

SPRING Working Group
Internet Draft
Intended status: Standards Track
Expires: October 2017

Dave Allan
Ericsson
Jeff Tantsura

April 2017

A Framework for Computed Multicast applied to MPLS based Segment
Routing
draft-allan-spring-mpls-multicast-framework-03

Abstract

This document describes a multicast solution for Segment Routing with MPLS data plane. It is consistent with the Segment Routing architecture in that an IGP is augmented to distribute information in addition to the link state. In this solution it is multicast group membership information sufficient to synchronize state in a given network domain. Computation is employed to determine the topology of any loosely specified multicast distribution tree.

Status of this Memo

This Internet-Draft is submitted to IETF in full conformance with the provisions of [BCP 78](#) and [BCP 79](#).

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 in October 2017.

Copyright and License Notice

Internet-Draft [draft-allan-spring-mpls-multicast-03](#)

April 2017

Copyright (c) 2017 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. Authors.....	3
1.2. Requirements Language.....	3
2. Conventions used in this document.....	3
2.1. Terminology.....	3
3. Solution Overview.....	4
3.1. Mapping source specific trees onto the segment routing architecture.....	5
3.2. Role of the Routing System.....	6
3.3. MDT Construction Requirements.....	6
3.4. Pruning – theory of operation.....	6
4. Elements of Procedure.....	7
4.1. Triggers for Computation.....	7
4.2. FIB Determination.....	7
4.2.1. Information in the IGP.....	7
4.2.2. Computation of individual segments.....	8
4.3. FIB Generation.....	11
4.4. FIB installation.....	11
5. Related work.....	12
5.1. IGP Extensions.....	12
5.2. BGP Extensions.....	12
6. Observations.....	12
7. Acknowledgements.....	13
8. Security Considerations.....	13
9. IANA Considerations.....	13
10. References.....	13
10.1. Normative References.....	13

10.2. Informative References.....	13
11. Authors' Addresses.....	14

[1. Introduction](#)

This memo describes a solution for multicast for Segment Routing with MPLS data plane in which source specific multicast distribution trees (MDTs) are computed from information distributed via an IGP.

Computation can use information in the IGP to determine if a given node in the network has a role as a root, leaf or replication point in a given MDT. Unicast tunnels are employed to interconnect the nodes determined to have a role. Therefore state only need be installed in nodes that have one of these three roles to fully instantiate an MDT.

Although this approach is computationally intensive, a significant amount of computation can be avoided when the computing agent determines that the node it is computing for has no role in a given MDT. This permits a computed approach to multicast convergence to be computationally tractable.

[1.1. Authors](#)

David Allan, Jeff Tantsura

[1.2. Requirements Language](#)

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

[2. Changes from the last version](#)

Clarifications in the pruning and simplification algorithm

[3. Conventions used in this document](#)

[3.1. Terminology](#)

Candidate replication point (CRP) - is a node that potentially needs to install state to replicate multicast traffic as determined at an intermediate step in multicast segment computation. It will either resolve to having no role or a role as a replication point once

multicast has converged.

Candidate role - refers to any potential combination of roles on a given multicast segment as determined at some intermediate step in MDT computation. For example, a node with a candidate role may be a leaf and may be a candidate replication point.

Downstream - refers to the direction along the shortest path to one or more leaves for a given multicast distribution tree

Multicast convergence - is when all computation and state installation to ensure the FIB reflects the multicast information in the IGP is complete.

MDT - multicast distribution tree. Is a tree composed of one or more multicast segments.

Multicast segment - is a portion of the multicast tree where only the root and the leaves have been specified, and computation based upon the current state of the IGP database is employed to determine and install the required state to implement the segment. For MPLS a multicast segment is implemented as a p2mp LSP. A multicast segment is identified by a multicast SID.

Multicast SID - Is the data plane identifier that is used to implement a multicast segment. As per a unicast MPLS segment, the rightmost 20 bits of a multicast SID is encoded as a label. It is drawn from an SRGB that is global to the SR domain.

Pinned path - Is a unique shortest path extending from a leaf upstream towards the root for a given multicast segment. Therefore is a component of the multicast segment that it has been determined must be there. It will not necessarily extend from the leaf all the way to the root during intermediate computation steps. A pinned path can result from pruning operations.

Role - refers specifically to a node that is either a root, a leaf, a replication node, or a pinned waypoint for a given MDT.

Unicast convergence - is when all computation and state installation to ensure the FIB reflects the unicast information in the IGP is

complete.

Upstream - refers to the direction along the shortest path to the root of a given MDT.

[4. Solution Overview](#)

This memo describes a multicast architecture in which multicast state is only installed in those nodes that have roles as a root, leaves, and replication points for a given multicast segment. The a-priori established segment routing unicast tunnels are used as interconnect between the nodes that have a role in a given multicast SID.

Allan et al.,

Expires October 2017

[Page 4]

Internet-Draft [draft-allan-spring-mpls-multicast-03](#)

April 2017

A loosely specified MDT is composed of a single multicast segment and the routing of the MDT is delegated entirely to computation driven by information in the IGP database.

Explicitly routed MDTs are expressed as a tree of concatenated multicast segments where both the leaves of each segment and the waypoints coupling a given segment to the upstream and/or downstream segment(s) is specified in information flooded in the IGP by the overall root of the MDT. The segments themselves will be computed as per a loosely specified MDT.

A PE acting as an overall root for a given tree is expected to be configured by the operator as to where to source multicast traffic from, be it an attachment circuit, interworking function for client technology or other. Similarly a leaf for a given tree is expected to be configured by the operator as to the disposition of received multicast traffic.

A computed segment is guaranteed to be loop free in a stable system. A concatenation of segments to construct an MDT will similarly be loop free as any collision of segments can be disambiguated in the data plane via the SIDs.

This architecture significantly reduces the amount of state that needs to be installed in the data plane to support multicast. This also means that the impact of many failures in the network on multicast traffic distribution will be recovered by unicast local repair or unicast convergence with subsequent multicast convergence acting in the role of network re-optimization (as opposed to

restoration).

[4.1.](#) Mapping source specific trees onto the segment routing architecture

A computed source specific tree for a given multicast group corresponds to one or more multicast segments in the SR architecture. Each multicast segment is assigned a SID, typically by management configuration of the node that will be the overall root for the source specific tree. The root node then uses the IGP to advertise this information to all nodes in the IGP area/domain.

A multicast group is implemented as the set of source specific trees from all nodes that have registered transmit interest to all nodes that have registered receive interest in a multicast group.

[4.2.](#) Role of the Routing System

The role of the IGP is to communicate topology information, multicast capability and associated algorithm, multicast registrations, unicast to SID bindings, multicast to SID bindings and waypoints in multi-segment MDTs. No changes to topology or unicast to SID binding advertisements are proposed by this memo.

The multicast registrations/bindings will be in the form of source, group, transmit/receive interest and the SID to use for the source specific multicast tree. Registrations are originated by any node that has send or receive interest in a given multicast group. Nodes will use the combination of topology and multicast registrations to determine the nodes that have a role in each source specific tree and the SID information to then derive the required FIB state.

[4.3.](#) MDT Construction Requirements

A multicast segment in an MDT is constructed such that between any pair of nodes that have a role in the segment and are connected by a unicast tunnel, there is not another node on the shortest path between the two with a role in that segment. This ensures that copies of a packet forwarded by an multicast segment will traverse a link only once in a stable system.

Note that this can be satisfied by a minimum cost shortest path tree, but is not an absolute requirement. The pruning rules specified in this memo will meet this requirement without necessarily producing absolutely minimum cost multicast segment (or incurring the associated computational cost).

4.4. Pruning - theory of operation

The role of nodes in a given multicast segment is determined by first producing an inclusive shortest path tree with all possible paths between the root and leaves, and then applying a set of pruning rules repeatedly until an acyclic tree is produced or no further prunes are possible.

For the majority of multicast segments these rules will authoritatively produce a minimum cost tree. For those segments that have not yet been authoritatively resolved, there is a set of pruning operations applied that are not guaranteed to produce a tree that meets the requirements of 3.3, therefore these trees require auditing and potential correction according to a further set of agreed rules. This avoids the necessity of an exhaustive search of the solution space.

A node during computation of a segment may conclude that it will absolutely not have a role at any of numerous points in the computation process and abandon computation of that segment.

5. Elements of Procedure

5.1. Triggers for Computation

MDT computation is triggered by changes to the IGP database. These are in the form of either changes in registered multicast group interest, addition or removal of a multi-segment MDT descriptor, or topology changes.

A change in registered interest for a group will require re-computation of all MDTs that implement the multicast group.

A topology change will require the computation of some number of multicast segments, the actual number will depend on the implementation of tree computation but at a minimum will be all trees

for which there is not an optimal shortest path solution as a result of the topology change.

[5.2.](#) FIB Determination

[5.2.1.](#) Information in the IGP

Group membership information for a multicast segment is obtained from the IGP. This is true for single segment MDTs as well as multi-segment MDTs. Included in the multi-segment MDT specification is the waypoint nodes in MDT and the upstream and downstream SIDs. The specified node is expected to cross connect the SIDs to join the segments together acting in the role of leaf for the upstream segment and root for the downstream segment.

When a waypoint in an MDT descriptor does not exist in the IGP, the assumption is that the node identified by the waypoint SID has failed. The response of the other nodes in the system in FIB determination is to add the leaves of the downstream segment to the upstream segment.

An example of this would be consider a node "x", and another node "y". At some point in time, "x" advertises a tree that identifies "y" as a waypoint that cross connects upstream SID "a" to downstream SID "b". At some later point node "y" fails. The other nodes in the network will compute segment "a" as if it included all leaves and waypoints in segment "b". All apriori state installed for segment "b"

would be removed as the failure of "y" has required "b" to be subsumed by "a".

[5.2.2.](#) Computation of individual segments

FIB generation for a multicast segment is the result of computation, ultimately as applied to all source specific trees in the network. All computing nodes implement a common algorithm for tree generation, as all MUST agree on the solution.

One algorithm is as follows:

All possible shortest paths to the set of leaves for the MDT is determined. Then pruning rules are repeatedly applied until no

further prunes are possible.

The philosophy of the application of these rules could be expressed as "simplify as much as possible, and prune that which cannot be". The rules are:

- 1) Eliminate any links and nodes not on a potential shortest path from the root to the leaves for the MDT under consideration.
- 2) Simplify via the replacement of any nodes that do not have a potential role in the MDT with links.

This will be nodes that are not a leaf, a root or a candidate replication point. For example:

Root-----A-----B

B is a leaf. A is not but is in a potential shortest path from root to B. However A will have no role in the MDT that serves B as it provides simple transit therefore is replaced with a direct connection between the root and B.

Root-----B

Note that such pruning also needs to avoid the creation of duplicate parallel links. For example:

```
      /-----A-----\  
Root                                B  
      \-----C-----/
```

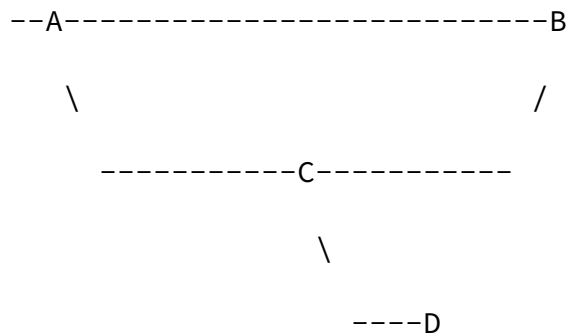
Where A and C have no role and the cost root-A-B = cost root-C-B, they can be replaced with a single link from Root to B.

- 3) Simplify via the elimination of fewer hop paths

When for a given set of leaves, a node has multiple downstream links that converge on a common downstream point, and that set of leaves is only a subset of the leaves reachable on one or more of the links, any link that only serves that subset of leaves can be

pruned.

For example:



Link AB is cost 2, link AC and CB are cost 1 (cost of link CD does not affect the example).

B and D are leaves of a root upstream of A. From A, link AB can reach leaf B. Path AC can reach leaf B and D. In this case path A-B can be pruned from consideration. The set of leaves reachable via link A-B is a subset of that reachable by A-C, and the paths from A that serves that subset converges at B.

4) Prune upstream links.

The normal procedure is to determine the closest upstream leaf or pinned path and then compare all upstream adjacencies with that metric

- a. If the upstream adjacency extends closer to the root than the closest leaf or pinned path, then that adjacency can be pruned.
- b. If the upstream adjacency extends the same distance towards the root then
 - i. If it is to a non-leaf or pinned path candidate replication point, it can be pruned

- ii. If it is to a pinned path, where there are equal upstream adjacencies that terminate on leaves, it can be pruned (considered inferior).

- iii. If there is more than one "equal" upstream adjacency, that is all terminate on nodes that are on pinned paths, or all terminate on nodes that are leaves, than one is selected. This is via the lowest node ID.
- c. If the upstream adjacency is a candidate replication point closer than the closest leaf, and upstream from it is a node that is a leaf or pinned path equidistant with the closest leaf, then all adjacencies that extend to leaves ranked lower than the leaf or pinned path behind the CRP may be pruned. Note that an upstream adjacency that has a CRP closer than the closest leaf or pinned path cannot be pruned.
- d. When for a given node all possible upstream adjacencies that can be pruned have been identified, each is removed, and any simplifications that can be performed as a result of the prune are performed. This is the equivalent of a localized check for 2 and 3 above and is then performed iteratively in response to changes to the graph as a result of pruning.

The procedure is to implement 1, 2 and 3 above, then loop on 4 until such time as the MDT is fully resolved, or no further prunes are possible. Step 4 is performed in a specific order. The nodes are processed according to a ranking from closest to the root to the farthest, and from lowest node ID to the highest within a given distance from the root.

If, at the end of pruning and simplification, all leaves in a multicast segment have a unique shortest path to the root, the tree is considered resolved, and the computation can progress directly to the FIB generation step.

If not all leaves have a unique shortest path, additional pruning steps are applied. These steps are NOT guaranteed to produce a lowest cost tree, and therefore require an additional audit and possible modification to ensure when forwarding a maximum of one copy of a packet will traverse an interface.

For segments not authoritatively resolved by the above rules, a prune that will not authoritatively result in a minimum cost tree is applied. For the purpose of interoperability, the following rule is proposed: A computing node will select the closest node to the root with a candidate role that does not have a unique shortest path to

the root. Where more than one such node exists, the one with the lowest unicast SID is selected. For that node, the best upstream link is selected and all other upstream links pruned. The best upstream link is defined as the link with the closest node with a candidate role that potentially serves the highest number of leaves. Where there is a tie, once again the node with the lowest SID is selected.

Once the links have been pruned, rules 2 through 4 are repeatedly applied until either the tree is fully resolved, or again no further prunes are possible, in which case the next closest remaining unresolved node has the same prune applied.

For all segments not resolved by the initial prune rules, they are audited to ensure all nodes that have a role in the tree do not have a node with a role between them and their upstream node on the tree. If they do, the old upstream adjacency is removed, and the superior one added.

[5.3.](#) FIB Generation

The topology components that remain at the end of the pruning operation will reflect all nodes that have a role in a given multicast segment plus the necessary tunnels (as all intervening multi-path scenarios will have been simplified away). From this the FIB can be generated:

All nodes that have a role in a given multicast segment and have nodes upstream in the segment will need to accept the SID for the MDT from at minimum, all upstream interfaces.

All nodes that have a role in a given segment and have nodes immediately downstream in the segment will need to replicate packets simply labelled with the multicast SID onto those interfaces.

All nodes that have a role in a given segment and have nodes reachable via a tunnel downstream set the FIB to push the tunnel unicast SID for the downstream node onto any replicated copies of a received packet, and identify the set of interfaces on the shortest path for the tunnel SID.

[5.4.](#) FIB installation

FIB installation needs to acknowledge two aspects of the hybrid tunnel and role model of multicast tree construction. The first is that because of the sparse state model simple tree adds, moves, and changes may require the installation of state where it did not previously exist, and such changes may impact existing services. The

second is that it is possible to retain the knowledge to prioritize computation of those trees impacted the failure of a node with a role.

To address this, there are three stages of state installation for multicast convergence:

1) Immediate:

- a. Installation of state for multicast segments impacted by the failure of a node in the network, and installation of state for segments in nodes that have not previously had a role in the given segment.

- b. Installation of state for waypoints in multi-segment MDTs.

2) After T1: Update state for nodes that both had and have a role in a given multicast segment.

3) After T2: Removal of state for nodes that transition from having a role to not having a role for a given multicast segment.

T1 and T2 are network wide configurable values.

[6.](#) Related work

[6.1.](#) IGP Extensions

The required IGP changes are documented in [[MCAST-ISIS](#)] and [[MCAST-OPSF](#)].

[6.2.](#) BGP Extensions

This memo will require the specification of a new PMSI Tunnel Attribute (SPRING P2MP tunnel, tentatively 0x09) to order to integrate into the multicast framework documented in [RFC 6514](#)

[7.](#) Observations

This technique is not confined to segment routing, and with the provision of a global label space (to be employed as per a multicast SID), an MPLS-LDP network would also provide the requisite mesh of unicast tunnels and be capable of implementing this approach to multicast.

This memo focuses on an implementation based upon nodes that are IGP

speakers and converge independently so is written in a form that

assumes a node, computing node and IGP speaker are one in the same. It should be observed that the relative frugality of data plane state would suggest that separation of computation from nodes in the data plane combined with management or "software defined networking" based population of the multicast FIB entries may also be useful modes of network operation.

[8.](#) Acknowledgements

Thanks to Uma Chunduri for his detailed review and suggestions.

[9.](#) Security Considerations

For a future version of this document.

[10.](#) IANA Considerations

This document requires the allocation of a PMSI tunnel type to identify a SPRING P2MP tunnel type from the P-Multicast Service Interface Tunnel (PMSI Tunnel) Tunnel Types registry.

[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.

[11.2.](#) Informative References

[MCAST-ISIS] Allan et.al., "IS-IS extensions for Computed Multicast applied to MPLS based Segment Routing", IETF work in progress, [draft-allan-isis-spring-multicast-00](#), July 2016

[MCAST-OSPF] Allan et.al., "OSPF extensions for Computed Multicast applied to MPLS based Segment Routing", IETF work in progress, [draft-allan-ospf-spring-multicast-00](#), July 2016

[RFC6514] Aggarwal et.al., "BGP Encodings and Procedures for Multicast in MPLS/BGP IP VPNs", IETF [RFC 6514](#), February 2012

[RFC7385] Andersson & Swallow "IANA Registry for P-Multicast Service Interface (PMSI) Tunnel Type Code Points", IETF [RFC 7385](#), October 2014

Allan et al., Expires October 2017 [Page 13]

Internet-Draft [draft-allan-spring-mpls-multicast-03](#) April 2017

[12.](#) Authors' Addresses

Dave Allan (editor)
Ericsson
300 Holger Way
San Jose, CA 95134
USA
Email: david.i.allan@ericsson.com

Jeff Tantsura
Email: jefftant.ietf@gmail.com

Allan et al.,

Expires October 2017

[Page 14]