PIM Working Group
Internet Draft
Intended status: Standards Track
Expires: December 1, 2018

Dave Allan
Ericsson
Jeff Tantsura
Nuage
Ian Duncan
Ciena

June 1, 2018

A Framework for Computed Multicast Applied to SR-MPLS
draft-allan-pim-sr-mpls-multicast-framework-00

Abstract

This document describes a multicast solution for SR-MPLS. 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.

Table of Contents

1. Introduction

This memo describes a solution for multicast for SR-MPLS in which source specific multicast distribution trees (MDTs) are computed from information distributed via an IGP. Computation uses information in the IGP to determine if a given node in the network has a role as a root, a leaf or replication point in a given MDT. Unicast tunnels are employed to interconnect the nodes determined to have a role. Therefore multicast topological instructions only need be installed in nodes that have one of these three roles to fully instantiate an MDT.

Although this approach might appear to be computationally intensive, a significant amount of computation can be avoided if and when the computing agent determines that the node it is computing for has no role in a given MDT. If there will be no need to install a multicast topological instruction in that node for the given MDT, the computing agent can abandon computation for the MDT and move on to other tasks, such as converging other MDTs. This permits a computed approach to multicast convergence to be computationally tractable.

This approach is proposed as a solution for networks for which an implementation of an alternative data plane, such as BIER, offers technical or economic challenges.

1.1. Authors

David Allan, Jeff Tantsura, Ian Duncan

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

Clarification in motivation.

Editorial corrections and improvements.

Clarification of the description of upstream pruning in section 5.2.2

Alignment of terminology with current segment routing practice.

3. Conventions used in this document

3.1. Terminology

Candidate replication point (CRP) - is a node that potentially needs to install a multicast topological instruction 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 also be a candidate replication point.

Computing agent- refers to the agent that will compute the FIB for the MDTs in a given network on behalf of one node (distributed model) or multiple nodes (SR controller(s) in a centralized model).

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 multicast topological instruction 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 topological instructions to implement the segment. For SR-MPLS a multicast segment is implemented as a p2mp LSP. A multicast segment is identified by a multicast SID.

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

Pinned path - Is a unique shortest path extending from a leaf upstream towards the root for a given multicast segment. Therefore, it 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 topological instruction 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 topological instructions are only installed in those nodes that have roles as a root, a leaf, or a replication point for a given multicast segment. The a-priori established mesh of unicast tunnels (using node-SIDs) are used as interconnect between the nodes that have a role in a given multicast SID. Hence on an outgoing interface where the next node in that path of the MDT is not immediately adjacent, the operation will typically be a CONTINUE of the multicast SID and a PUSH of the node-SID.

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 fault free 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 number of multicast topological instructions 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 node-SID bindings, multicast to SID bindings and waypoints in multi-segment MDTs. No changes to topology or unicast to node-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 a multicast segment will traverse a link only once in a stable system and avoids the potential scenario whereby a packet needs to be replicated twice on a given interface.

Note that this can be satisfied by a minimum cost shortest path tree, but this is not an absolute requirement. The pruning rules specified

in this memo will meet this requirement without necessarily producing an absolute minimum cost multicast segment (or incurring the associated computational cost).

4.4. Simplification and 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 simplification and pruning rules repeatedly until either 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 are not able to be 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 and computational overhead of an exhaustive search of the solution space.

A computing agent during computation of a segment may conclude that none of the nodes that it is computing on behalf of will have a role at any point 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 agents in a given network computing a tree for a given multicast segment must 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 simplification and pruning rules are repeatedly applied until no further prunes are possible or the MDT is determined to be resolved.

The distinction between simplification rules and pruning rules is the former will not change the candidate role of a node with respect to the MDT under consideration and therefore can be performed in any order, while the latter will affect candidate node roles and must be

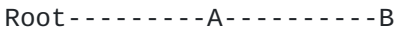
performed in an agreed order between all participating computing agents.

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:

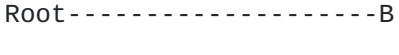
1) Simplification: Eliminate any links and nodes not on a potential shortest path from the root to the leaves for the MDT under consideration.

2) Simplification: Replace any nodes that do not have a potential role in the MDT with links.

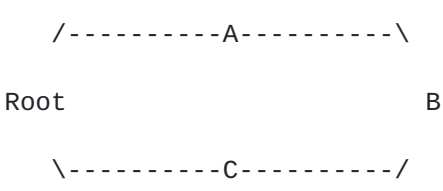
This will be nodes that are not a leaf, a root or a candidate replication point. For example:



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.



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

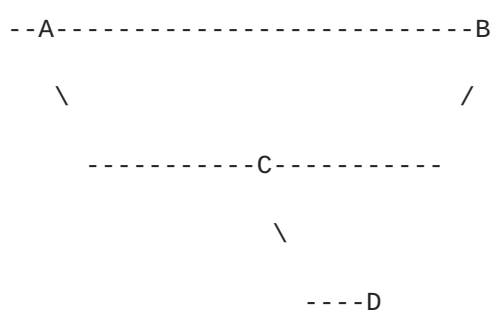


Where A and C have no role and the cost root-A-B = cost root-C-B, they can be replaced with a single link from Root to B.

3) Simplification: Eliminate 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 eliminated.

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 eliminated 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 best-closest upstream leaf or pinned path and then compare all upstream adjacencies with that metric. Note that the best-closest upstream leaf or pinned path may not be directly connected to the node under consideration. Where there is more than one equally close upstream leaf or pinned path, the highest ranked is selected with the ranking being that a leaf is ranked superior to a pinned path, and the lowest unicast SID is selected when the leaf/pinned path ranking is equal.

Then examine each of the remaining upstream adjacencies:

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 as the best-closest adjacency, then it can be eliminated as it has already been ranked lower than the best-closest adjacency. Note that this would include non-leaf and non-pinned path candidate replication points.

c. If the upstream adjacency is a candidate replication point closer than the best-closest leaf or pinned path, then it is left alone.

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 all simplifications of type 1, 2 and 3 above, then loop on type 4 prunes until such time as the MDT is fully resolved from the point of view of the node under consideration, or no further prunes are possible. Step 4 is required to be performed in a specific order if there is more than one computing agent generating topological instructions for a given multicast segment. This memo suggests that the nodes are processed according to a ranking of nodes from closest to the root to the farthest, and from lowest unicast SID to the highest within a given distance from the root.

At the end of pruning and simplification, either:

1) The node whom the computing agent is computing for has no role in the multicast segment under consideration

2) A unique shortest path to the root has been determined for all leaves in the multicast segment that are downstream of the node under consideration (also termed as a pinned path from the root to every leaf).

3) A unique shortest path to the root has not been determined for all leaves downstream of the node under consideration in the multicast segment.

If 1 or 2 then the multicast segment is considered to be resolved, and for 2, the computation can progress directly to the topological instruction generation step for that segment.

If 3 (not all downstream 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 applied: A computing agent 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 node-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 unicast 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 simplification and pruning operations 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 topological instructions to put in 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 multicast 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 topological instructions where they 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 in the distributed model, there are three stages of topological instruction installation for multicast convergence:

1) Immediate:

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

   b. Installation of topological instructions for waypoints in multi-segment MDTs.

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

3) After T2: Removal of topological instructions 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.

When an SR-Controller is used, it is only necessary to properly sequence the installation of state. This also suggests that when there is more than one SR-Controller, the division of responsibility should be on the basis of MDT ownership.

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 0x0c) to order to integrate into the multicast framework documented in RFC 6514

7. Observations

This technique is not confined to SR-MPLS:

- 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.

- It is also possible to envision an SRv6 implementation but would require the ability to rewrite the SRH at each hop.

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 agent 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

[SR-ARCH] Filsfils et.al., "Segment Routing Architecture", IETF work in progress, draft-ietf-spring-segment-routing-15, January 2018

[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

12. Authors' Addresses

Dave Allan (editor)
Ericsson
2455 Augustine Drive
Santa Clara 95054
USA
Email: david.i.allan@ericsson.com

Jeff Tantsura
Email: jefftant.ietf@gmail.com

Ian Duncan
Ciena
iduncan@ciena.com
5050 Innovation Drive
Kanata, ON K2K 0J2