

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: July 10, 2020

A. Przygienda
C. Bowers
Juniper
Y. Lee
A. Sharma
Comcast
R. White
Juniper
January 7, 2020

IS-IS Flood Reflection
draft-przygienda-lsr-flood-reflection-01

Abstract

This document describes an optional ISIS extension that allows the creation of IS-IS flood reflection topologies. Flood reflection allows the creation of topologies where L1 areas provide transit forwarding for L2 destinations within an L2 topology. It accomplishes this by creating L2 flood reflection adjacencies within each L1 area. The L2 flood reflection adjacencies are used to flood L2 LSPDUs, and they are used in the L2 SPF computation. However, they are not used for forwarding. This arrangement gives the L2 topology better scaling properties. In addition, only those routers directly participating in flood reflection have to support the feature. This allows for the incremental deployment of scalable L1 transit areas in an existing network, without the necessity of upgrading other routers in the network.

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 <https://datatracker.ietf.org/drafts/current/>.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any

time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on July 10, 2020.

Copyright Notice

Copyright (c) 2020 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 (<https://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. Description	2
2. Further Details	8
3. Flood Reflection TLV	8
4. Flood Reflection Discovery Sub-TLV	10
5. Flood Reflection Adjacency Sub-TLV	10
6. Flood Reflection Discovery	11
7. Flood Reflection Adjacency Formation	12
8. Redistribution of Prefixes	12
9. Route Computation	13
10. Special Considerations	13
11. IANA Considerations	14
11.1. New IS-IS TLV Codepoint	14
11.2. Sub TLVs for TLV 242	14
11.3. Sub TLVs for TLV 22, 23, 25, 141, 222, and 223	15
12. Security Considerations	15
13. Acknowledgements	15
14. References	15
14.1. Informative References	15
14.2. Normative References	15
Authors' Addresses	16

1. Description

Due to the inherent properties of link-state protocols the number of IS-IS routers within a flooding domain is limited by processing and flooding overhead on each node. While that number can be maximized

by well written implementations and techniques such as exponential back-offs, IS-IS will still reach a saturation point where no further routers can be added to a single flooding domain. In some L2 backbone deployment scenarios, this limit presents a significant challenge.

The traditional approach to increasing the scale of an IS-IS deployment is to break it up into multiple L1 flooding domains and a single L2 backbone. This works well for designs where an L2 backbone connects L1 access topologies, but it is limiting where a large L2 is supposed to span large number of routers. In such scenarios, an alternative approach is to consider multiple L2 flooding domains connected together via L1 flooding domains. In other words, L2 flooding domains are connected by "L1/L2 lanes" through the L1 areas to form a single L2 backbone again. Unfortunately, in its simplest implementation, this requires the inclusion of most, or all, of the transit L1 routers as L1/L2 to allow traffic to flow along optimal paths through such transit areas. Consequently, this approach fails to reduce the number of L2 routers involved, so it fails to increase the scalability of the L2 backbone.

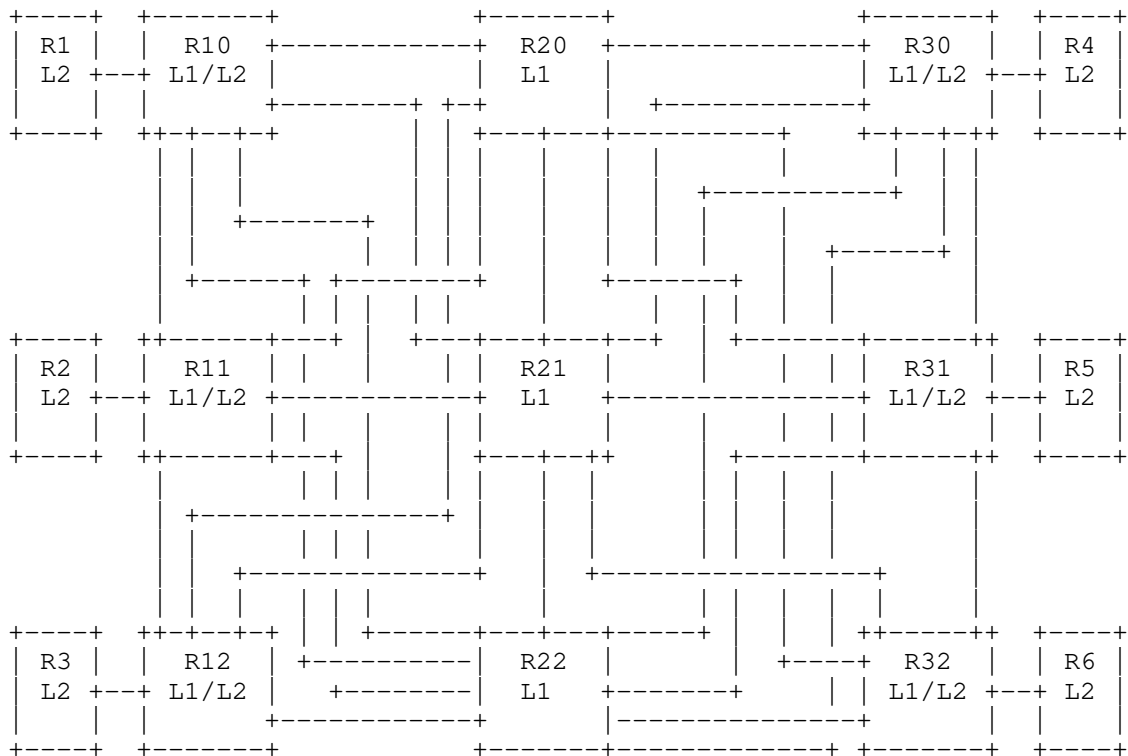


Figure 1: Example topology

Figure 1 is an example of a network where a topologically rich L1 area is used to provide transit between six different L2-only routers (R1-R6). Note that the six L2-only routers do not have connectivity to one another over L2 links. To take advantage of the abundance of paths in the L1 transit area, all the intermediate systems could be placed into both L1 and L2, but this essentially combines the separate L2 flooding domains into a single one, triggering again maximum L2 scale limitation we try to address in first place.

A more effective solution would allow to reduce the number of links and routers exposed in L2, while still utilizing the full L1 topology when forwarding through the network.

[RFC8099] describes Topology Transparent Zones (TTZ) for OSPF. The TTZ mechanism represents a group of OSPF routers as a full mesh of adjacencies between the routers at the edge of the group. A similar mechanism could be applied to ISIS as well. However, a full mesh of adjacencies between edge routers (or L1/L2 nodes) significantly

limits the scale of the topology. The topology in Figure 1 has 6 L1/L2 nodes. Figure 2 illustrates a full mesh of L2 adjacencies between the 6 L1/L2 nodes, resulting in $(5 * 6)/2 = 15$ L2 adjacencies. In a somewhat larger topology containing 20 L1/L2 nodes, the number of L2 adjacencies in a full mesh rises to 190.

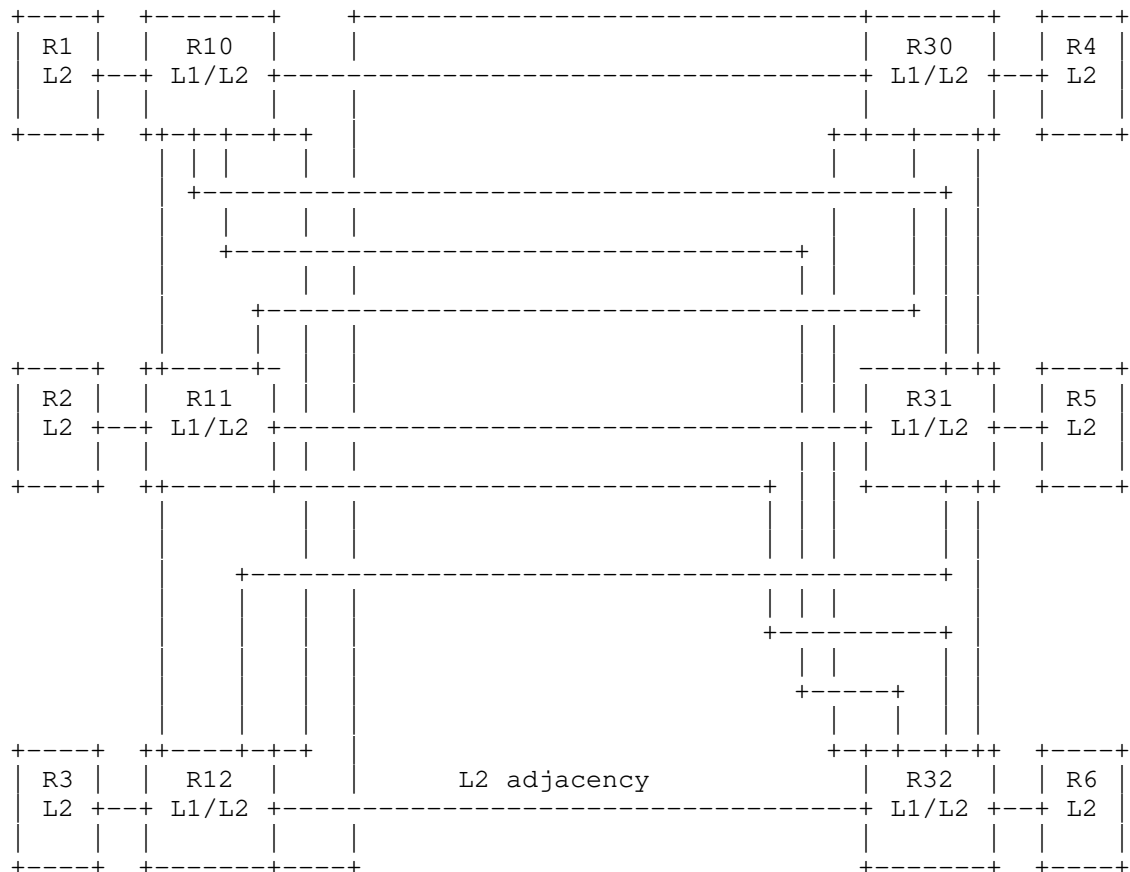


Figure 2: Example topology represented in L2 with a full mesh of L2 adjacencies between L1/L2 nodes

BGP, as specified in [RFC4271], faced a similar scaling problem, which has been solved in many networks by deploying BGP route reflectors [RFC4456]. We note that BGP route reflectors do not necessarily have to be in the forwarding path of the traffic. This incongruity of forwarding and control path for BGP route reflectors

allows the control plane to scale independently of the forwarding plane.

We propose here a similar solution for IS-IS. A simple example of what a flood reflector control plane approach would look like is shown in Figure 3, where router R21 plays the role of a flood reflector. Each L1/L2 ingress/egress router builds a tunnel to the flood reflector, and an L2 adjacency is built over each tunnel. In this solution, we need only 6 L2 adjacencies, instead of the 15 needed for a full mesh. In a somewhat larger topology containing 20 L1/L2 nodes, this solution requires only 20 L2 adjacencies, instead of the 190 need for a full mesh. Multiple flood reflectors can be used, allowing the network operator to balance between resilience, path utilization, and state in the control plane. The resulting L2 adjacency scale is $R \cdot n$, where R is the number of flood reflectors used and n is the number of L1/L2 nodes. This compares quite favorably with $n \cdot (n-1)/2$ L2 adjacencies required in a fully meshed L2 solution.

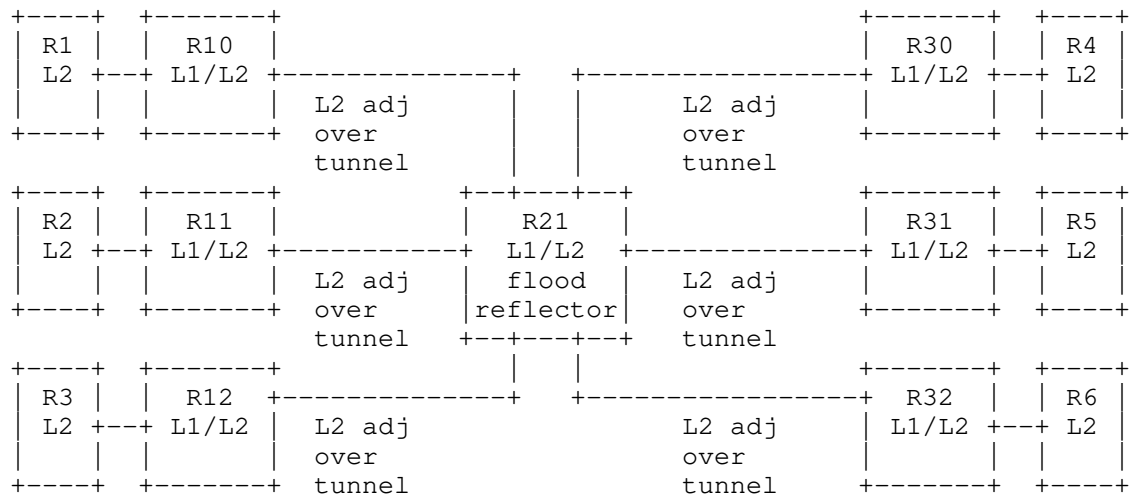


Figure 3: Example topology represented in L2 with L2 adjacencies from each L1/L2 node to a single flood reflector

As illustrated in Figure 3, when R21 plays the role of flood reflector, it provides L2 connectivity among all of the previously disconnected L2 islands by refloding all L2 LSPDUs. At the same time, R20 and R22 remain L1-only routers. L1-only routers and L1-only links are not visible in L2. In this manner, the flood

reflector allows us provide L2 control plane connectivity in a scalable manner.

As described so far, the solution illustrated in Figure 3 relies only on currently standardized ISIS functionality. Without new functionality, however, the data traffic will traverse only R21. This will unnecessarily create a bottleneck at R21 since there is still available capacity in the paths crossing the L1-only routers R20 and R22.

Hence, some new functionality is necessary to allow the L1/L2 edge nodes (R10-12 and R30-32 in Figure 3) to recognize that the L2 adjacency to R21 should not be used for forwarding. The L1/L2 edge nodes should forward traffic that would normally be forwarded over the L2 adjacency to R21 over L1 links instead. This would allow the forwarding within the L1 area to use the L1-only nodes and links shown in Figure 1 as well. It allows networks to be built that use the entire forwarding capacity of the L1 areas, while at the same time introducing control plane scaling benefits provided by L2 flood reflectors.

This document defines all extensions necessary to support flood reflector deployment:

- o A 'flood reflector adjacency' for all the adjacencies built for the purpose of reflecting flooding information. This allows these 'flood reflectors' to participate in the IS-IS control plane without being used in the forwarding plane. This is a purely local operation on the L1/L2 ingress; it does not require replacing or modifying any routers not involved in the reflection process. Deployment-wise, it is far less tricky to just upgrade the routers involved in flood reflection rather than have a flag day on the whole ISIS domain.
- o A full mesh of L1 tunnels between the L1/L2 routers, ideally load-balancing across all available L1 links. This harnesses all forwarding paths between the L1/L2 edge nodes without injecting unneeded state into the L2 flooding domain or creating 'choke points' at the 'flood reflectors' themselves. A solution without tunnels is also possible by judicious scoping of reachability information between the levels.
- o Some way to support reflector redundancy, and potentially some way to auto-discover and advertise such adjacencies as flood reflector adjacencies. Such advertisements may allow L2 nodes outside the L1 to perform optimizations in the future based on this information.

2. Further Details

Several considerations should be noted in relation to such a flood reflection mechanism.

First, this allows multi-area IS-IS deployments to scale without any major modifications in the IS-IS implementation on most of the nodes deployed in the network. Unmodified (traditional) L2 routers will compute reachability across the transit L1 area using the flood reflector adjacencies.

Second, the flood reflectors are not required to participate in forwarding traffic through the L1 transit area. These flood reflectors can be hosted on virtual devices outside the forwarding topology.

Third, astute readers will realize that flooding reflection may cause the use of suboptimal paths. This is similar to the BGP route reflection suboptimal routing problem described in [ID.draft-ietf-idr-bgp-optimal-route-reflection-19]. The L2 computation determines the egress L1/L2 and with that can create illusions of ECMP where there is none. And in certain scenarios lead to an L1/L2 egress which is not globally optimal. This represents a straightforward instance of the trade-off between the amount of control plane state and the optimal use of paths through the network often encountered when aggregating routing information.

One possible solution to this problem is to expose additional topology information into the L2 flooding domains. In the example network given, links from router 01 to router 02 can be exposed into L2 even when 01 and 02 are participating in flood reflection. This information would allow the L2 nodes to build 'shortcuts' when the L2 flood reflected part of the topology looks more expensive to cross distance wise.

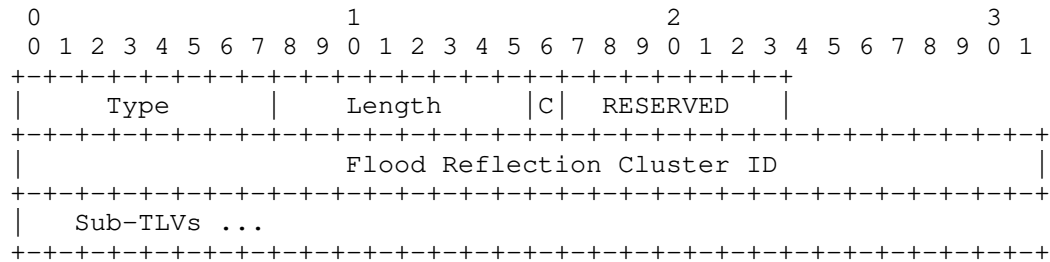
Another possible variation is for an implementation to approximate with the L1 tunnel cost the cost of the underlying topology.

Redundancy can be achieved by building multiple flood reflectors in the L1 area. Multiple flood reflectors do not need any synchronization mechanisms amongst themselves, except standard ISIS flooding and database maintenance procedures.

3. Flood Reflection TLV

The Flood Reflection TLV is a new top-level TLV that MAY appear in IIHs. The Flood Reflection TLV indicates the flood reflector cluster (based on Flood Reflection Cluster ID) that a given router is

configured to participate in. It also indicates whether the router is configured to play the role of either flood reflector or flood reflector client. The Flood Reflection Cluster ID and flood reflector roles advertised in the IIHs are used to ensure that flood reflector adjacencies are only formed between a flood reflector and flood reflector client, and that the Flood Reflection Cluster IDs match. The Flood Reflection TLV has the following format:



Type: TBD

Length: The length, in octets, of the following fields.

C (Client): This bit is set to indicate that the router acts as a flood reflector client. When this bit is NOT set, the router acts as a flood reflector. On a given router, the same value of the C-bit MUST be advertised across all interfaces advertising the Flood Reflection TLV in IIHs.

RESERVED: This field is reserved for future use. It MUST be set to 0 when sent and MUST be ignored when received.

Flood Reflection Cluster ID: Flood Reflection Cluster Identifier. These same 32-bit value MUST be assigned to all of the flood reflectors and flood reflector clients in the L1 area. The value MUST be unique across different L1 areas within the IGP domain. On a given router, the same value of the Flood Reflection Cluster ID MUST be advertised across all interfaces advertising the Flood Reflection TLV in IIHs.

Sub-TLVs: Optional sub-TLVs. For future extensibility, the format of the Flood Reflection TLV allows for the possibility of including optional sub-TLVs. No sub-TLVs of the Flood Reflection TLV are defined in this document.

The Flood Reflection TLV MUST NOT appear more than once in an IIH. A router receiving multiple Flood Reflection TLVs in the same IIH SHOULD use the values in the first TLV.

4. Flood Reflection Discovery Sub-TLV

Flood Reflection Discovery sub-TLV is advertised as a sub-TLV of the IS-IS Router Capability TLV-242, defined in [RFC7981]. The Flood Reflection Discovery sub-TLV is advertised in L1 LSPs with area flooding scope in order to enable the auto-discovery of flood reflection capabilities and the automatic creation of L2 tunnels to be used as flood reflector adjacencies. The Flood Reflection Discovery sub-TLV has the following format:

```

      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      |      Length      |C|   Reserved   |
+-----+-----+-----+-----+-----+-----+-----+-----+
|                                     Flood Reflection Cluster ID
+-----+-----+-----+-----+-----+-----+-----+-----+

```

Type: TBD

Length: The length, in octets, of the following fields.

C (Client): This bit is set to indicate that the router acts as a flood reflector client. When this bit is NOT set, the router acts as a flood reflector.

RESERVED: This field is reserved for future use. It MUST be set to 0 when sent and MUST be ignored when received.

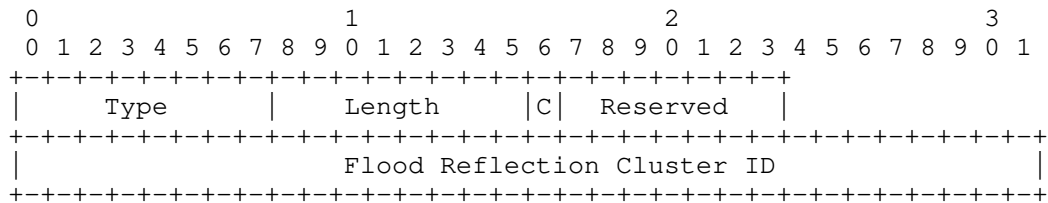
Flood Reflection Cluster ID: The Flood Reflection Cluster Identifier is the same as that defined in the Flood Reflection TLV.

The Flood Reflection Discovery sub-TLV MUST NOT appear more than once in TLV 242. A router receiving multiple Flood Reflection Discovery sub-TLVs in TLV 242 SHOULD use the values in the first sub-TLV.

5. Flood Reflection Adjacency Sub-TLV

The Flood Reflection Adjacency sub-TLV is advertised as a sub-TLV of TLVs 22, 23, 25, 141, 222, and 223. Its presence indicates that a given adjacency is a flood reflector adjacency. It is included in L2

area scope flooded LSPs. Flood Reflection Adjacency sub-TLV has the following format:



Type: TBD

Length: The length, in octets, of the following fields.

C (Client): This bit is set to indicate that the router advertising this adjacency is a flood reflector client. When this bit is NOT set, the router advertising this adjacency is a flood reflector.

RESERVED: This field is reserved for future use. It MUST be set to 0 when sent and MUST be ignored when received.

Flood Reflection Cluster ID: The Flood Reflection Cluster Identifier is the same as that defined in the Flood Reflection TLV.

The Flood Reflection Adjacency sub-TLV MUST NOT appear more than once in a given TLV. A router receiving multiple Flood Reflection Adjacency sub-TLVs in a TLV SHOULD use the values in the first sub-TLV.

6. Flood Reflection Discovery

A router participating in flood reflection MUST be configured as an L1/L2 router. It originates the Flood Reflection Discovery sub-TLV with area flooding scope in L1 only. Normally, all routers on the edge of the L1 area (those having traditional L2 adjacencies) will advertise themselves as route reflector clients. Therefore, a flood reflector client will have both traditional L2 adjacencies and flood reflector L2 adjacencies.

A router acting as a flood reflector MUST NOT have any traditional L2 adjacencies. It will be an L1/L2 router only by virtue of having flood reflector L2 adjacencies. A router desiring to act as a flood reflector will advertise itself as such using the Flood Reflection Discovery sub-TLV in L1.

A given flood reflector or flood reflector client can only participate in a single cluster, as determined by the value of its Flood Reflection Cluster ID.

Upon reception of Flood Reflection Discovery sub-TLVs, a router acting as flood reflector client MUST initiate a tunnel towards each flood reflector with which it shares an Flood Reflection Cluster ID. The L2 adjacencies formed over such tunnels MUST be marked as flood reflector adjacencies. If the client has a direct L2 adjacency with the flood reflector it SHOULD use it instead of instantiating a new tunnel.

Upon reception of Flood Reflection Discover TLVs, a router acting as a flood reflector client MAY initiate tunnels with L1-only adjacencies towards all the other flood reflector clients in its cluster. These tunnels MAY be used for forwarding to improve the load-balancing characteristics of the L1 area.

7. Flood Reflection Adjacency Formation

In order to simplify both implementations and network deployments, we do not allow the formation of complex hierarchies of flood reflectors and clients. All flood reflectors and flood reflector clients in the same L1 area MUST share the same Flood Reflector Cluster ID. A flood reflector MUST only form flood reflection adjacencies with flood reflector clients. A flood reflector MUST NOT form any traditional L2 adjacencies. Flood reflector clients MUST only form flood reflection adjacencies with flood reflectors. Flood reflector clients may form traditional L2 adjacencies with flood reflector clients or nodes not participating in flood reflection.

The Flood Reflector Cluster ID and flood reflector roles advertised in the Flood Reflection TLVs in IIHs are used to ensure that flood reflection adjacencies that are established meet the above criteria.

Once a flood reflection adjacency is established, the flood reflector and the flood reflector client MUST advertise the adjacency by including the Flood Reflection Adjacency Sub-TLV in the Extended IS reachability TLV or MT-ISN TLV.

8. Redistribution of Prefixes

In some scenarios, L2 prefixes need to be redistributed into L1 by the route reflector clients. However, if a L1 area edge router doesn't have any L2 flood reflector adjacencies, then it cannot be the shortest path egress in the L2 topology. Therefore, flood reflector client SHOULD only redistribute L2 prefixes into L1 if it has an L2 flood reflector adjacency. The L2 prefixes advertisements

redistributed into L1 SHOULD be normally limited to L2 intra-area routes (as defined in [RFC7775]), if the information exists to distinguish them from other L2 prefix advertisements.

On the other hand, in topologies that make use of flood reflection to hide the structure of L1 areas while still providing transit forwarding across them, we generally do not need to redistribute L1 prefixes advertisements into L2.

In deployment scenarios where L1 tunnels are not used, all L1/L2 edge nodes MUST be flood reflector clients.

9. Route Computation

To ensure loop-free routing, the route reflection client MUST follow the normal L2 computation to determine L2 routes. This is because nodes outside the L1 area will generally not be aware that flood reflection is being performed. The flood reflection clients need to produce the same result for the L2 route computation as a router not participating in flood reflection. However, a flood reflector client will not necessarily use a given L2 route for forwarding. For an L2 route that uses a flood reflection adjacency as a next-hop, the flood reflection client may use the next-hop from an L1 route instead.

On the reflection client, after L2 and L1 computation, all flood reflector adjacencies used as next-hops for L2 routes MUST be examined and replaced with the correct L1 tunnel next-hop to the egress. Alternatively, if the ingress has adequate reachability information to ensure forwarding towards destination via L1 routes, L2 routes using flood reflector adjacencies as next-hops can be omitted entirely. Due to the rules in Section 7 the computation in the resulting topology is relatively simple, the L2 SPF from a flood reflector client is guaranteed to reach within a hop the Flood Reflector and in the following hop the L2 egress to which it has a L1 forwarding tunnel. However, if the topology has L2 paths which are not route reflected and look "shorter" than the path through the Flood Reflector then the computation will have to track the egress out of the L1 domain by a more advanced algorithm.

10. Special Considerations

In pathological cases setting the overload bit in L1 (but not in L2) can partition L1 forwarding, while allowing L2 reachability through flood reflector adjacencies to exist. In such a case a node cannot replace a route through a flood reflector adjacency with a L1 shortcut and the client can use the L2 tunnel to the flood reflector for forwarding while it MUST initiate an alarm and declare misconfiguration.

A flood reflector with directly L2 attached prefixes should advertise those in L1 as well since based on preference of L1 routes the clients will not try to use the L2 flood reflector adjacency to route the packet towards them. A very, very corner case is when the flood reflector is reachable via L2 flood reflector adjacency (due to underlying L1 partition) only in which case the client can use the L2 tunnel to the flood reflector for forwarding towards those prefixes while it MUST initiate an alarm and declare misconfiguration.

Instead of modifying the computation procedures one could imagine a flood reflector solution where the Flood Reflector would re-advertise the L2 prefixes with a 'third-party' next-hop but that would have less desirable convergence properties than the solution proposed and force a fork-lift of all L2 routers to make sure they disregard such prefixes unless in the same L1 domain as the Flood Reflector.

Depending on pseudo-node choice in case of a broadcast domain with multiple flood reflectors attached this can lead to a partitioned LAN and hence a router discovering such a condition MUST initiate an alarm and declare misconfiguration.

11. IANA Considerations

This document requests allocation for the following IS-IS TLVs and Sub-TLVs.

11.1. New IS-IS TLV Codepoint

This document requests the following IS-IS TLV:

Value	Name	IIH	LSP	SNP	Purge
TBD1	Flood Reflection	y	n	n	n

11.2. Sub TLVs for TLV 242

This document request the following registration in the "sub-TLVs for TLV 242" registry.

Type	Description
TBD2	Flood Reflection Discovery

11.3. Sub TLVs for TLV 22, 23, 25, 141, 222, and 223

This document requests the following registration in the "sub-TLVs for TLV 22, 23, 25, 141, 222, and 223" registry.

Type	Description	22	23	25	141	222	223
----	-----	---	---	---	---	---	---
TBD3	Flood Reflector Adjacency	y	y	y(s)	y	y	y

12. Security Considerations

This document introduces no new security concerns to ISIS or other specifications referenced in this document.

13. Acknowledgements

The authors thank Shraddha Hegde, Peter Psenak, and Les Ginsberg for their thorough review and detailed discussions.

14. References

14.1. Informative References

- [ID.draft-ietf-idr-bgp-optimal-route-reflection-19]
Raszuk et al., R., "BGP Optimal Route Reflection", July 2019.
- [RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, DOI 10.17487/RFC4271, January 2006, <<https://www.rfc-editor.org/info/rfc4271>>.
- [RFC4456] Bates, T., Chen, E., and R. Chandra, "BGP Route Reflection: An Alternative to Full Mesh Internal BGP (IBGP)", RFC 4456, DOI 10.17487/RFC4456, April 2006, <<https://www.rfc-editor.org/info/rfc4456>>.
- [RFC8099] Chen, H., Li, R., Retana, A., Yang, Y., and Z. Liu, "OSPF Topology-Transparent Zone", RFC 8099, DOI 10.17487/RFC8099, February 2017, <<https://www.rfc-editor.org/info/rfc8099>>.

14.2. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<https://www.rfc-editor.org/info/rfc2119>>.
- [RFC7775] Ginsberg, L., Litkowski, S., and S. Previdi, "IS-IS Route Preference for Extended IP and IPv6 Reachability", RFC 7775, DOI 10.17487/RFC7775, February 2016, <<https://www.rfc-editor.org/info/rfc7775>>.
- [RFC7981] Ginsberg, L., Previdi, S., and M. Chen, "IS-IS Extensions for Advertising Router Information", RFC 7981, DOI 10.17487/RFC7981, October 2016, <<https://www.rfc-editor.org/info/rfc7981>>.

Authors' Addresses

Tony Przygienda
Juniper
1137 Innovation Way

Sunnyvale, CA

USA

Email: prz@juniper.net

Chris Bowers
Juniper
1137 Innovation Way

Sunnyvale, CA

USA

Email: cbowers@juniper.net

Yiu Lee
Comcast
1800 Bishops Gate Blvd
Mount Laurel, NJ 08054
US

Email: Yiu_Lee@comcast.com

Alankar Sharma
Comcast
1800 Bishops Gate Blvd
Mount Laurel, NJ 08054
US

Email: Alankar_Sharma@comcast.com

Russ White
Juniper
1137 Innovation Way

Sunnyvale, CA

USA

Email: russw@juniper.net