Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: May 5, 2020

A. Przygienda
C. Bowers
Juniper
Y. Lee
A. Sharma
Comcast
R. White
Juniper

November 2, 2019

Flood Reflectors
draft-przygienda-lsr-flood-reflection-00

Abstract

This document provides specification of an optional ISIS extension that allows to create l2 flood reflector topologies allowing forwarding through all paths within L1 areas when they are used as 'transit' to guarantee L2 connectivity between L2 'islands'. Only routers participating in the flood reflection have to be upgraded and with that the feature allows to significantly increase practical size of ISIS L2 backbone without forklifting the whole domain or complex configuration requirements.

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 May 5, 2020.

Table of Contents

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 certain deployment scenarios of L2 backbones, this limit presents an obstacle.

While the standard solution to increase the scale of an IS-IS deployement is to break it up into multiple L1 flooding domains and a single L2 backbone, and alternative way is to think about "multiple" L2 flooding domains connected via L1 flooding domains. In such a solution, the L2 flooding domains are connected by "L1/L2 lanes" through the L1 areas to form a single L2 backbone again. However, in the 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 and with that ultimately does not help to reduce number of L2 routers and increase the scalability of L2 backbone.

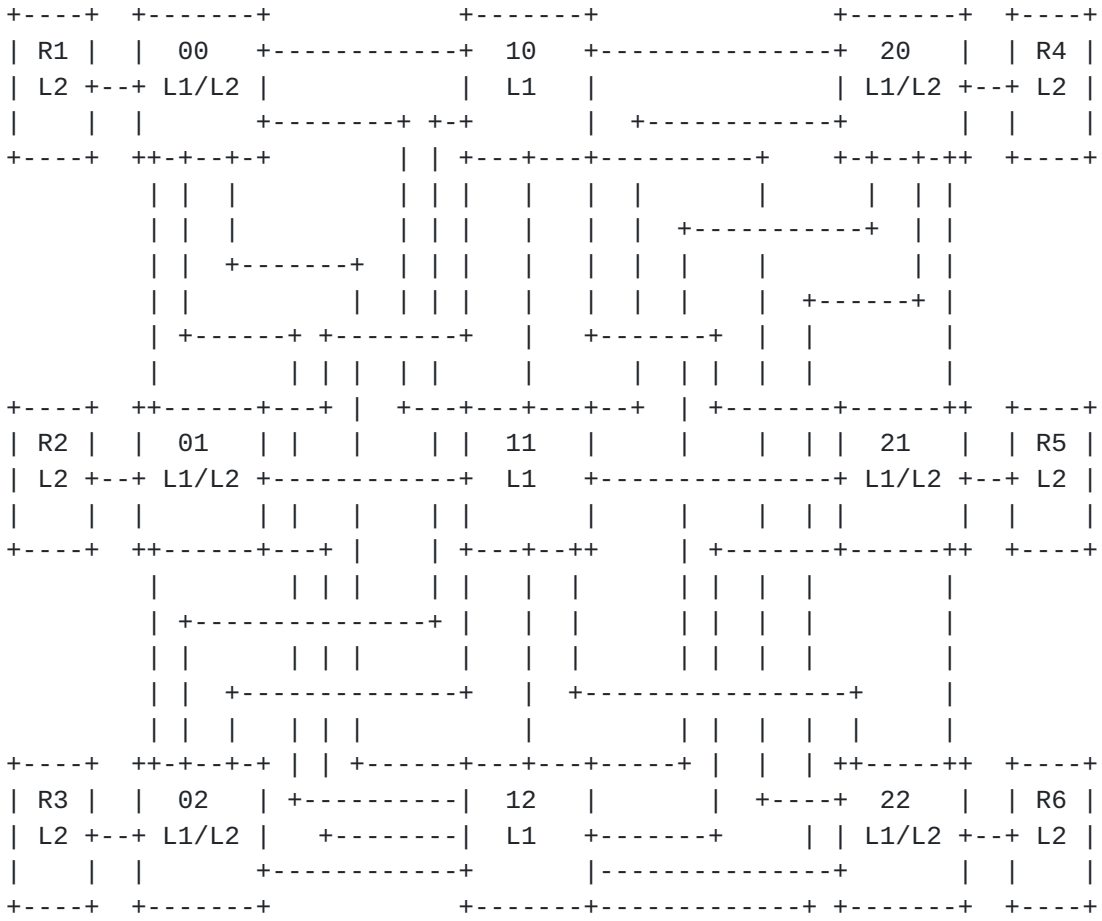


Figure 1

Figure 1 is an example of a network where a topologically rich L1 area is used to provide transit between six different routers in L2 "partitions" (R1-R6). To take advantage of the cornucopia of paths in the L1 transit, 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.

The mechanism described in [RFC8099] could be used in ISIS to build a full mesh of tunnels over the L1 transit, but a full mesh of tunnels can also quickly limit the scaling. The network in Figure 2 would expose 6 L1/L2 nodes and (5 * 6)/2 = 15 L2 tunnels. In a slightly larger network, however, in a comparable topology containing 15 L1/L2 edge nodes the number grows very quickly to 105 tunnels.

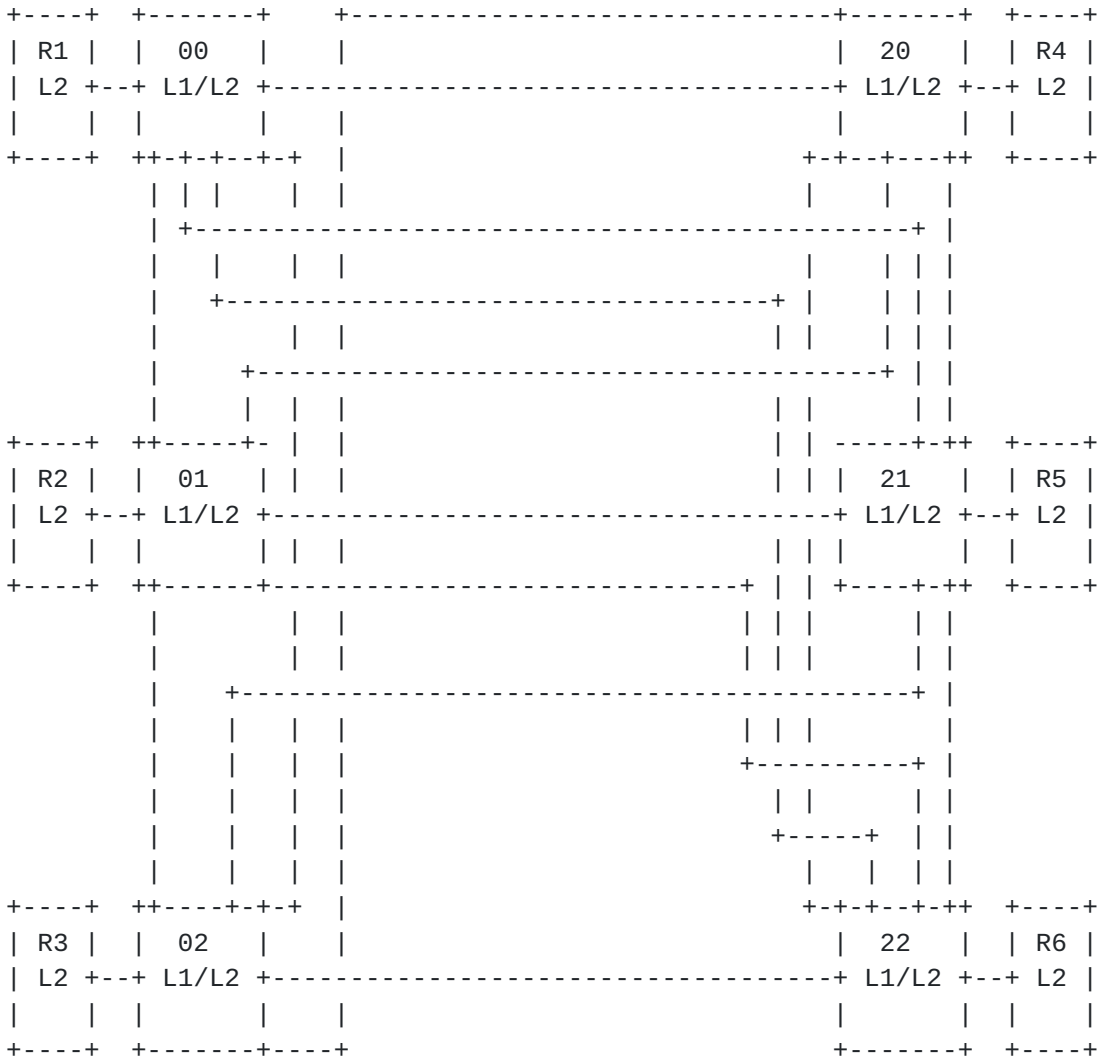


Figure 2

BGP, as specified in [RFC4271], faced a similar scaling problem, which has been solved in many networks by deploying BGP route reflectors [RFC4456]. And, as another crucial observation, BGP route reflectors do not necessarily have to be in the forwarding path of the traffic. Such incongruity of forwarding and control path is allowing conceptually to scale the control plane independently of the number of nodes participating in the forwarding path.

We propose here a similar solution for IS-IS. A good approximation of what a "flood reflector" control plane approach would look like is shown in Figure 3, where router 11 is used as 'reflector.' All L1/L2 routers build an L2 tunnel to such reflectors, so we end up with only 6 L2 tunnels instead of 15 needed for a full mesh. Multiple such reflectors can be used, of course, allowing the network operator to balance between resilience, path utilization, and state in the control plane. The resulting L2 tunnel scale is roughly R * n where R is the redundancy factor or in other words, number of flood reflectors used. This compares quite favorably with n^2 / 2 tunnels used in a fully meshed L2 solution.

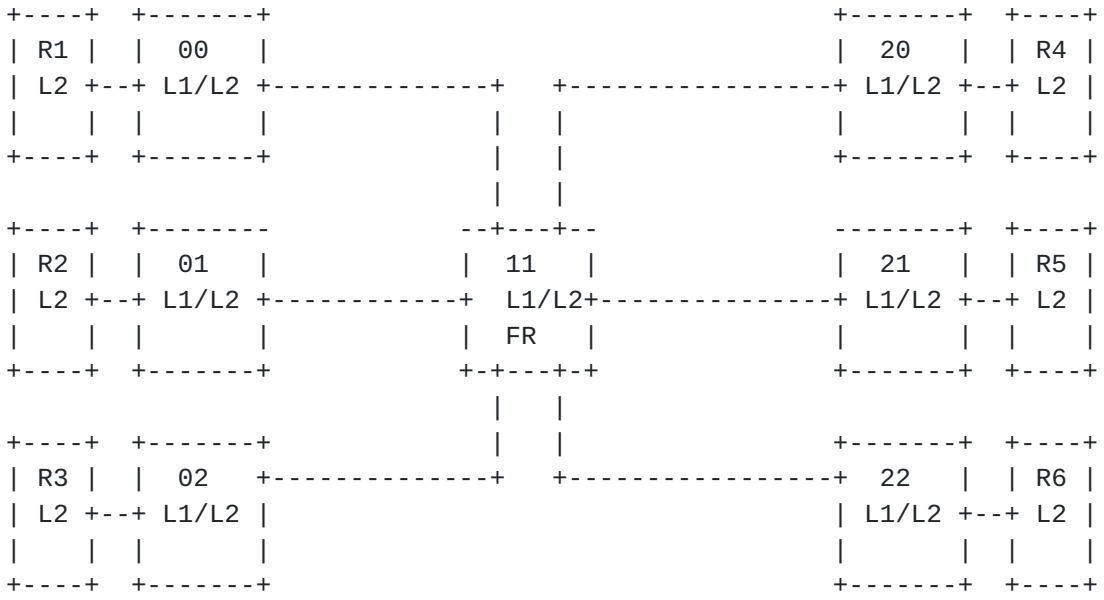


Figure 3

And thus, as suggested already by Figure 3, to scale L2 better we decouple the forwarding plane from the control plane in a first step. Router 11 can reflood L2 information while the other routers in Level 1 are not visible in Level 2. Without further additions, however, with such a change all the data traffic will traverse through 11

creating a bottleneck and disregarding available capacity in the paths crossing the 'hidden' routers 10 and 12.

To use the whole L1 capacity with flood reflectors, multiple pieces will be necessary, only one of which is a local protocol extension on the L1/L2 leafs and the 'flood reflectors'. In first approximation these extensions include:

- 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.

- 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.

- 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.

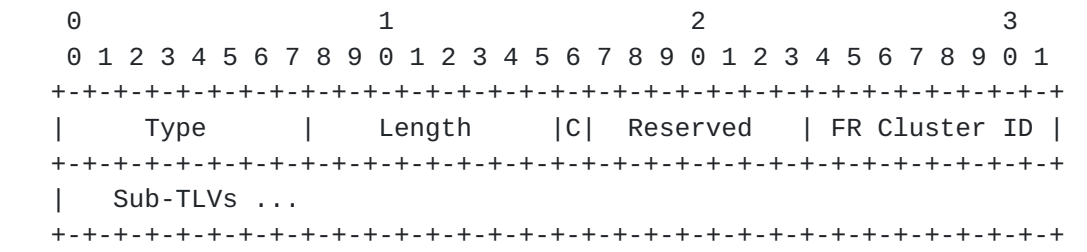
One possible solution to this problem is to expose additional topology information into the L2 flooding domains. In the example network given, links from router 01 to router 02 can be exposed into L2 even when 01 and 02 are participating in flood reflection. This information would allow the L2 nodes 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 in the solution is trivial to achieve by building multiple flood reflectors into the L1 area while all reflectors are still remaining completely stateless and do not need any kind of synchronized algorithms amongst themselves except standard ISIS flooding procedures and database.

3. Flood Reflection TLV

The Flood Reflection TLV is a new top-level TLV that SHOULD appear in IIHs. The Flood Reflection TLV indicates the flood reflector cluster (based on Flood Reflector Cluster ID) that a given router interface 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 Reflector Cluster ID and flood reflector roles advertised in the IIHs on a given interface are used to ensure that flood reflector adjacencies are only formed between a flood reflector and flood reflector client, and that the Flood Reflector Cluster IDs match. The Flood Reflection TLV has the following format:



Type: TBD

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

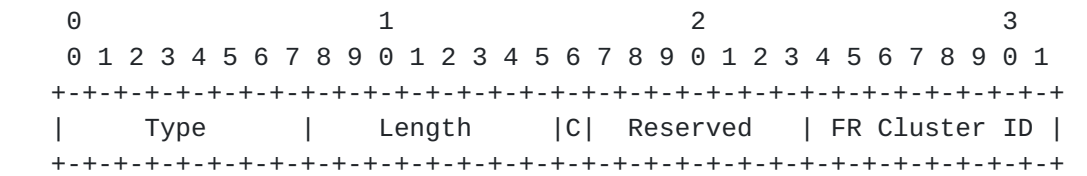
C (Client): This bit is set to indicate that the router acts as a flood reflector client. When this bit is NOT set, the router acts as a flood reflector.

Flood Reflector Cluster ID: Flood Reflector Cluster Identifier. This value allows a flood reflector client to establish flood reflector adjacencies with multiple flood reflectors. Each flood reflector is the "hub" of a flood reflector cluster. Each flood reflector cluster is distinguished by a Flood Reflector Cluster Identifier unique within the IGP domain.

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.

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:



Type: TBD

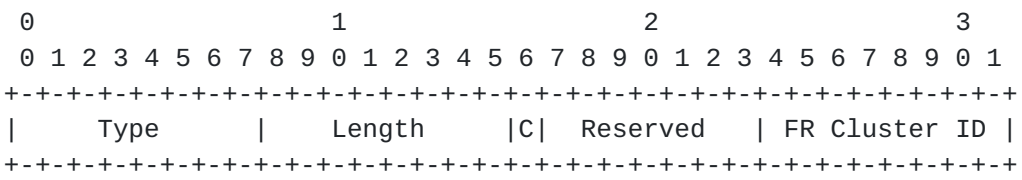
Length: The length, in octets, of the following fields.

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.

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

5. Flood Reflector Adjacency Sub-TLV

The Flood Reflector 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 Reflector 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.

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

6. Procedures

There are a number of points to consider when implementing and deploying flood reflection, including:

A router participating in flood reflection MUST be configured as L1L2 router. It originates the Flood Reflection Discovery sub-TLV with area flooding scope in L1 only. Normally routers on the edge of the area, i.e. with non-Flood Reflector L2 adjacencies, will advertise themselves as clients. Any L1L2 non-client router in the area can act as flood reflector.

A flood reflector can participate in a single cluster only, the clients are free to participate in multiple clusters at the same time.

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 Reflector 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 (in case it doesn't have such direct L1 adjacencies already) SHOULD initialize tunnels towards all the other clients in its clusters. L1 *only* adjacencies SHOULD be built over such tunnels to ensure their liveliness, but other means can be used (since those adjacencies are used for L1 forwarding, it is prudent to advertise them into L1 as forwarding links).

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. Alternately, 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.

A node, when advertising the L2 flood reflector adjacency MUST include the Flood Reflector Adjacency Sub-TLV in Extended IS reachability TLV and MT-ISN TLV.

7. Adjacency Forming Procedures

To ensure loop-free routing the ingress routers MUST follow normal L2 computation to generate L2 routes. This is because nodes outside the L1 area may not be aware that flooding reflection is performed. The resulting short cuts through the L1 area needs to be able to easily

calculate the egress L1/L2 router where the tunnel tail-end is located.

To prevent complex scenarios of flood reflectors building L2 adjacencies within a cluster or across clusters or hierarchies of reflectors, a flood reflector MUST never form an L2 adjacency with a peer if the peer is not a client in the same Cluster ID. This ensures a L2 computation on an ingress link or adjacency following a flood reflector adjacency will always traverse a client of the flood reflector to exit the flooding domain. This allows shortcuts through the L1 area to be used without any danger of forwarding loops.

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

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.

8. 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.

9. IANA Considerations

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

9.1. 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

9.2. 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

9.3. 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

10. Security Considerations

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

11. Acknowledgements

Thanks to Shraddha Hegde and others for thorough review.

12. References

12.1. 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>.

12.2. 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>.

[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