Internet Engineering Task Force
Internet-Draft
Intended status: Informational
Expires: December 30, 2018

T. Li
Arista Networks

June 28, 2018

Level 1 Area Abstraction for IS-IS
draft-li-area-abstraction-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, 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 topology into level 2.

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 December 30, 2018.

Copyright Notice

Table of Contents

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 is protected.

This works well if the Level 1 area is tangential to the Level 2 area. This also works well if there are a number of 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 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 link state database. This implies that all routers that could possibly 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 onerous.

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 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 to 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 provide by the spines.

Following the 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 link state database held in the L3LS topology and eliminates any benefits of the hierarchy.

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 RFC 2119 [RFC2119].

2. Area Abstraction

We propose to completely abstract away the Level 2 topology of the Level 1 area, making the entire area look like a single 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.

We propose to implement Area Abstraction by having a Level 2 pseudonode that represents the entire Level 1 area. This is the only LSP from the area that will be injected into the overall Level 2 link state database.

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

Area Leader The Area Leader is a router in the Level 1 area that is elected to represent the Level 1 area by injecting an LSP into the Level 2 link state database.

Area Edge Router An Area Edge Router is a router that is part of the Level 1 area and has at least one Level 2 interface outside of the Area.

Area Neighbor An Area Neighbor is a Level 2 router that is outside of the Level 1 Area.

The Area Leader has several responsibilities. First, it must inject a pseudonode identifier into the Level 1 link state database. This is the Area Pseudonode Identifier. Second, the Area Leader must generate the pseudonode LSP for the Area.

All Area Edge Routers learn the Area Pseudonode Identifier from the Level 1 link state database and use that as the identifier in their Level 2 IS-IS Hello PDUs on interfaces outside the Level 1 area. The Area Edge Routers MUST also maintain an Level 2 adjacency with the Area Leader, either via a direct link or via a tunnel.

Area Edge Routers MUST be able to provide transit to Level 2 traffic. We propose that the Area Edge Routers use Segment Routing (SR) [I-D.ietf-spring-segment-routing] and, during Level 2 SPF computation, use the SR forwarding path to reach the exit Area Edge Routers.

2.1. Area Leader Election

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

2.2. LSP Generation

Area Edge Routers generate a Level 2 LSP that includes adjacencies to any Area Neighbors and the Area Leader. This LSP is not advertised outside of the area.

The Area Leader uses the Level 2 LSPs generated by the Area Edge Routers to generate the Area Pseudonode LSP. This LSP is originated using the Area Pseudonode Identifier and includes adjacencies for all of the Area Neighbors that have been advertised by the Area Edge Routers. The Area Pseudonode LSP is the only LSP that is injected into the overall Level 2 link state database, with all other Level 2 LSPs from the area being filtered out at the area boundary.

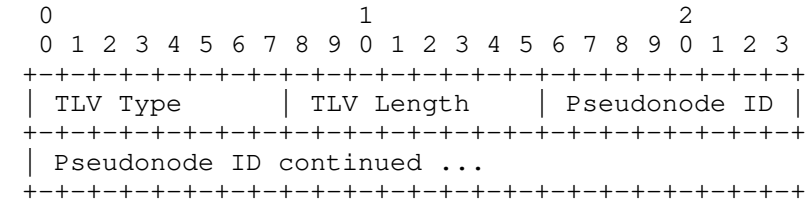
2.3. Redundancy

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

3. Area Pseudonode TLV

The Area Pseudonode TLV allows the Area Leader to advertise the existence of an Area Pseudonode Identifier. This TLV is injected into one of the Area Leader's Level 1 LSPs.

The format of the Area Pseudonode TLV is:



TLV Type: XXX

TLV Length: 2 + (length of a system ID + 1)

Pseudonode ID: A pseudonode ID, which is the length of a system ID plus one octet. field.

4. Acknowledgements

To be written.

5. 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.

6. 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.

7. References

7.1. Normative References

[I-D.ietf-spring-segment-routing]
Filsfils, C., Previdi, S., Ginsberg, L., Decraene, B., Litkowski, S., and R. Shakir, "Segment Routing Architecture", draft-ietf-spring-segment-routing-15 (work in progress), January 2018.

[I-D.li-dynamic-flooding]
Li, T. and P. Psenak, "Dynamic Flooding on Dense Graphs", draft-li-dynamic-flooding-05 (work in progress), June 2018.

[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>.

7.2. 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>.

Author's Address

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

Email: tony.li@tony.li