction of a packet may be reversed at most once along a given ARC.

Collapsed ARC: An ARC that is formed of a single node. This node is altogether the Cursor and both Edge Nodes. This implies that the node has at least 2 outgoing links to 2 different Safe Nodes.



E: Edge ARC Node	-	collapsed in a single node
C: Cursor	-	

Figure 1: Collapsed ARC

Infrastructure ARC: An ARC that is formed of more than one node, which also means that the Edge Nodes are different nodes.

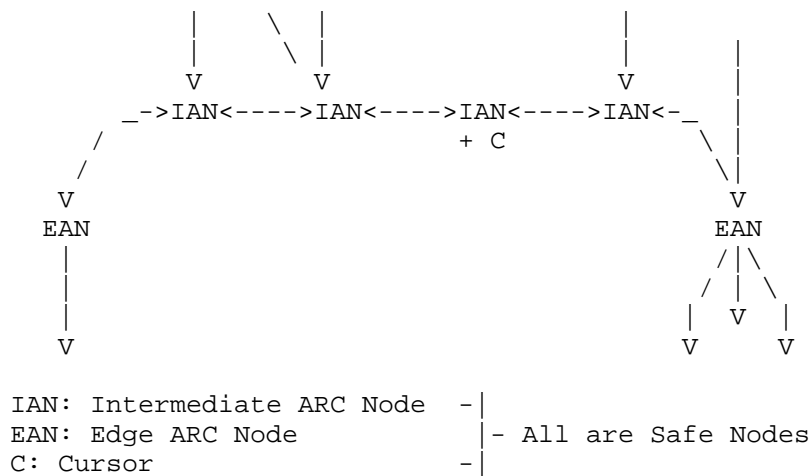


Figure 2: Infrastructure ARC

DAG: Directed Acyclic Graph.

ARC Set (or Cascade): A DAG with ARCs as vertices. In the DAG, an edge between ARC A and ARC B corresponds to a link from an Edge ARC Node in ARC A and an arbitrary ARC Node in ARC B. Note that by definition, an ARC has at least 2 outgoing Edge Links, one per Edge Node, and maybe more if an Edge Node has multiple outgoing Edge Links. All vertices in the DAG have 2 forwarding solutions, even the ARC closest to the destination.

Omega: the abstract destination (== root) of an ARC Set. Omega is also referred as a complex destination in that it typically comprises more than one node and/or more than one link on a node. if Omega has a single node, then the plural interfaces on that node are considered as as many virtual node for the sake of the ARC computation algorithm.

ARC Height: An arbitrary distance from Omega that is associated to an ARC. The Height of an ARC must be more than the Height of any of the ARCs it terminates into. The order of ARC formation by a given algorithm can be used as a Height whereby an ARC is always strictly higher or lower than another.

Buttressing ARC: A split ARC that is merged into another ARC at one edge. An ARC and one or more Buttressing ARCs form a Comb construct that is resilient to additional breakages. A Buttressing ARC may be applied to an ARC or a Comb iff traffic

outgoing the Buttreassing ARC Edge always reaches in an ARC that is lower than this ARC, or Omega.

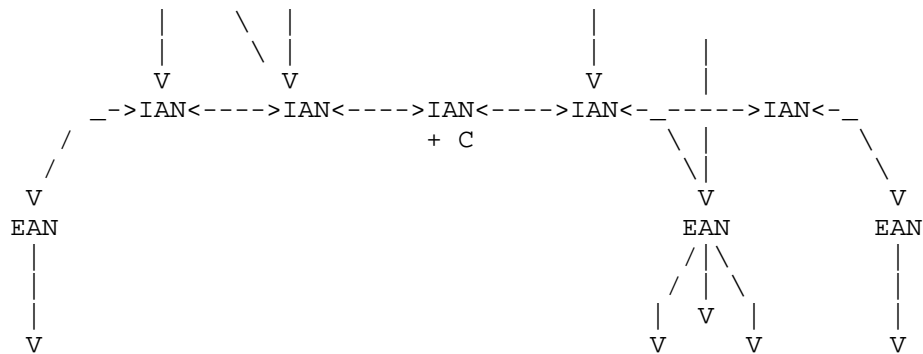


Figure 3: Comb with Buttreassing ARC

**Safe Node:** A node is Safe if there is no single point of failure - apart from the node itself - on its way to Omega. From this definition, a node is Safe if it has at least two non-congruent paths to two different other Safe Nodes. It results that a Safe node that is not Omega has at least two completely disjunct paths to Omega. When an ARC has been successfully constructed, all its nodes become safe with respect to the Omega for which the ARC was constructed. By extension for a collapsed path Omega is deemed to be Safe, that is any node that pertains in Omega is a Safe Node.

**?-S:** A node N is deemed dependent on a node S or S-dependent (denoted as ?-S) if S is the last single point of failure along N's shortest path to Omega.

### 3. ARC Set representations

An ARC Set can be represented in a number of fashions:

Graph View:

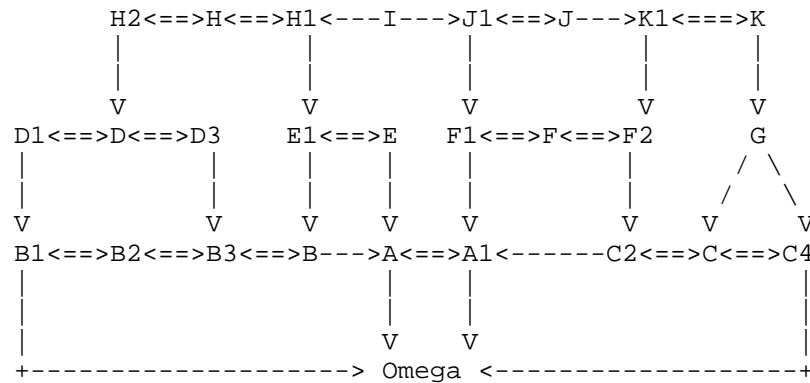


Figure 4: Routing Graph View

This representation is similar to a classical routing graph with the peculiarity that some Links are marked reversible. An ARC is represented as a sequence of reversible links. The node that holds the Cursor is also indicated somehow.

ARC View:

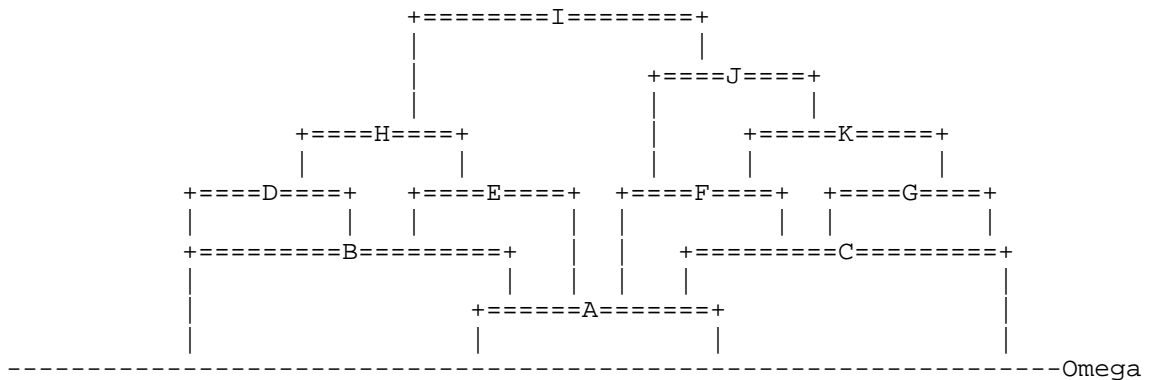


Figure 5: ARC Representation

This ARC representation abstracts a whole ARC as a single vertex. An ARC ends in one or more other ARCs, but it has to be noted that even if both edges of an ARC end in a same other ARC, it ends in



fact in 2 different nodes, or Omega. This is turn can be represented as a DAG as described in Paragraph 3.

Collapsed DAG view:

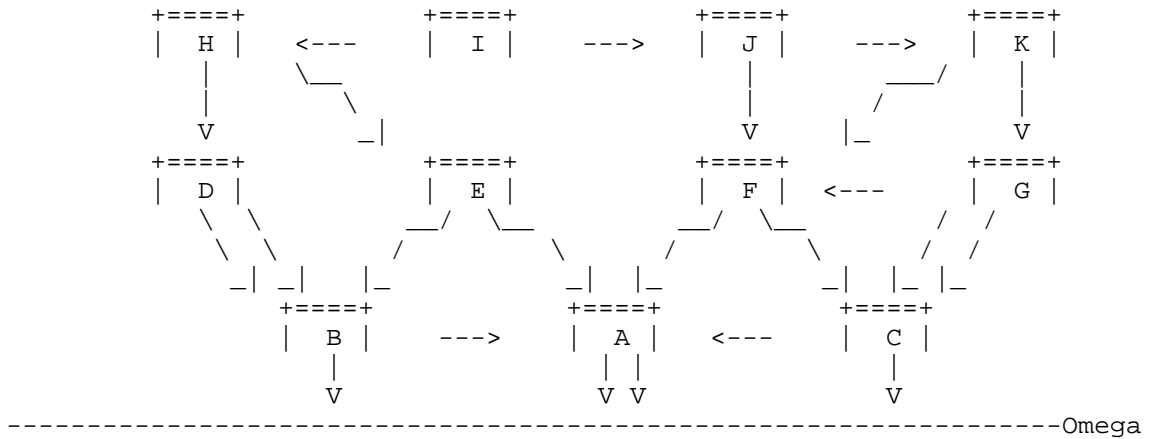


Figure 6: ARC DAG

In the DAG representation, an ARC is abstracted as a vertex and links between ARCs are shown as directed edges. This way, the reversible links are omitted and the graph is simplified. It can be noted that even the vertex closest to Omega has 2 non-congruent forwarding solutions, that is Heir Links to Omega.

#### 4. Lowest ARC First

The open Lowest ARC First(oLAF) algorithm is presented below in such a way as to help the reader figure how an ARC Set can be obtained but not in a computer-optimized fashion that is left to be determined. oLAF is based on Dijkstra's algorithm for Shortest Path First (SPF) computation, and is designed in such a fashion that the reverse SPF tree towards a destination is conserved and preferred for forwarding along the resulting ARC Set.

We make the computation on behalf of Omega, that is an abstraction, but could represent the node or the set of nodes that we want to reach with an ARC Set. If Omega is instantiated as an actual destination node, then that node may be a fine location for an ARC Computing Engine.

#### 4.1. Init

So we start with an proverbial Initial Set of Nodes that are interconnected by Links, and Omega that is the destination that we want to reach with an ARC Set.

If there is no Heir, we're done. If there is a single Heir then the graph is mono-connected, so we restart the computation taking that Heir off the Set of Nodes and making it Omega.

Else, if Omega is a single Node, or if Omega is composed of multiple nodes but we are willing to accept that both ends of an ARC terminate in a same node in Omega, then we create virtual Omega nodes, a minimum of two and at most one per Heir, and we make them the new Omega. Note: we need at least two destinations because both ends of an ARC cannot terminate in a same node.

Now we can start building an ARC Set towards the resulting Omega.

In this process, we create so-called Dependent Sets of nodes, each owned by a Safe Node  $S$ ,  $DSet(S)$ .  $DSet(S)$  contains nodes that are not determined to be Safe at the current stage of the computation and for which  $S$ , the owner Safe Node, is the last single point of failure on the shortest path tree to Omega. It results that a given node can be at most in one  $DSet$ , and that a Safe Node belongs to its own  $DSet$ .

For each node  $S$  in Omega we create a  $DSet(S)$  in which we place  $S$ .

#### 4.2. Growing Trees

And then the process goes like this:

We select the node in the Set of Nodes that is closest to Omega using the cost towards Omega as if we were building a traditional reverse SPF tree and we place the selected node in the same Dependent Set as its parent in the reverse SPF tree. Note that for a Heir, the parent might be a real node in Omega, or a virtual Omega node.

If we kept it at that, we would be building subtrees that are hanging off a Safe Node and together would represent the reverse shortest path tree towards Omega, each subtree being grown separately inside  $DSet(S)$  where  $S$  is the (virtual) Safe node that is the root of the subtree.

#### 4.3. Being Safe

But once we have placed the selected node in a DSet, we consider its neighbors one by one. If at least one of the neighbors is already in a different DSet than this node, we select the neighbor that provides the shortest alternate path to Omega for the selected node.

Doing so, we have isolated two paths:

- o one along its own shortest path that is contained within its own Dependent Set and that leads to the owner Safe Node of this set.
- o and one via the selected neighbor, along its own shortest path within the selected neighbor's Dependent Set and that leads to the owner Safe Node of that other set.

Because the two sets are different and have no intersection, these paths are non-congruent. And because the two non-congruent paths lead to two different Safe Nodes, this node is Safe.

It might happen that:

- o the selected node's parent is already a Safe Node, in which case the selected node is the Edge AN on its shortest path side.
- o It might also happen that the selected neighbor is already a Safe Node, in which case selected node is the Edge AN on its alternate side.

If both conditions are met for a same AN, then that AN forms a collapsed ARC by itself.

#### 4.4. Bending An ARC

Now we form an ARC as follows:

- o A height is attributed to this ARC that must be strictly more than that of the ARCs it terminates into, if any. The order in which the ARCs are built may be used in some cases.
- o The ARC terminates in the two Safe Nodes that are the owners of the two DSets. The normal behaviour is to make a Edge Link the link to the Safe Node.
- o If the Safe Node at one end forms a collapsed ARC by itself, it may be absorbed in the ARC in order to build a multi-edged ARC.

- o If one of the two Safe Nodes pertains in a ARC or a Comb construct that is higher than the other end, then this ARC may be merged at the Safe Node with its original ARC, in order to form a Comb construct whereby this ARC is a Buttreassing ARC of the Comb. The resulting Comb conserves the height on the original ARC or Comb that it extends.
- o The ARC is built by adjoining the two non-congruent paths that we isolated for the selected node.
- o The selected node is the node farthest from Omega in the resulting ARC, so we make it the Cursor.
- o The link between the selected node and the selected neighbor would not have been used in a classical reverse SPF tree. Here, we have determined that this link is in fact critical to connect 2 zones of the network (the DSets) that can act as a back up for one another in case of the failure of their respective single points of failure (the Safe Nodes).
- o Because the ARC can be used in both directions, each AN along the ARC has two non-congruent paths to the Safe Nodes that the ARC terminates into. So it is a Safe Node. We create individual DSets for each AN and we move the AN to its own DSet.

#### 4.5. Orienting Links

For each ARC Node along the ARC:

- o any link (there can be zero for a collapsed ARC, one for an Edge AN or two of them for a Intermediate AN) between this AN and a next AN along this ARC is made an Intermediate Link, that is, reversible. The normal direction, away from the Cursor, preserves the shortest path.
- o If this AN is an Edge AN for this ARC, than all links off this node that terminate in a Safe Node are made Edge Links, that is, outgoing but not reversible.
- o All the other links left undetermined.

The nodes left in the Dependent Sets but the owner Safe Node are still not Safe. They are moved back to the original Set of Nodes to enable forming additional ARCs which might depend on this ARC in the ARC Set.

#### 4.6. Looping or recursing

We are done processing the particular node we had picked in the original Set of Nodes. If the Set of Nodes as it stands now is not empty, we continue from Section 4.2.

If the Set of Nodes went empty, we are done with this pass and we consider the Dependent Sets that we have put together. In a biconnected graph, there should be one set per node and one node per set, denoting that every node is a Safe Node.

If some portion of the graph is mono-connected, then each mono-connected portion forms the Dependent Set of the Safe Node that is its single point of failure. In order to be maximally redundant, we need to form the ARCs again, within the Dependent Set.

To do so, we remove the Safe Node from the Dependent set and make it Omega. We make the resulting DSet our Set of Nodes and run the algorithm again.

This may recurse a number of times if the graph has mono-connected zones within others.

#### 5. Forwarding Along An ARC Set

Under normal conditions, the traffic flows away from the Cursor of the current ARC and cascades into the next ARC on that side of the Cursor, with the Height of the current ARC decreasing monotonically from ARC to ARC till Omega is reached.

The same goes for a generic Comb construct. When Buttrressing ARCs are applied on a main ARC or other Buttrressing ARCs, the final construct assumes the shape of a tree. The tree may be walked in different manners but the shortest path requires to start going down the current ARC or Buttrressing ARC to its Edge.

In case of Label forwarding, the same recursive technique is applied and a multiple ARC label path is constructed. Each ARC has its own set of label path per Omega, each ARC Set label path being merged into the lower ARC label set, thus at the interconnection point. At minimum, ARC label path should be built from the Cursor toward each edge, but this would require label path recompilation upon Cursor move, the proposed approach is then to build for the normal flow to an Omega one pair of label path from edge to edge.

As this label construct maps the ARC topology with local significant label, the Label Distribution Protocol (LDP) could be reused to announce label association to neighbors on the ARC.

Upon a breakage inside an ARC, until a corrective action takes place, some traffic will be lost. The corrective action might be either operated at the control plane or the data plane, if immediate action and near-zero packet loss is required.

#### 5.1. Control Plane Recovery

Upon a first breakage in an ARC, the Cursor is moved to the breakage point, either a node or a link, so that traffic flows away from the Cursor again.

Upon a second breakage within a same ARC, a segment of the ARC is now isolated. Both breakage points become sinks till an additional corrective action, such as modifying the ARC Set, takes place. All incoming links in the isolated segment are blocked, causing the traffic to exit at the other end of the incoming ARCs.

Blocking an Edge Link in the incoming ARC may create an isolated segment in the incoming ARC as well if it is a second breakage there too, or if both edges of the incoming ARC terminate in the broken segment. In that case the process recurses and the broken zone can be determined as the collection of the isolated segments.

If a segment of an ARC is getting isolated by a dual failure but that ARC segment has incoming Edges then the ARC can be reversed. This reversal is done by reversing of all the incoming Edges, which become outgoing. The segment that was isolated now benefits from multiple exits in a loop free fashion. This process might in turn isolate a segment of an ARC that was incoming and the process recurses and some links flap. If a real exit exists the process will stabilize, but a count to infinity must be put in place to avoid a permanent flapping when a whole ARC Subset is physically isolated. One may consider that this process is in fact the classical link reversal technique, as applied to the DAG of ARCs.

#### 5.2. Data Plane Recovery

Upon a breakage inside an ARC, it is possible in the data plane to reverse the direction of -to turn- a given packet once along the ARC so the packets exits over the other Edge Link. But in order to avoid loops, it is undesirable to reverse the direction of a given packet a second time.

Note that once a given packet leaves an ARC to enter the next, it is free to bounce again in the next ARC. In other Words, the domain that is impacted by a turn is limited to the current ARC itself; the ARC forms the event horizon wherein the notion that a turn happened may cause a loop.

So a local strategy must be put in place inside an ARC to allow a given packet to bounce once upon a breakage, and get dropped upon a second breakage.

In the case of IP packet forwarding, a packet can be tagged when it bounces inside an ARC, or when it passes the Cursor, for instance by reserving a TOS bit for that purpose. When the packet bounces, the bit is set and when the packet leaves the ARC, the bit is reset and may be used again in the next ARC. In the generic case of a Comb, a strategy must be put in place to walk the structure and drop a packet that tries all the Edges. it attempts to pass the Cursor twice in a same direction, meaning that more than a full walk was already accomplished.

#### 5.2.1. Label Switched ARCs

In the case of MultiProtocol Label Switching (MPLS) forwarding, the same result can be achieved with Label Switched ARCs (LSARCs), that are composed of either 3 or 4 Labels Switched Paths (LSPs) along the ARC.

3-Labels method: In this case we lay a primary LSP from the cursor to the Edge in each direction, and a backup LSP Edge to Edge in each direction. So a node along the way has three labels, one primary and two backup, one in each direction. Should the primary path fail, the packet can be placed along the backup LSP in the other direction. We'll note that this method constraints the location of the Cursor. Should the Cursor move, The primary LSPs have to be recomputed, at a minimum between the old and the new location of the Cursor where the direction is reversed.

4-Labels method: In this case we have a primary and a backup LSPs in each direction all of them Edge to Edge, 4 labels total. The labels are independent of the location of the Cursor, so the Cursor can be moved from a node to the next in control plane with no impact on labels. This method consumes an additional label but is more amenable to load balancing techniques and allows each node that inject a packet inside an ARC to make its own decision of the exit edge for a given packet or flow.

#### 5.2.2. Segment Routed ARCs

In the case of an infrastructure that is capable of Segment Routing (SR) [I-D.ietf-spring-segment-routing], the tag in the packet is in essence a Routing Header (RH) via the cursor. The RH forces routing to the destination all the way back up the broken ARC and then down on the other end. via the cursor of a broken ARC.

Upon a breakage, the node detecting the failure reroutes the packet towards the other edge, which means going backwards up to the cursor, and following normal routing from there.

The Routing Header may indicate, as consumed, an entry that points on the broken edge, if that is necessary for the cursor to figure out which is the broken edge so as to route towards the other edge.

### 5.3. Flooding

ARCs probably apply to both unicast and multicast traffic, as illustrated by [I-D.thubert-rtgwg-arc-bicast]. In particular, ARCs enable a redundant flooding of a packet. The flooded packet is injected at all edges ending in Omega, and from there swims upriver along the reverse ARC direction. The packet is then forwarded from the incoming edge of the ARC to the other edge where it is absorbed. On the way along the ARC, the packet is copied into all the ARCs that terminate in this ARC where the process recommences.

Since a packet is finally injected from both edges of any ARC, it should get to all nodes in an ARC even if there is one breakage in that ARC. In normal conditions, at least two copies of the packet circulate in an ARC, one in each direction, and a mechanism should be put in place to make sure that only one copy is injected in an incoming edge.

## 6. ARC Signaling

### 6.1. Serial ARC Representation

A single ARC can be serialized as the sets of endpoints at both edges and the ordered list of nodes in the ARC between the edges. Since the endpoints are effectively nodes in downstream ARCs, the set of all serialized ARCs provides a full description of the topology.

### 6.2. Centralized vs. Distributed computation

An ARC set can be computed with a slightly altered Shortest Path First algorithm, as further explained in Section 4. It results that any node, or all nodes participating to a Link State protocol, may learn the topology and compute an ARC Set. If all nodes compute the topology on their own and asynchronously, micro-loops will follow till the network converges.

It makes more sense to limit the computation of an ARC Set to specific nodes, typically a Path Computation Element (PCE), a Network Management Entity (NME), or nodes in Omega. This is typically what happens in a Software Define Networking (SDN) environment.



### 6.3. ARC Topology Injection

Regardless of the central entity that computes the ARC set, the new or updated ARC Set is serialized in a control message and flooded over itself from Omega as described in Section 5.3.

The new ARC Set can be used as soon as it is received, in the direction from which it is received, since a path along nodes in that direction exists already, through the nodes that forwarded the control messages. The full ARC redundancy is only available when a control message has been received along an ARC in both directions. In that model, there is no micro-loop.

### 6.4. ARC Operations, Administration, and Maintenance

Operations, Administration, and Maintenance (OAM) frames are used within an ARC and flow periodically or asynchronously from an edge to the other. Such frame may carry indications such as a breakage or a congestion, and may be used to control the load balancing, or link reversal operations.

## 7. Other ARC Operations

### 7.1. Node-Local vs. ARC-Wide reaction

ARCs enable forwarding plane reactions to breakages. In the simple case of a single breakage in an ARC, the reaction can be immediate to the discovery of the breakage and consists in rerouting the packet towards the other edge across the cursor, as explained in Section 5.2.

More complex situations require the coordination of all the nodes along an ARC. For instance, load balancing requires the knowledge of the congestion level at multiple points along the ARC, whereas the solution to the Shared Risk Link Group (SRLG) problem discussed in Section 7.3 requires all incoming edges in an isolated ARC segment to be blocked before they can be returned. For such ARC-Wide coordinated reactions, OAM frames are necessary to enable forwarding plane rapid reactions.

### 7.2. Load Balancing

In normal conditions, only the Cursor may distribute its traffic between the two Edge Nodes. If an Edge Node is still congested after the Cursor forwards all its traffic towards the other Edge Node, then the Cursor can be moved towards the congested Edge in order to derive even more traffic towards the other Edge. If both Edges are congested, then a back-pressure can be applied on the incoming ARCs

so that they move their own traffic towards their own alternate Edge. The process may recurse.

It is expected that control frames similar to those defined for MPLS Fault Management Operations, Administration, and Maintenance (OAM) [RFC6291] will echo from Edge Node to Edge Node provide information such as liveliness and load. In order to establish a control loop between the Edge Nodes and the Cursor, the OAM frame would carry at least a logical information whether:

The Edge Node is capable of forwarding data down to the next ARC

the load may be increased (e.g. rate below threshold including hysteresis)

the load should be decreased (e.g. congestion observed as increased latency or buffer bloat)

If the load should be decreased towards of congested Edge Node and the load may be increased towards the other then the Cursor may adjust its balancing of the load, or move Cursor ownership towards the congested Edge if it is already redirecting all the traffic towards the non-congested Edge.

If the Cursor is balancing traffic away from the default position due to a past congestion notification and the Edge that was congested now reports that the load may be increased, then the reverse operation can happen and the Cursor may balance the load back to the original position taking the reverse steps as above.

If the OAM can not be forwarded due to a link or a node failure, then the last node towards the broken Edge becomes Cursor and echoes the OAM frames advertising that it is an Edge node that is blocked, not capable of forwarding data down to the next ARC.

If both Edges are experiencing a congestion then the condition should be reported to the Edge Nodes of all incoming ARCs. Same goes when both Edges are blocked.

### 7.3. Shared Risk Link Group

Essentially, the Shared Risk Link Group (SRLG) problem is that a physical breakage may end up breaking more than one apparently unrelated IP links. such a breakage may end up breaking an ARC in more than one place, effectively creating isolated segments.

The basic approach to solve that problem is the classical link reversal technique. Since ARCs form a DAG as illustrated in

Figure 6, it is possible to return all the incoming edges in an isolated ARC segment so that traffic that circulates inside the segment is actually fed back in incoming ARCs. The incoming ARCs are considered broken on that edge so all the traffic is fed into the other edge. If this causes the incoming ARC to be doubly broken, the process recurses. in that incoming ARC. Over a number of iterations, if there is an exit, it will be found and the traffic will be funneled that way. If there is none, after a certain number of iterations, the process counts to infinity and stops.

If the iterations are performed too quickly, the process may cause micro-loops. OAM frames circulating within the broken segment can solve that issue. On the way in, the OAM frame should block all incoming ARCs, which effectively causes the edges to appear broken in incoming ARCs. On the way back, the OAM frame returns the incoming edges to be used the other way.

#### 7.4. Olympic Rings

By Olympic Ring problem we mean how to optimally reuse the multiple path opportunities that interconnecting 2 rings enable. ARCs can simply be deployed inside a ring A to reach a connected ring B by installing an ARC between adjacent interconnections on the rings. If the rings only connect at one point, there is a single ARC going all the way around the ring, with the cursor at the far side. If there are more than one interconnection, then you always end up with as many ARCs as there are interconnections.

A packet being forwarded inside a ring picks the side of the ring that is away from the cursor, taking effectively the shortest path to the next ring. If a hop is broken, then the packet is returned to the other edge of the ARC, which is the adjacent interconnection between the ARCs.

Note: There is no need in that model to artificially disable one hop in the ring and re-enable it in case of breakage.

#### 7.5. Routing Hierarchies

The ARC methods may be used to build and/or leverage routing hierarchies, allowing high availability at multiple hierarchical levels. In one hand, the view of an ARC Set can be simplified by abstracting an ARC as a node in a DAG. The view of the routing topology is thus simplified, as illustrated in Figure 6.

In the case of connected rings, abstracting a full ring as a node, ARCs can be applied to a graph of rings, providing another level of redundancy and an abstract end-to-end path computation, ring to ring

to ring. ARCs may be used to make that computation resilient as well.

## 8. Manageability

This specification describes a generic model. Protocols and management will come later

## 9. IANA Considerations

This specification does not require IANA action.

## 10. Security Considerations

This specification is not found to introduce new security threat.

## 11. Acknowledgments

The authors wishes to thank Dirk Anteunis, Stewart Bryant, IJsbrand Wijnands, George Swallow, Eric Osborne, Clarence Filsfils and Eric Levy-Abegnoli for their participation and continuous support to the work presented here.

## 12. References

### 12.1. Normative References

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

### 12.2. Informative References

[I-D.ietf-spring-segment-routing]  
Filsfils, C., Previdi, S., Bashandy, A., Decraene, B., Litkowski, S., Horneffer, M., Shakir, R., Tantsura, J., and E. Crabbe, "Segment Routing Architecture", draft-ietf-spring-segment-routing-00 (work in progress), December 2014.

[I-D.thubert-rtgwg-arc-bicast]  
Thubert, P. and I. Wijnands, "Applying Available Routing Constructs to bicasting", draft-thubert-rtgwg-arc-bicast-01 (work in progress), October 2013.

[I-D.thubert-rtgwg-arc-vs-mrt]  
Thubert, P., Enyedi, G., and S. Ramasubramanian, "Available Routing Constructs", draft-thubert-rtgwg-arc-vs-mrt-01 (work in progress), January 2014.

- [RFC5921] Bocci, M., Bryant, S., Frost, D., Levrau, L., and L. Berger, "A Framework for MPLS in Transport Networks", RFC 5921, July 2010.
- [RFC6291] Andersson, L., van Helvoort, H., Bonica, R., Romascanu, D., and S. Mansfield, "Guidelines for the Use of the "OAM" Acronym in the IETF", BCP 161, RFC 6291, June 2011.

Authors' Addresses

Pascal Thubert (editor)  
Cisco Systems, Inc  
Building D  
45 Allee des Ormes - BP1200  
MOUGINS - Sophia Antipolis 06254  
FRANCE

Phone: +33 497 23 26 34  
Email: pthubert@cisco.com

Patrice Bellagamba  
Cisco Systems  
214 Avenue des fleurs  
Saint-Raphael 83700  
FRANCE

Phone: +33.6.1998.4346  
Email: pbellaga@cisco.com

RTGWG  
Draft  
standards Track

P. Thubert, Ed. Internet-  
ciscoIntended status: S  
IJ. Wijnands Expires: April 19, 2014

Cisco Systems

October 18, 2013

Applying Available Routing Constr

cts to bicasting draft-thubert-rtgwg-arc-bicast-01Abstract Th  
is draft introduces methods that leverage the concept of ARC to enable bicastin  
g operations. Requirements Language The key words "MUST", "MUST NOT", "REQUIRED"  
, "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED"  
, "MAY", and "OPTIONAL" in this document are to be interpreted as described in  
RFC 2119 [RFC2119]. Status of this Memo This Internet-Draft is submitted in f  
ull conformance with the provisions of BCP 78 and BCP 79. Internet-Drafts are  
working documents of the Internet Engineering Task Force (IETF). Note that ot  
her 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 inap  
propriate to use Internet-Drafts as reference material or to cite them other th  
an as "work in progress." This Internet-Draft will expire on April 19, 2014. Cop  
yright Notice Copyright (c) 2013 IETF Trust and the persons identified as the  
document authors. All rights reserved. This document is subject to BCP 78 and  
the IETF Trust's Legal Provisions Relating to IETF Documents ([http://trustee.ietf.org/](http://trustee.ietf.org/license-info)  
[license-info](http://trustee.ietf.org/license-info)) in effect on the date of publication of this document.

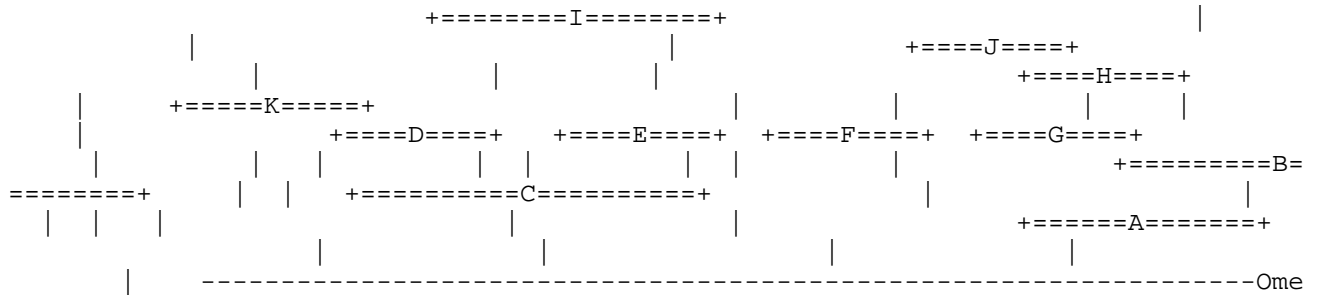
Please review these documents carefully, as they describe your rights and rest  
rictions with respect to this document. Code Components extracted from this do  
cument must include Simplified BSD License text as described in Section 4.e of  
the Trust Legal Provisions and are provided without warranty as described in th  
e Simplified BSD License. Thubert & Wijnands Expires April 19, 2014

[Page 1]

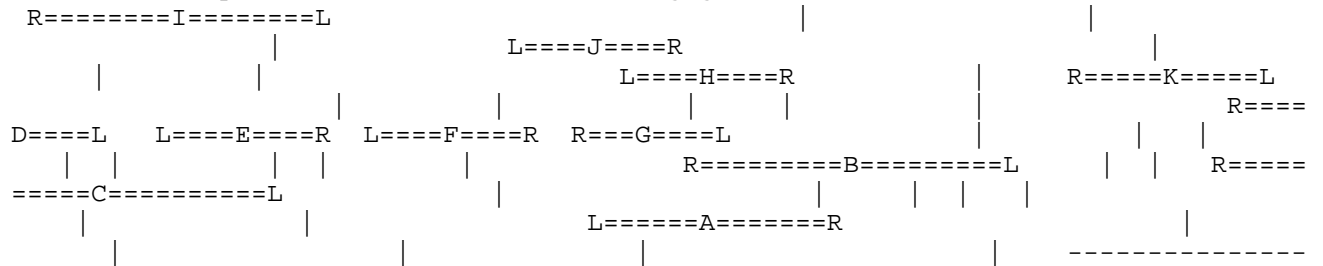
Internet-Draft	ARC bicasting	October 2013	Table of
Contents	1. Introduction	2	
2. Terminology		3	3. Downward Bicast
3. Downward Bicast Operation		4	4. Upward Bicast
4. Upward Bicast Operations		5	4.1. Resolving crossing
5. Applicability		6	4.2. Single Point of Failure
5.1. In conjunction with Protocol Independent Multicast		7	
6. Manageability		7	
7. IANA Considerations		7	
8. Security Considerations		7	9. Acknowledgements
9. Acknowledgements		7	10. References
10. References		8	10.1. Normative References
10.1. Normative References		8	10.2. Informative References
10.2. Informative References		8	Authors' Addresses
Authors' Addresses		8	
1. Introduction			

Traditional routing and forwarding uses the concept of path as the basic routing paradigm to get a packet from a source to a destination by following an ordered sequence of arrows between intermediate nodes. In this serial design, a path is broken as soon as a single arrow is, and getting around a breakage can require path recomputation, network reconvergence, and incur delays to till service is restored. Available Routing Constructs [I-D.thubert-rtgwg-arc] (ARC) introduces the concept of ARC as a routing construct made of a sequence of nodes and links with 2 outgoing edges, that is this resilient to one breakage so that an ARC topology is resilient to one breakage per ARC. The routing graph to reach a certain destination is expressed as a cascade of ARCs, which terminates in an abstract destination Omega, each ARC providing its own independent domain of fault isolation and recovery:

Thubert & Wijnands Expires April 19, 2014 [Page 2]



This cascade of ARCs appears ideally suited to the operation of bica sting (a.k.a. duocasting), which consists in sending two copies of a single packet, if possible over divergent - that is fully or partially non-congruent - paths, in order to augment the chances that at least one of the copies reaches the destination timely.2. Terminology The draft uses the terminology defined in [I-D.thubert-rtgwg-arc]. This specificatin also introduces Sided ARCs, that is ARCs with at least an Edge that is known as Left and an Edge that is known as Right. The sense of Left and Right adds up to the existing sense of height that is already defined in [I-D.thubert-rtgwg-arc].

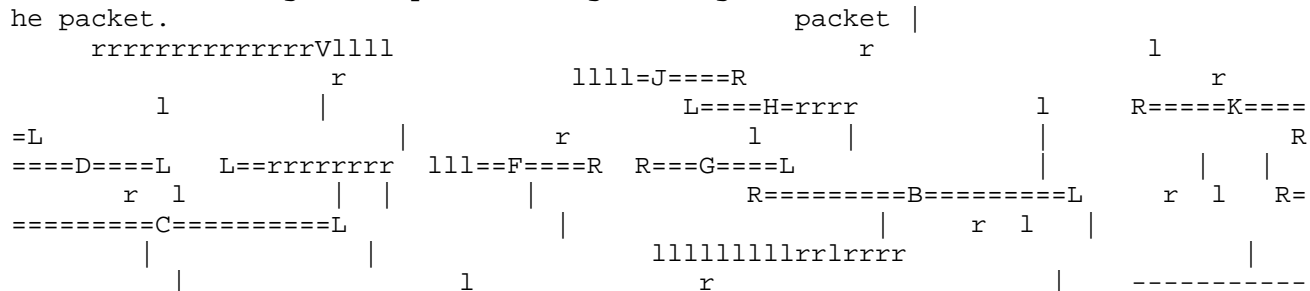


One way of doing this is to The basic rule is that an ARC MUST have at least one Left and one Right Edge. Thubert & Wijnands Expires April 19, 2014 [Page 3]



leg of an ARC between the cursor and the Edge inherits the side of the Edge. In a Comb, the whole buttressing ARC inherits the side of the Edge. o An Edge ending in Omega can arbitrarily become Left or Right as long as the basic rule is satisfied. o An Edge that does not end in Omega inherits the side of an ARC it terminates into, again as long as the basic rule is satisfied. o A collision occurs if all the Edges end up on the same side. The shortest path is used to resolve the collision and restore the basic rule: the Edge closer to Omega and all buttressing ARCs on the same side of the cursor keep the side, and the other Edges are toggled. In case of equal cost, an other tie breaker must be used. o For instance, this situation occurs in the representation above for ARC F, which has both ends ending in a Right side of ARCs, and since the Edge closer to Omega is the one that ends in ARC C, that Edge becomes Right and the other becomes Left.

3. Downward Bicasting Operation Two copies of a same packet from a given node are forwarded downwards along opposite side of the cascading ARCs, each packet bearing an indication (such as a tag or a label) of its intended side, Left or Right. The packets exit the ARCs along their paths through an Edge that matches the indication in the packet.



-----Omega As it goes, the method does not guarantee a full non congruence of the paths, as illustrated above. In case of a breakage, this can be compensated by the capability to return a packet along an ARC upon a failure, that is already used to protect unicast traffic.

Thubert & Wijnands Expires April 19, 2014 [Page 4]

```

packet | 1 Left packet path rrrrrrrrrrrrrrrr
Vlllll R Right packet path r 1
r      llll=J====R      r      1
|      L====H=rrrr      1      R====K====L      1
|      |      r      1      |      R====D====L      L=rrrr
rrrrrr lll==F====R R===G====L      |      |      r 1
|      |      |      R=====B=====L      r 1      R=====C=====
=L      |      |      rrrrrrrrr\lllll      |      |
|      |      r      /\      1      |      |

```

-----Omega 4. Upward Bicasting Operations It is also possible with a downward bicasting to place states in the intermediate routers in order to provision an upward bicast path from Omega to a source D. In that case, if the graph is biconnected, it is possible to resolve the pathological cases so as to ensure a real separation of the left and Right paths.4.1. Resolving crossing ARCs The first pathological case occurs when both Left and Right packet cross over the same ARC, as illustrated below. Say that the Right reservation comes first and sets up the right path:

```

r      |      R====D====L L=rrrrrrrrr L====F====R R===G
====L      |      |      r      |      |      |
R=====B=====L      r      |      R=====C=====L      |
|      r      |      |      |      |      |      L===
===Arrrrrrrrr      |      |      |      |      |      r

```

-----Omega Then comes the left reservation which collisions:Thubert & Wijnands Expires April 19, 2014 [Page 5]

```

      r      l      R====D====L      L==rrrrrrrrr
111==F====R  R===G====L      |      |      |
      |      R=====B=====L      r  l  R=====C=====L
      |      |      r  l  |      |      |
      L=====Arrrr?rrrr      |      |
      |      r      |      |
-----Omega
t of the common ARC is common to both path and expose the bicasted traffic. Th
e resolution is to leave the second packet through but prune the unwanted state
s along the collision segment of the ARC afterwards.
      r      l      R====D====L      L==rrrrrrrrr 111==F====R R=
==G====L      |      |      |      |      |
      R=====B=====L      r  l  R=====C=====L      |
      |      r  l  |      |      |      |
11111111==rrrrr      |      |      |      |      |
      r      |      |
-----Omega

```

States along the ARC between the two incoming points are cleaned, up and the paths that were generated by the Left and Right packets are now crossed. This results in two non-congruent upward paths.

4.2. Single Point of Failure The second pathological case occurs when both Left and Right packet reach a same ARC at the same node, which is this a Single Point Of Failure (SPoF) for both paths.

```

      r      |
      R====D====L  L==rrrrrrrrr  L====F====R  R===G====L
      |      |      |      |      |
      r      |      |      |      R=====B=====L      r
/  R=====C=====L      |      |      r/
      |      |      |      L=====A==rrrrrrr
      |      |      |      r      |
-----Omega

```

The resolution is to reject the second packet and send it back along the incoming ARC to exit on the other side. The rejected packet cleans up the states that it has created on its way back and then creates states on the other side of the ARC.

Thubert & Wijnands Expires April 19, 2014 [Page 6]

[illegible]

-----Omega This is in fact what happens also in the case of a monoconnected zone, or if a breakage cause the downward packet to bounce.5. A pplicability5.1. In conjunction with Protocol Independent Multicast Upwards bi casting can be used for Protocol Independent Multicast PIM [RFC4601] and Point-to-Multipoint and Multipoint-to-Multipoint Label Switched Paths mLDP [RFC6388].

A bicasteds downwards Join message would establish two non congruent return paths, each path joining the receiver and Omega that is the set of existing receivers.6. Manageability This specification describes a generic model. Protocols and management will come later7. IANA Considerations This specification does not require IANA action.8. Security Considerations This specification is not found to introduce new security threat.9. AcknowledgementsThubert & Wijnands

Expires April 19, 2014

[Page 7]

Internet-Draft                      ARC bicasting                      October 2013                      The authors wishes to thank Dirk Anteunis, Stewart Bryant, Patrice Bellagamba, George Swallow, Eric Osborne, Clarence Filsfils and Eric Levy-Abegnoli for their participation and continuous support to the work presented here.

10. References

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

11. Informative References

[I-D.thubert-rtgwg-arc] Thubert, P. and P. Bellagamba, "Available Routing Constructs", Internet-Draft draft-thubert-rtgwg-arc-00, October 2012.

[RFC4601] Fenner, B., Handle y, M., Holbrook, H. and I. Kouvelas, "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)", RFC 4601 , August 2006.

[RFC6388] Wijnands, IJ., Minei, I., Kompella, K. and B. Thomas, "Label Distribution Protocol Extensions for Point-to-Multipoint and Multipoint-to-Multipoint Label Switched Paths", RFC 6388, November 2011.

Authors' Addresses

Pascal Thubert, editor  
Cisco Systems, Inc  
Village d'Entreprises Green Side  
400, Avenue de Roumanille  
Batiment T3  
Biot - Sophia Antipolis, 06410 FRANCE  
Phone: +33 497 23 26 34 Email: pt.hubert@cisco.com

IJsbrand Wijnands  
Cisco Systems  
De kleetlaan 6a  
Diegem,  
1831 Belgium  
Email: ice@cisco.com

Thubert & Wijnands Expires April 19, 2014

[Page 8]

Routing Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: July 28, 2014

IJ. Wijnands, Ed.  
L. De Ghein  
Cisco  
G. Enyedi, Ed.  
A. Csaszar  
J. Tantsura  
Ericsson  
January 24, 2014

Tree Notification to Improve Multicast Fast Reroute  
draft-wijnands-rtgwg-mcast-frr-tn-02

Abstract

This draft proposes dataplane triggered Tree Notifications to support multicast fast reroute for PIM and mLDP. These Tree Notifications are initiated by a node detecting the failure to a Repair Node downstream. A Repair Node is a node that has a pre-built backup path that can circumvent the failure. Using this mechanism, a Repair Node has the ability to learn about non-local failures quickly without having to wait for the IGP to convergence. This draft also covers an optional method to avoid bandwidth usage on the pre-built backup path.

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 July 28, 2014.

Copyright Notice

Copyright (c) 2014 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal

Provisions Relating to IETF Documents  
(<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

## Table of Contents

1. Terminology and Definitions . . . . .	3
2. Introduction . . . . .	4
3. Improving Non-local failures . . . . .	4
3.1. Downstream Tree Notifications . . . . .	5
3.2. DTN processing logic . . . . .	5
3.3. Repair Node discovery . . . . .	7
3.3.1. Repair Node Information item . . . . .	8
4. Reduce the bandwidth consumption in networks with fast failover response times . . . . .	8
4.1. Joining a secondary tree in blocking mode . . . . .	9
4.2. Upstream Tree Notifications . . . . .	9
5. MRT/MCI-Only Mode . . . . .	10
6. TN Authentication . . . . .	10
7. The TN Packet . . . . .	11
7.1. TN Packet Format . . . . .	11
7.1.1. TN TimeStamp TLV Format . . . . .	13
7.1.2. TN Signature TLV Format . . . . .	13
8. PIM Specific TN Components . . . . .	14
8.1. RNI item in PIM Join Message . . . . .	14
8.2. Tree Information Item . . . . .	16
8.3. Incremental deployment . . . . .	17
9. mLDLP Specific TN Components . . . . .	17
9.1. RNI item in mLDLP Label Mapping . . . . .	18
9.2. Tree Information Item . . . . .	19
10. Acknowledgements . . . . .	19
11. IANA Considerations . . . . .	20
12. Security Considerations . . . . .	20
13. References . . . . .	20
13.1. Normative References . . . . .	20
13.2. Informative References . . . . .	21
Authors' Addresses . . . . .	21

## 1. Terminology and Definitions

MoFRR : Multicast only Fast Re-Route.

LFA : Loop Free Alternate.

mLDP : Multi-point Label Distribution Protocol.

PIM : Protocol Independent Multicast.

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

tree : Either a PIM (S,G)/(\*,G) tree or a mLDP P2MP or MP2MP LSP.

OIF : Outgoing InterFace, an interface used to forward multicast packets down the tree towards the receivers. Either a PIM (S,G)/(\*,G) tree or a mLDP P2MP or MP2MP LSP.

IIF : Incoming InterFace, an interface where multicast traffic is received by a router.

MCE : MultiCast Egress, the last node where the multicast stream exits the current transport technology (MPLS-mLDP or IP-PIM) domain or administrative domain. This maybe the router attached to a multicast receiver.

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

DTN : Downstream Tree Notification.

UTN : Upstream Tree Notification.

TN : Tree Notification, Upstream or Downstream

JM : Join Message, the message used to join to a multicast tree, i.e. to build up the tree. In PIM, this is a JOIN message, while in mLDP this corresponds to a Label Mapping message.

MRT : Maximally Redundant Trees.

Repair Node : The node performing a dual-join to the tree through two different UMHs. Sometimes also called as dual-joining node or merging node (it merges the secondary and primary tree).



RNI : The Repair Node Information is an item included in the TN which holds the necessary repair information when the TN is sent to the Repair Node.

Branching Node : A node, (i) which is considered as being on the primary tree by its immediate UMH and (ii) which has at least one OIF on the secondary tree installed for a multicast tree.

## 2. Introduction

Both [I-D.karan-mofrr] and [I-D.atlas-rtgwg-mrt-mc-arch] describe "live-live" multicast protection, where a node joins a tree via different candidate upstream multicast hops (UMH). With MoFRR the list of candidate UMHS can come from either ECMP or Loop Free Alternate (LFA) paths towards the MultiCast Ingress node (MCI). With MRT, the candidate UMHS are determined by looking up the MCI in two different (Red and Blue) topologies. In either case, the multicast traffic is simultaneously received over different paths/topologies for the same tree. The node 'dual-joining' the tree needs a mechanism to prevent duplicate packets being forwarded to the end user. For that reason a node 'dual-joining' the tree only accepts packets from one of the UMHS at the time. Which UMH is preferred is a local decision that can be based on IGP reachability, link status, BFD, traffic flow monitoring, etc...

Should the node detect a local failure on the primary UMH, the node has an instantly available secondary UMH that it can switch to, simply by unblocking the secondary UMH. The dual-joining node is also called Repair Node in the following.

This draft attempts to improve these solutions by:

- o Improving fail-over time and the reliability of failure detection for non-local failures; and
- o Reducing the bandwidth consumption in a network with fast failover response times, by avoiding sending the multicast traffic over the secondary path.

## 3. Improving Non-local failures

If a failure is not local and happens further upstream, the dual-joining node needs a fast mechanism (i) to detect the upstream failure and (ii) to learn that other upstream nodes cannot circumvent the failure. Existing methods based on traffic monitoring are limited in scope and work best with a steady state packet flow.

Therefore, we propose a method which can trigger the unblocking the secondary UMH independently of the packet flow.

Figure 1 shows an example. Consider that, e.g., node A goes down. Nodes C, D and E cannot detect that locally, so they need to resort to other means. After detecting the failure, node C should not change to its secondary UMH (node J) as it won't help for the failure of A. Node D, on the other hand, will have to unblock its secondary UMH (node I). Yet again, with MoFRR, node E should not unblock its secondary UMH (node K): (i) this won't help in resolving the failure of node A, and (ii) one of its upstream nodes (node D in this case) will be able to restore the stream with a fail-over action.

### 3.1. Downstream Tree Notifications

When a node detects a local failure of its primary UMH it MUST originate a Downstream Tree Notification (DTN) to all the Repair Nodes directly below it in the multicast tree. The method of discovering such nodes is described in Section 3.3. When a Repair Node receives a DTN containing the primary UMH of the node, it must switch to the secondary UMH.

DTN packets are sent to the Repair Node via unicast. The packet may be forwarded using any transport that is available (MPLS or IP) to reach the destination. The IP precedence in the IP header should have a value of 6 (Internetwork Control). The EXP field (Traffic Class field) in the MPLS header should have a value of 6. The DTN packets are identified by a well known port number (to be allocated). Using a well-known port number it is easy for the Repair Node to identify the DTN packet and invoke the procedures as described in this draft. We are proposing to allocate different port numbers for PIM and mDLP since it will be easier to dispatch the packet to the right process dealing with this request.

When a router detects a local failure, it should sent out the DTN packet to the Repair Node as fast as possible. The sooner the Repair Node gets the packet, the sooner the traffic can be restored. It is recommended that the DTN packet is pre-created and originated from the data-plane. The same is true for receiving the DTN packet on the Repair Node, the faster it can be processed, the faster the traffic is restored. For both forwarding and processing the DTN, control-plane interaction SHOULD be avoided to get the best failover results.

### 3.2. DTN processing logic

When a DTN packet is received on the Repair Node it must determine which tree and UMH the notification is for. The information encoded in the DTN is specific for the type of tree being used, i.e. PIM vs

mLDP. For details on the specific encoding see Section 8 and Section 9 for the details. Once the Repair node has determined the tree and the UMH, the following rules are use for processing the DTN.

1. If the UMH encoded in the DTN packet is the primary UMH in the tree, the secondary UMH MUST become the new primary UMH and the old primary MUST become the secondary.
2. If the UMH encoded in the DTN packet is the secondary UMH in the tree, no action needs to be taken.
3. If a DTN notification has been received for both the primary and secondary UMH in the tree, a new DTN notification MUST be originated to the Repair Node(s) downstream from this node.

In order for the Repair Node to determine that a DTN notification was received for both the primary and secondary UMH, it must store the fact a DTN was received for a particular UMH.

Consider the example in Figure 1 below. MCI is the root of a tree that includes the nodes as follows (based on the primary UMH).

```

->F->G->H->I
MCI
->A->B->C->D->E

```

Node C, D and E are candidate Repair Nodes.

```

-- Primary UMH
++ Secondary UMH

```

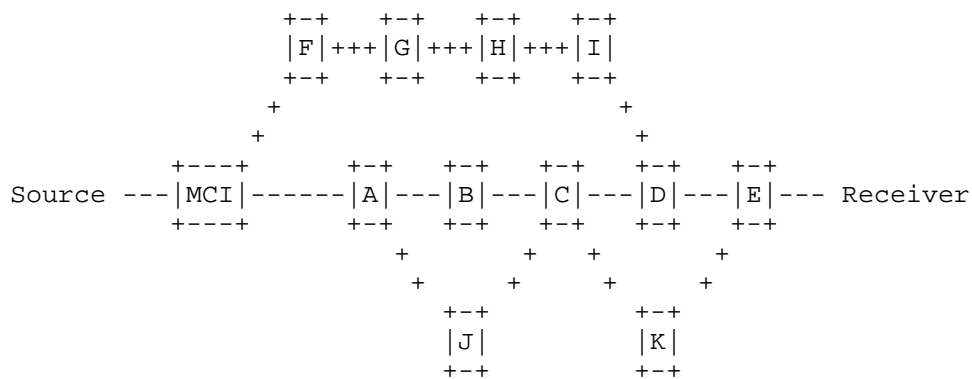


Figure 1: Remote failure example

Suppose that the link between node A and B failed, B is directly connected and will detect the failure locally. In this case, node B is the only node that detects the failure and will originate a DTN to its downstream repair node C. Node C will receive the DTN for the UMH that is the primary UMH. Following rule 1 (Section 3.2), node C will make the backup UMH the new primary. No further action is needed because C has repaired the tree via node J. Note, J would not have sent a DTN to node C because J is not directly connected to the failing link.

Suppose that node A fails, B and J are directly connected and detect the failure locally. A DTN packet is triggered to first downstream repair node of A, which is node C. Node C is an unusable Repair Node because it will receive DTN for both the primary UMH (from B) and the secondary UMH (from J). Following rule 3 (Section 3.2), C can't repair the tree and must send a new DTN packet towards the Repair Nodes of C, which are D, on the primary path, and E, on the secondary path.

Suppose that the link between A and the MCI failed. Node A is directly connected to the failure and will trigger a DTN packet to its downstream repair node(s). In this case, node A has learned about the downstream repair node C twice, the primary UMH (via node B) and secondary UMH (via node J). Node A will therefore send a DTN packet including both the primary and secondary UMH to node C (see Section 7 for details on the encoding). Following rule 3 (Section 3.2), C can't repair the tree and must send a new DTN packet towards the Repair Nodes of C, which are D, on the primary path, and E, on the secondary path.

The DTN packet that D received from C will match against the primary UMH. Following rule 1, D will activate the backup path to I. The DTN packet that E received from C will match against the backup UMH, following rule 2, no action is taken. In the example one can see that we recovered from the failure because node D started accepting the data packets from node I and is forwarding them to node E.

### 3.3. Repair Node discovery

In example Figure 1 we wrote that nodes C, D and E are the repair nodes. How does a node determine that it is a Repair Node? The rule is straightforward, a node that is enabled to join two UMH's, one in active the other in backup ([I-D.karan-mofrr]), is a repair node on the tree. A Repair node has the ability to repair the tree for the nodes upstream from this node. In order for the Repair Node to get notified of upstream failures (ie DTN), the nodes upstream from the Repair Node need to learn about it.

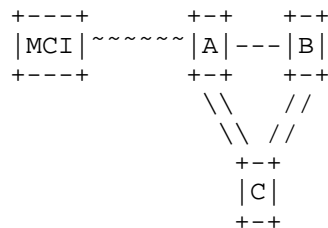
### 3.3.1. Repair Node Information item

A Repair Node MUST advertise its own address (either a router ID or any directly connected address) and an UMH identifier to the nodes upstream on the tree. This address and UMH are part of the RNI (Repair Node Information) item that is included in the JM. The RNI is carried hop by hop in the JM upstream. If a node along the path is not a Repair Node, it will save the RNI and forward it further upstream. If the node is a Repair Node, it will save the RNI and include its own RNI in the JM sent further upstream. If a Repair Node changes one of its UMH's, it needs to trigger a new RNI to its upstream node(s) to notify them of the changed UMH. If a RNI is received and it does not match the saved RNI, the new RNI overrides the old RNI and triggers a JM with the new RNI to its upstream node(s). A RNI includes protocol specific information on how to identify the tree and UMH. For that reason it is documented in the protocol specific sections Section 8 and Section 9.

The Repair Node MAY include additional information in the RNI for reasons of security and robustness, please see Section 6 and Section 7.1.

## 4. Reduce the bandwidth consumption in networks with fast failover response times

In some of networks, such as aggregation networks, bandwidth is more sparse than, e.g., in core networks. Live-live multicast protection results in more bandwidth consumption in the network as it continuously pulls traffic on both trees. In such networks it is relevant if the capacity serving backup purposes can be used, most of the time, by best-effort or even by lower-than-best-effort traffic.



Nodes A and B have receivers. Double lines show bandwidth consumption that is superfluous when there is no failure in the network.

Figure 2: Example for secondary segments occupying bandwidth in MoFRR

In live-standby mode the aim is that the secondary tree is not forwarding multicast traffic as long as there is no failure. In order to achieve such a "live-standby" multicast protection the following procedures must be followed:

- o Upstream nodes block their OIF when they are part of a standby tree.
- o If all of the OIF's of the node are marked as blocking, the node joins the tree in blocking mode further upstream.
- o A procedure so that the upstream node can quickly unblock its OIF and starts to forward.

#### 4.1. Joining a secondary tree in blocking mode

The JM sent to the secondary UMH includes an identifier to indicate the upstream node MUST not forward packets down this branch of the tree. The identifier is TBD. The mechanism to join a secondary path is identical to what the MRT and MoFRR drafts describe, i.e. a Repair Node simply sends a secondary JM through another UMH (on another topology, in case of MRT). If a node receives a JM without a blocking identifier for an OIF that previously was in blocking mode, the blocking mode is reset and the node starts forwarding out of this interface. If this node joined the tree in blocking mode further upstream, a new JM MUST be originated to reset the blocking state further upstream.

#### 4.2. Upstream Tree Notifications

In order to make an upstream node start forwarding on the backup path quickly after a failure was detected on the primary UMH, we send an Upstream Tree Notification (UTN) to the upstream node on the backup UMH. The failure on the primary UMH may be local or detected using a DTN. The UTN received by the upstream node should be processed in the data-plane and reset the blocking state of the OIF. If this node also joined the tree in blocking mode upstream, a UTN has to be forwarded further upstream. This procedure is repeated until we find a node that is not in blocking mode or we reached the MCI.

When the upstream node resets the blocking mode in the data-plane, the control plane will still have the blocking mode set. In order for the control plane to get in sync with the data-plane, the node that originated the UTN MUST also trigger a JM without blocking mode.

The upstream node receiving the UTN must be able to identify the tree which the notification is sent for, as well as the downstream interface it applies to. The information is encoded in a same RNI

item that is used for DTN packets. For details please see the protocol specific sections Section 8 and Section 9.

Like DTN packets, UTN packets are sent via unicast to the upstream node.

## 5. MRT/MCI-Only Mode

If each node in the network supports UTN and also all nodes support MRT, the nodes may work in "MRT/MCI-only" mode.

In MRT/MCI-only mode, there is one single Repair Node for all failures, the MCI. Other nodes MUST NOT consider themselves as Repair Nodes. MRT ensures the necessary maximally disjoint secondary tree up to the MCI, on a second topology. Only the MCI MUST keep its OIFs corresponding to the secondary tree blocked. Similarly, only MCEs MUST keep their secondary backup IIFs blocked. Any other nodes MUST NOT block their (secondary) IIFs or OIFs.

In MRT/MCI-only mode, the UTNP MUST be forwarded directly to the MCI. This mode enables that a node detecting a downstream failure of the primary tree MAY send a UTNP upstream towards the source/MCI on the primary tree.

If an UTNP is received by the MCI on the secondary topology in "MRT/MCI-only" mode, the MCI MUST unblock the OIF where the UTNP was received. This activates a whole sub-tree of the secondary tree.

If an UTNP is received by the MCI on the primary topology in "MRT/MCI-only" mode, the MCI gets no information on which leg to activate on the secondary tree, so it MUST activate (unblock) all secondary legs.

## 6. TN Authentication

If a malicious attacker can reproduce the TN packet format, unwanted reconvergence can be triggered. In order to avoid such attack, a TN packet MAY contain a digital signature. Having authentication is optional, it can be enabled or disabled in the network. If however security is enabled, all the nodes must share the same secret key, which they get either by configuration or from the multicast routing protocol. Moreover, for protection against reply attacks, each TN packet must contain a sequence number.

The sequence numbers in the network are not necessarily synchronised, instead, each node can have its own. Sequence numbers can be

generated arbitrarily, it can be even some random value; the only requirement is to create a new sequence number each time a reconvergence was triggered due to a TN (i.e. the sequence number was used).

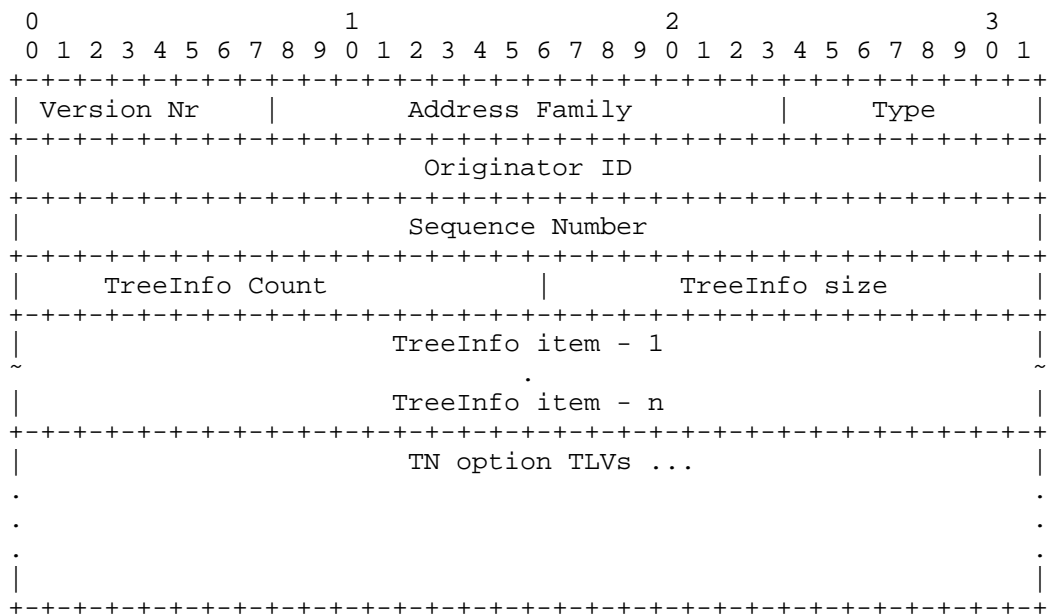
The originator of the DTN packet MUST use the sequence number of the Repair Node to create a TN signature TLV (see Section 7.1.2). For UTN packet the sender MUST use its own sequence number, what it sent previously to its UMH. The destination in this case must check validity based on the sequence number of the sender.

A sequence number is learned from JM and part of the RNI. It is the responsibility of multicast routing protocol to protect JM against malicious change.

## 7. The TN Packet

### 7.1. TN Packet Format

A Tree Notification is an IPv4 or IPv6 UDP packet with the following format.





Version number: This is a 1 octet field encoding the version number, currently 0.

Address Family: This is a 2 octet field encoding a value from ADDRESS FAMILY NUMBERS in [RFC3232] that encodes the address family for the Root Address of the tree.

Type: This is a 1 octet field encoding the message type, currently two are defined;

Type 0: Downstream Tree Notification.

Type 1: Upstream Tree Notification.

Originator ID: 4 bytes long unique ID of the originator. That can be some loopback IPv4 address if there is such, or can be set by the operator.

Sequence Number: Number unique for each failure case. It is recommend to start at 0, and to be increased by 1 each time a new TN is originated. The Sequence number may differ at each node, thus the sender and the receiver must know the same sequence number.

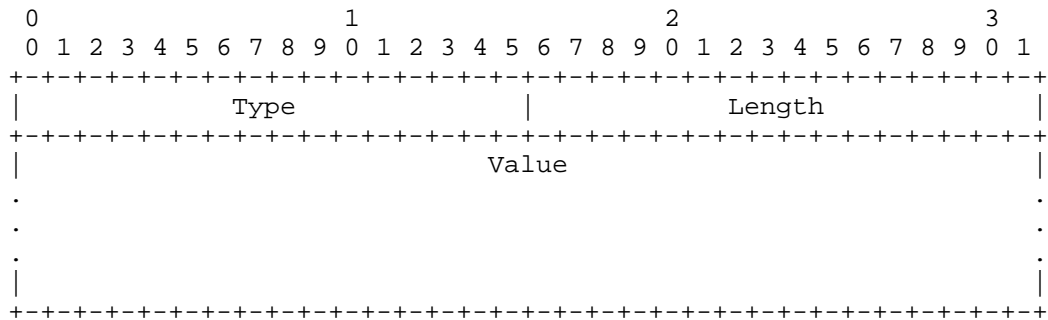
TreeInfo count: 2 octet field encoding the number of TreeInfo items includes.

TreeInfo size: 2 octet field encoding the number of octets use to encode the TreeInfo's following.

TreeInfo item: The encoding of this field is protocol specific, see Section 8 and Section 9.

TN option TLVs: TLVs (Type-Length-Value tuples) describing additional options for TN packets.

The TLV's have the following format.



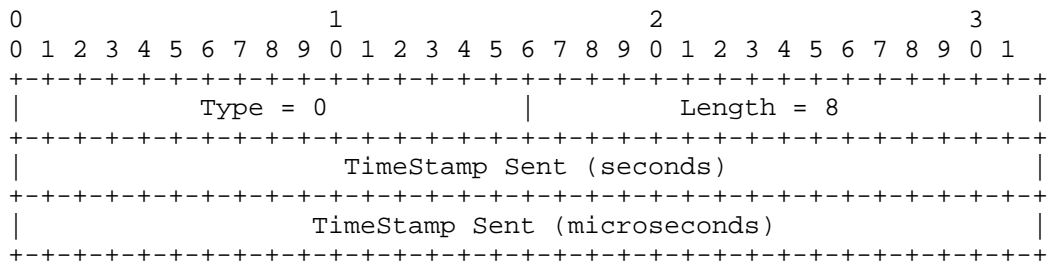
Type: This is a 2 octet field encoding the type number of the TLV.

Length: This is a 2 octet field encoding the length of the Value in octets.

Value: String of Length octets, to be interpreted as specified by the Type field.

#### 7.1.1.1. TN TimeStamp TLV Format

The TimeStamp is an optional TLV that MAY be included when the TN was originated, it has the following format.



TimeStamp: The TimeStamp is the time-of-day (in seconds and microseconds, according to the sender's clock) in NTP format [NTP] when the Tree Notification is sent.

#### 7.1.1.2. TN Signature TLV Format

TN Signature is an optional TLV, which protects the whole TNP (including other TLVs) against attacks thus it must be the last TLV if present. The signature is SHA-512 hash value. The input of the hash function is as follows:

```

+-----+
| Complete packet content without signature TLV |
+-----+
|                               Secret key       |
+-----+

```

Signature input: The input of the hash function is the packet extended with TN security key

The build up of the TLV is as follows:

```

0                               1                               2                               3
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-----+-----+-----+-----+-----+-----+-----+-----+
|                               Type = 1                               |
+-----+-----+-----+-----+-----+-----+-----+-----+
|                               Hash function result                     |
|                               .                                         |
|                               .                                         |
|                               .                                         |
+-----+-----+-----+-----+-----+-----+-----+-----+

```

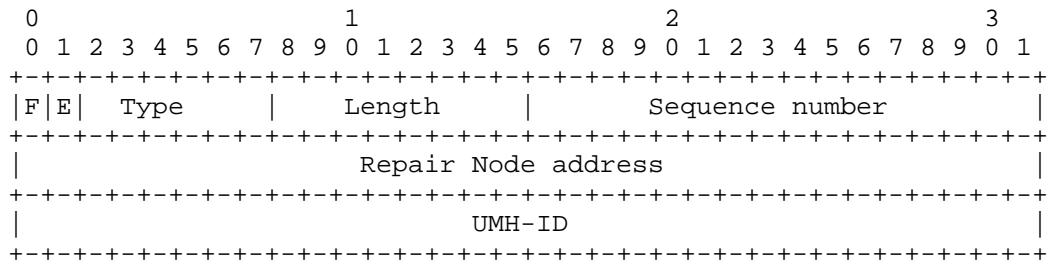
Signature: SHA-512 signature protecting TN packets.

## 8. PIM Specific TN Components

In this section we are documenting the PIM specific data-structures and procedures (if they are different from the generic procedures are defined in this document). As described in this document, TN packets are UDP/IP packets sent via unicast to its destination. The UDP port number for PIM is set to the (to be) assigned IANA port number for PIM-TN.

### 8.1. RNI item in PIM Join Message

As described previously, PIM must insert the RNI when sending a PIM join to its UMH. The RNI includes its router ID, sequence number and UMH Identifier. The UMH-ID can be locally unique identifier since its has only local significance on the Repair Node. A good ID to use would be the IP address of the interface associated with the UMH the PIM join is sent to. The RNI is carried in the PIM Join as a new PIM Attribute following [RFC5384]. The PIM RNI attribute has the following format for IPv4.



F Forward if not understood.

E End of Attributes, following [RFC5384].

Type: This 6 bit field should be assigned by IANA for TN specific JOIN messages.

Length: Length = 10 octets.

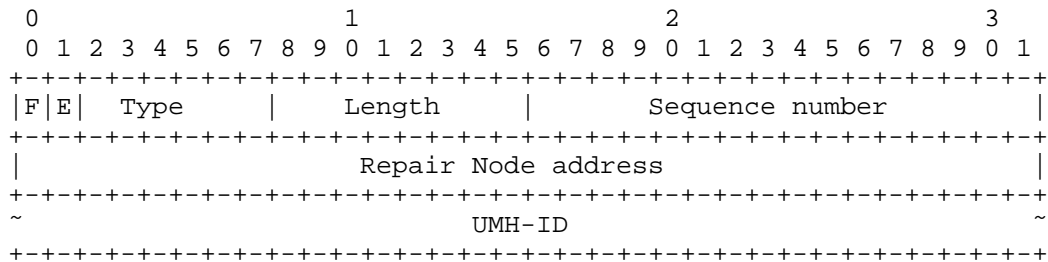
Sequence number: 2 octets long field, describing the sequence number of the sending Repair Node.

Repair Node address: The router ID of the Repair Node, in IPv4 address format.

UMH-ID: This is a 4 octet field encoding UMH identifier. This is the IPv4 address of the interface associated with the UMH the PIM join is sent to.

Figure 3: PIM IPv4 RNI attribute TLV

The PIM RNI attribute has the following format for IPv6.



F Forward if not understood.

E End of Attributes, following [RFC5384].

Type: This 6 bit field should be assigned by IANA for TN specific JOIN messages.

Length: Length = 16 octets.

Sequence number: 2 octets long field, describing the sequence number of the sending Repair Node.

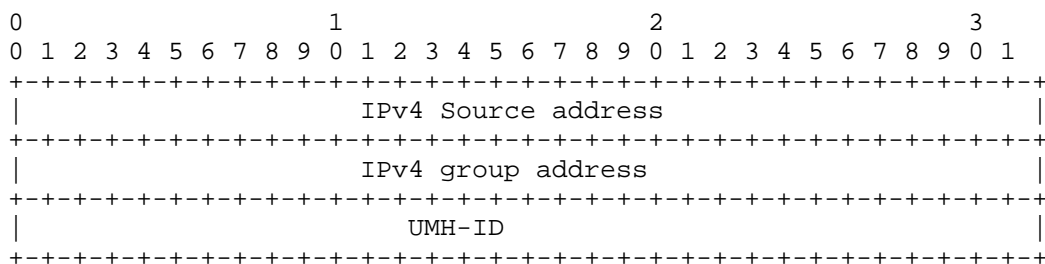
Repair Node address: The router ID of the Repair Node, in IPv4 address format.

UMH-ID: This is a 16 octet field encoding UMH identifier. This is the IPv6 address of the interface associated with the UMH the PIM join is sent to.

Figure 4: PIM IPv6 RNI attribute TLV

## 8.2. Tree Information Item

A TN packet contains one or more TreeInfo items that allows a Merge Node to identify which tree(s) and interface(s) are effected by the TN. The same encoding is used for DTN and UTN packets. The PIM TreeInfo items are defined for IPv4 and IPv6. Which version is to be included in the TN packet depends on Address Family in the TN packet. The UMH-ID included in the DTN MUST be taken from the RNI that was signalled for that tree. The UMH-ID for UTN packets is the PIM neighbor address for that tree. The TreeInfo item has the following format:

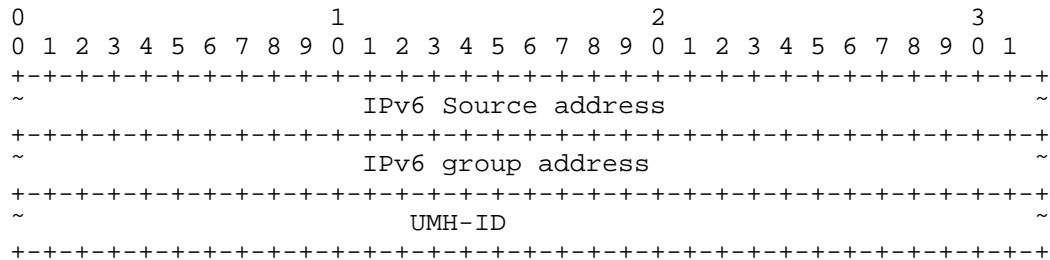


Source Address: This is a 4 octet field encoding the IPv4 source address of the tree. A source address of 0.0.0.0 means that this TN relates to a (\*,G) tree.

Group Address: This is a 4 octet field encoding the IPv4 group address of the tree.

UMH-ID: This is a 4 octet field encoding UMH identifier.

Figure 5: PIM IPv4 TreeInfo item



Source Address: This is a 16 octet field encoding the IPv6 source address of the tree. A source address of 0:0:0:0:0:0:0:0 means that this TN relates to a (\*,G) tree.

Group Address: This is a 16 octet field encoding the IPv6 group address of the tree.

UMH-ID: This is a 16 octet field encoding UMH identifier.

Figure 6: PIM IPv6 TreeInfo item

### 8.3. Incremental deployment

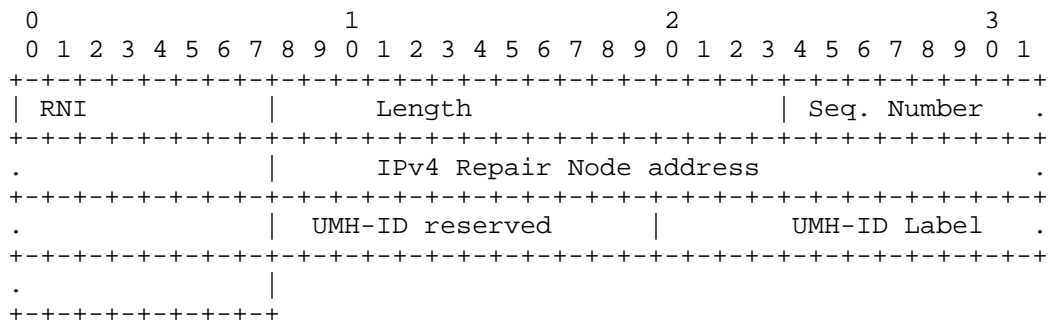
Joins with a RNI can be forwarded through legacy nodes if the Transitive Attribute (see [RFC5384]) has the F bit set to 1. It is up to the network operator to determine this. The DTN functionality can be deployed incrementally as long as the node detecting the failure and Repair Nodes support it.

## 9. mLDP Specific TN Components

In this section we are documenting the mLDP specific data-structures and procedures (if they are different from the generic procedures are defined in this document). As described in this document, TN packets are UDP/IP packets sent via unicast to its destination. The UDP port number for mLDP is set to the (to be) assigned IANA port number for mLDP-TN.

### 9.1. RNI item in mLDP Label Mapping

The RNI item for mLDP is encoded in a LDP MP Status TLV as documented in [RFC6388] section 5. A new LDP MP Status Value Element is created for this purpose and called the RNI Status. The RNI Status includes the router ID, sequence number and UMH Identifier. The UMH-ID can be locally unique identifier since its has only local significance on the Repair node. For mLDP the value that MUST be used is the Local Label associated with the UMH the mLDP Label Mapping is sent to. The RNI status is carried in Label Mapping messages and has the following format.



**RNI Type:** This 1 octet field assigned by IANA for RNI Status Value Element Types.

**Length:** This is a 2 octet field, describing the length of the Value, Length = 10 octets.

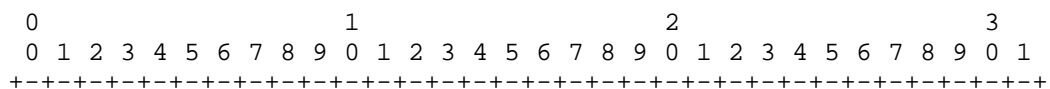
**Sequence number:** 2 octets long field, describing the sequence number of the sending Repair Node.

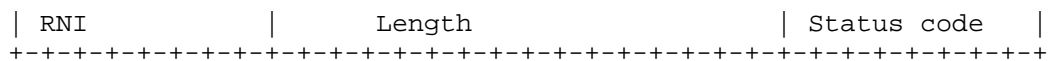
**IPv4 Repair Node address:** The IPv4 address of the Repair Node.

**UMH-ID reserved:** 12 bit field, reserved.

**UMH-ID Label:** This is a 20 bit field encoding a Label as UMH identifier.

Figure 7: mLDP RNI Status Value Element





MBB Type: Type 1 (to be assigned by IANA)

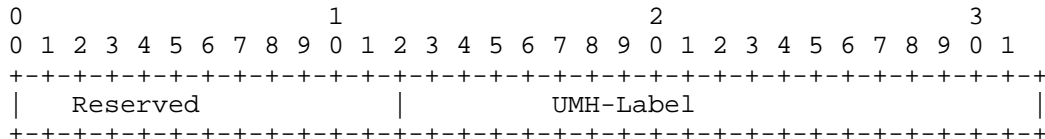
Length: 1

Status code: 1 = MBB request

2 = MBB ack

## 9.2. Tree Information Item

A TN packet contains one or more TreeInfo items that allows a Merge Node to identify which tree(s) and interface(s) are effected by the TN. The same encoding is used for DTN and UTN packets. Following [RFC6388], mLDP will assign a unique Label to each upstream node per MP-LSP. This label identifies the UMH AND the LSP. Since we are using a label to identify the UMH and LSP, there is no need to define a IPv4 and IPv6 specific encoding. The Label included in the DTN MUST be taken from the RNI that was signalled for that tree. The Label for UTN packets is the Local Label that was allocated for that tree. The TreeInfo item has the following format:



Reserved: This is a 12 bits field, set to zero on sending, ignored when received.

UMH-Label: This is a 20 bit field encoding MPLS Label of the UMH.

Figure 8: mLDP TreeInfo item

## 10. Acknowledgements

The authors would like to thank Stefan Olofsson, Javed Asghar and Greg Sheperd for their comments on the draft.



## 11. IANA Considerations

IANA is requested to allocate UDP port numbers to TN messages. One port number for TN in IP/PIM context, and another one for MPLS/mLDP context. The separation of UDP port numbers between IP and MPLS is requested to prevent problems when a PIM multicast tree is transported partly through an mLDP multicast tree.

IANA is requested to allocate a value from "PIM Join Attribute" to make routers capable to advertisement their Tree Notification capability.

IANA is requested to allocate a value from "PIM Join Attribute Types" for TN's join command extra information.

A new IANA registry is needed for "TN option TLVs". This describes the types of TLVs containing extra options for TN messages.

## 12. Security Considerations

Two types of security problems can be foreseen by the authors:

- o Handling illegally injected TN packets
- o Handling replay attacks (re-injecting previous TN messages)
- o TN messages propagating outside an operator's domain

Illegal TN packets can be detected with authentication check. Providing authentication for TN messages is described in Section 6. Prevention of replay attacks needs authentication in combination with sequence numbering, which is also described at the same section.

Preventing TN messages that travel inline with data packets MUST be solved by nodes egressing the operator's domain. Solutions for IP and MPLS are described in sections Section 8 and Section 9, respectively.

## 13. References

### 13.1. Normative References

- [I-D.karan-mofrr]  
Karan, A., Filsfils, C., Farinacci, D., Decraene, B.,  
Leymann, N., and W. Henderickx, "Multicast only Fast Re-  
Route", draft-karan-mofrr-02 (work in progress),

March 2012.

[RFC5384] Boers, A., Wijnands, I., and E. Rosen, "The Protocol Independent Multicast (PIM) Join Attribute Format", RFC 5384, November 2008.

[RFC6388] Wijnands, IJ., Minei, I., Kompella, K., and B. Thomas, "Label Distribution Protocol Extensions for Point-to-Multipoint and Multipoint-to-Multipoint Label Switched Paths", RFC 6388, November 2011.

### 13.2. Informative References

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

### Authors' Addresses

IJsbrand Wijnands (editor)  
Cisco  
De kleetlaan 6a  
Diegem, 1831  
Belgium

Phone:  
Email: ice@cisco.com

Luc De Ghein  
Cisco  
De kleetlaan 6a  
Diegem, 1831  
Belgium

Phone:  
Email: ldeghein@cisco.com

Gabor Sandor Enyedi (editor)  
Ericsson  
Konyves Kalman Krt 11/B  
Budapest, 1097  
Hungary

Phone:  
Email: Gabor.Sandor.Enyedi@ericsson.com

Andras Csaszar  
Ericsson  
Konyves Kalman Krt 11/B  
Budapest, 1097  
Hungary

Phone:  
Email: Andras.Csaszar@ericsson.com

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

Email: Jeff.Tantsura@ericsson.com

