

Internet Engineering Task Force
Internet-Draft
Intended status: Standards Track
Expires: April 5, 2020

T. Li
S. Chen
Arista Networks
October 3, 2019

Area Proxy for IS-IS
draft-li-lsr-isis-area-proxy-00

Abstract

Link state routing protocols have hierarchical abstraction already built into them. However, when lower levels are used for transit, they must expose their internal topologies to each other, leading to scale issues.

To avoid this, this document discusses extensions to the IS-IS routing protocol that would allow level 1 areas to provide transit, yet only inject an abstraction of the level 1 topology into level 2. Each level 1 area is represented as a single level 2 node, thereby enabling greater scale.

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 April 5, 2020.

Copyright Notice

Copyright (c) 2019 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. Introduction	2
1.1. Requirements Language	3
2. Area Proxy	4
3. Inside Router Functions	5
3.1. The Area Proxy Router Capability	5
3.2. Level 2 SPF Computation	5
4. Area Leader Functions	6
4.1. Area Leader Election	6
4.2. Redundancy	6
4.3. Area Proxy System Identifier TLV	6
4.4. Area Proxy LSP Generation	7
5. Inside Edge Router Functions	8
5.1. Generating L2 IIHs to Outside Routers	8
5.2. Filtering LSP information	8
6. Acknowledgments	9
7. IANA Considerations	9
8. Security Considerations	9
9. References	9
9.1. Normative References	9
9.2. Informative References	10
9.3. URIs	10
Authors' Addresses	10

1. Introduction

The IS-IS routing protocol IS-IS [ISO10589] currently supports a two-level hierarchy of abstraction. The fundamental unit of abstraction is the 'area', which is a (hopefully) connected set of systems running IS-IS at the same level. Level 1, the lowest level, is abstracted by routers that participate in both Level 1 and Level 2, and they inject area information into Level 2. Level 2 systems seeking to access Level 1, use this abstraction to compute the shortest path to the Level 1 area. The full topology database of Level 1 is not injected into Level 2, only a summary of the address space contained within the area, so the scalability of the Level 2 Link State Database (LSDB) is protected.

This works well if the Level 1 area is tangential to the Level 2 area. This also works well if there are several routers in both Level 1 and Level 2 and they are adjacent, so Level 2 traffic will

never need to transit Level 1 only routers. Level 1 will not contain any Level 2 topology, and Level 2 will only contain area abstractions for Level 1.

Unfortunately, this scheme does not work so well if the Level 1 only area needs to provide transit for Level 2 traffic. For Level 2 shortest path first (SPF) computations to work correctly, the transit topology must also appear in the Level 2 LSDB. This implies that all routers that could provide transit, plus any links that might also provide Level 2 transit must also become part of the Level 2 topology. If this is a relatively tiny portion of the Level 1 area, this is not overly painful.

However, with today's data center topologies, this is problematic. A common application is to use a Layer 3 Leaf-Spine (L3LS) topology, which is a folded 3-stage Clos [Clos] fabric. It can also be thought of as a complete bipartite graph. In such a topology, the desire is to use Level 1 to contain the routing dynamics of the entire L3LS topology and then to use Level 2 for the remainder of the network. Leaves in the L3LS topology are appropriate for connection outside of the data center itself, so they would provide connectivity for Level 2. If there are multiple connections to Level 2 for redundancy, or other areas, these too would also be made to the leaves in the topology. This creates a difficulty because there are now multiple Level 2 leaves in the topology, with connectivity between the leaves provided by the spines.

Following the current rules of IS-IS, all spine routers would necessarily be part of the Level 2 topology, plus all links between a Level 2 leaf and the spines. In the limit, where all leaves need to support Level 2, it implies that the entire L3LS topology becomes part of Level 2. This is seriously problematic as it more than doubles the LSDB held in the L3LS topology and eliminates any benefits of the hierarchy.

This document discusses the handling of IP traffic. Supporting MPLS based traffic is a subject for future work.

1.1. 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 BCP 14 [1] [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

2. Area Proxy

To address this, we propose to completely abstract away the details of the Level 1 area topology within Level 2, making the entire area look like a single proxy system directly connected to all of the area's Level 2 neighbors. By only providing an abstraction of the topology, Level 2's requirement for connectivity can be satisfied without the full overhead of the area's internal topology. It then becomes the responsibility of the Level 1 area to ensure the forwarding connectivity that's advertised.

For this discussion, we'll consider a single Level 1 IS-IS area to be the Inside Area, and the remainder of the Level 2 area is the Outside Area. All routers within the Inside Area speak Level 1 and Level 2 IS-IS on all of the links within the topology. We propose to implement Area Proxy by having a Level 2 Proxy Link State Protocol Data Unit (PDU, LSP) that represents the entire Inside Area. This is the only LSP from the area that will be flooded into the overall Level 2 LSDB.

There are four classes of routers that we need to be concerned with in this discussion:

Inside Router A router within the Inside Area that runs Level 1 and Level 2 IS-IS.

Area Leader The Area Leader is an Inside Router that is elected to represent the Level 1 area by injecting the Proxy LSP into the Level 2 LSDB. There may be multiple candidates for Area Leader, but only one is elected at a given time.

Inside Edge Router An Inside Edge Router is an Inside Area Router that has at least one Level 2 interface outside of the Inside Area.

Outside Edge Router An Outside Edge Router is a Level 2 router that is outside of the Inside Area that has an adjacency with an Inside Edge Router.

All Inside Edge Routers learn the Area Proxy System Identifier from the Level 1 LSDB and use that as the system identifier in their Level 2 IS-IS Hello PDUs (IIHs) on all Outside interfaces. Outside Edge Routers should then advertise an adjacency to the Area Proxy System Identifier. This allows all Outside Routers to use the Proxy LSP in their SPF computations without seeing the full topology of the Inside Area.

Area Proxy functionality assumes that all circuits are either Level 1-2 circuits within the Inside Area, or Level 2 circuits between Outside Routers and a single Inside Edge Router. Multi-access circuits (i.e. Ethernet in LAN mode) with multiple Inside Edge Routers and an Outside Router are not supported.

3. Inside Router Functions

All Inside Routers run Level 1-2 IS-IS and must be explicitly instructed to enable the Area Proxy functionality. To signal their readiness to participate in Area Proxy functionality, they will advertise the Area Proxy Router Capability as part of its Level 1 Router Capability TLV.

3.1. The Area Proxy Router Capability

The Area Proxy Router Capability is a sub-TLV of the Router Capability TLV [RFC7981] and has the following format:

```

      0                               1
      0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5
      +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
      | TLV Type           | TLV Length   |
      +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+

```

TLV Type: YYY

TLV Length: 0

A router advertising this TLV indicates that it is running Level 1-2 and is prepared to perform Area Proxy functions.

3.2. Level 2 SPF Computation

When Outside Routers perform a Level 2 SPF computation, they will use the Area Proxy LSP for computing a path transiting the Inside Area. Because the topology has been abstracted away, the cost for transiting the Inside Area will be zero.

When Inside Routers perform a Level 2 SPF computation, they must ignore the Area Proxy LSP. Further, because these systems do see the Inside Area topology, the link metrics internal to the area are visible. This could lead to different and possibly inconsistent SPF results, potentially leading to forwarding loops.

To prevent this, the Inside Routers must consider the metrics of links outside of the Inside Area (inter-area metrics) separately from the metrics of the Inside Area links (intra-area metrics). Intra-

area metrics are always less than any inter-area metric. Thus, if two paths have different total inter-area metrics, the path with the lower inter-area metric would be preferred, regardless of any intra-area metrics involved. However, if two paths have equal inter-area metrics, then the intra-area metrics would be used to compare the paths.

4. Area Leader Functions

The Area Leader has several responsibilities. First, it must inject the Area Proxy System Identifier into the Level 1 LSDB. Second, the Area Leader must generate the Proxy LSP for the Inside Area.

4.1. Area Leader Election

The Area Leader is selected using the election mechanisms and TLVs described in Dynamic Flooding for IS-IS [I-D.ietf-lsr-dynamic-flooding].

4.2. Redundancy

If the Area Leader fails, another candidate may become Area Leader and MUST regenerate the Area Proxy LSP. The failure of the Area Leader is not visible outside of the area and appears to simply be an update of the Area Proxy LSP.

4.3. Area Proxy System Identifier TLV

The Area Proxy System Identifier TLV allows the Area Leader to advertise the existence of an Area Proxy System Identifier. This TLV is injected into the Area Leader's Level 1 LSP.

The format of the Area Proxy System Identifier TLV is:

0										1										2									
0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1	2	3						
TLV Type										TLV Length										Proxy SysID									
Proxy System Identifier continued ...																													

TLV Type: XXX

TLV Length: length of a system ID (6)

Proxy System Identifier: the Area Proxy System Identifier.

The Area Leader MAY advertise the Area Proxy System Identifier TLV when it observes that all Inside Routers are advertising the Area Proxy Router Capability. Their advertisements indicate that they are individually ready to perform Area Proxy functionality. The Area Leader then advertises the Area Proxy System Identifier TLV to indicate that the Inside Area should enable Area Proxy functionality.

Other candidates for Area Leader MAY also advertise the Area Proxy System Identifier when they observe that all Inside Routers are advertising the Area Proxy Router Capability. All candidates advertising the Area Proxy System Identifier TLV MUST be advertising the same system identifier. Multiple proxy system identifiers in a single area is a misconfiguration.

The Area Leader and other candidates for Area Leader MAY withdraw the Area Proxy System Identifier when one or more Inside Routers are not advertising the Area Proxy Router Capability. This will disable Area Proxy functionality. However, before withdrawing the Area Proxy System Identifier, an implementation should protect against unnecessary churn from transients by delaying the withdrawal. The amount of delay is implementation-dependent.

4.4. Area Proxy LSP Generation

Each Inside Router generates a Level 2 LSP, and the Level 2 LSPs for the Inside Edge Routers will include adjacencies to Outside Edge Routers. Unlike normal Level 2 operations, these LSPs are not advertised outside of the Inside Area and must be filtered by all Inside Edge Routers to not be flooded to Outside Routers.

The Area Leader uses the Level 2 LSPs generated by the Inside Edge Routers to generate the Area Proxy LSP. This LSP is originated using the Area Proxy System Identifier and includes adjacencies for all of the Outside Edge Routers that have been advertised by the Inside Edge Routers. Since the Outside Edge Routers also advertise an adjacency to the proxy system identifier, this will result in a bi-directional adjacency. The Area Proxy LSP is the only LSP that is injected into the overall Level 2 LSDB, with all other Level 2 LSPs from the Inside Area being filtered out at the Inside Area boundary.

The Area Leader may also insert additional TLVs into the Area Proxy LSP for additional information for the Outside Area. It is RECOMMENDED that the Area Leader insert the Dynamic Hostname TLV [RFC5301] into the Area Proxy LSP. The Area Leader SHOULD insert additional TLVs describing any routing prefixes that should be advertised on behalf of the area. These prefixes may be learned from the Level 1 LSDB, statically configured, or redistributed from

another routing protocol, using the usual TLVs for prefix advertisement. [RFC5305] [RFC5308] [RFC5120]

5. Inside Edge Router Functions

The Inside Edge Router has two additional and important functions. First, it must generate IIHs that appear to have come from the Area Proxy System Identifier. Second, it must filter the L2 LSPs, Partial Sequence Number PDUs (PSNPs), and Complete Sequence Number PDUs (CSNPs) that are being advertised to Outside Routers.

5.1. Generating L2 IIHs to Outside Routers

The Inside Edge Router has one or more Level 2 interfaces to Outside Routers. These may be identified by explicit configuration or by the fact that they are not also Level 1 circuits. On these Level 2 interfaces, the Inside Edge Router MUST NOT send an IIH until it has learned the Area Proxy System Id from the Area Leader. Then, once it has learned the Area Proxy System Id, it should generate its IIHs on the circuit using the Proxy System Id as the source of the IIH.

Using the Proxy System Id causes the Outside Router to advertise an adjacency to the Proxy System Id, not to the Inside Edge Router, which supports the proxy function. The normal system id of the Inside Edge Router MUST NOT be used as it will cause unnecessary adjacencies to form and subsequently flap.

5.2. Filtering LSP information

For the proxy abstraction to be effective the L2 LSPs generated by the Inside Routers MUST be restricted to the Inside Area. The Inside Routers know which system ids are members of the Inside Area based on the Level 1 LSDB. To prevent unwanted LSP information from escaping the Inside Area, the Inside Edge Router MUST perform filtering of LSP flooding, CSNPs, and PSNPs. Specifically:

A Level 2 LSP with a source system identifier that is found in the Level 1 LSDB should never be flooded to an Outside Router.

A Level 2 CSNP sent to an Outside Router MUST NOT contain any information about an LSP with a system identifier found in the Level 1 LSDB. If an Inside Edge Router filters a CSNP and there is no remaining content, then the CSNP MUST NOT be sent. The source address of the CSNP should be the Area Proxy System Id.

A Level 2 PSNP sent to an Outside Router MUST NOT contain any information about an LSP with a system identifier found in the Level 1 LSDB. If an Inside Edge Router filters a PSNP and there

is no remaining content, then the PSNP MUST NOT be sent. The source address of the PSNP should be the Area Proxy System Id.

6. Acknowledgments

The authors would like to thank Bruno Decraene and Gunter Van De Velde for their many helpful comments. The authors would also like to thank a small group that wishes to remain anonymous for their valuable contributions.

7. IANA Considerations

This memo requests that IANA allocate and assign one code point from the IS-IS TLV Codepoints registry for the Area Pseudonode TLV (XXX).

IANA is also requested to allocate and assign one code point from the IS-IS Router Capability TLV sub-TLV registry for the Area Proxy Capability (YYY).

8. Security Considerations

This document introduces no new security issues. Security of routing within a domain is already addressed as part of the routing protocols themselves. This document proposes no changes to those security architectures.

9. References

9.1. Normative References

- [I-D.ietf-lsr-dynamic-flooding]
Li, T., Psenak, P., Ginsberg, L., Chen, H., Przygienda, T., Cooper, D., Jalil, L., and S. Dontula, "Dynamic Flooding on Dense Graphs", draft-ietf-lsr-dynamic-flooding-03 (work in progress), June 2019.
- [ISO10589]
International Organization for Standardization, "Intermediate System to Intermediate System Intra-Domain Routing Exchange Protocol for use in Conjunction with the Protocol for Providing the Connectionless-mode Network Service (ISO 8473)", ISO/IEC 10589:2002, Nov. 2002.
- [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>>.

- [RFC5120] Przygienda, T., Shen, N., and N. Sheth, "M-ISIS: Multi Topology (MT) Routing in Intermediate System to Intermediate Systems (IS-ISs)", RFC 5120, DOI 10.17487/RFC5120, February 2008, <<https://www.rfc-editor.org/info/rfc5120>>.
- [RFC5301] McPherson, D. and N. Shen, "Dynamic Hostname Exchange Mechanism for IS-IS", RFC 5301, DOI 10.17487/RFC5301, October 2008, <<https://www.rfc-editor.org/info/rfc5301>>.
- [RFC5305] Li, T. and H. Smit, "IS-IS Extensions for Traffic Engineering", RFC 5305, DOI 10.17487/RFC5305, October 2008, <<https://www.rfc-editor.org/info/rfc5305>>.
- [RFC5308] Hopps, C., "Routing IPv6 with IS-IS", RFC 5308, DOI 10.17487/RFC5308, October 2008, <<https://www.rfc-editor.org/info/rfc5308>>.
- [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>>.
- [RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <<https://www.rfc-editor.org/info/rfc8174>>.

9.2. Informative References

- [Clos] Clos, C., "A Study of Non-Blocking Switching Networks", The Bell System Technical Journal Vol. 32(2), DOI 10.1002/j.1538-7305.1953.tb01433.x, March 1953, <<http://dx.doi.org/10.1002/j.1538-7305.1953.tb01433.x>>.

9.3. URIs

- [1] <https://tools.ietf.org/html/bcp14>

Authors' Addresses

Tony Li
Arista Networks
5453 Great America Parkway
Santa Clara, California 95054
USA

Email: tony.li@tony.li

Sarah Chen
Arista Networks
5453 Great America Parkway
Santa Clara, California 95054
USA

Email: sarahchen@arista.com