Internet Engineering Task Force
Internet-Draft
Intended status: Informational
Expires: July 11, 2018

T. Li
Arista Networks

January 7, 2018

Dynamic Flooding for IS-IS
draft-li-dynamic-flooding-isis-00

Abstract

Routing with link state protocols in dense network topologies can result in sub-optimal convergence times due to the overhead associated with flooding. This can be addressed by decreasing the flooding topology so that it is less dense.

This document discusses extensions to the IS-IS routing protocol to support a solution to flooding in dense subgraphs.

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 11, 2018.

Table of Contents

1. Introduction 2
1.1. Requirements Language 3
2. Area Leader TLV 3
3. Area System IDs TLV 3
4. Flooding Adjacency Matrix TLV 4
5. Acknowledgements 6
6. IANA Considerations 6
7. Security Considerations 6
8. References 7
8.1. Normative References 7
8.2. Informative References 7
Author's Address 7

1. Introduction

In recent years, there has been increased focused on how to address the dynamic routing of networks that have a bipartite (a.k.a. spine-leaf), Clos [Clos], or Fat Tree [Leiserson] topology. Conventional Interior Gateway Protocols (IGPs, i.e. IS-IS [ISO10589], OSPF [RFC5340]) under-perform, redundantly flooding information throughout the dense topology, leading to overloaded control plane inputs and thereby creating operational issues. For practical considerations, network architects have resorted to applying unconventional techniques to address the problem, applying BGP in the data center [RFC7938], however it is very clear that using an Exterior Gateway Protocol as an IGP is sub-optimal, if only due to the configuration overhead.

This problem is discussed in more detail in [Architecture], along with an architectural solution for the problem. The remainder of this document is focused on describing extensions to the IS-IS protocol to implement that architecture. Three additions appear to be necessary.

1. A TLV that an IS may inject into its LSP to indicate its preference for becoming Area Leader.

2. A TLV to carry the list of system IDs that compromise the flooding topology for the area.

3. A TLV to carry the adjacency matrix for the flooding topology for the area.

1.1. Requirements Language

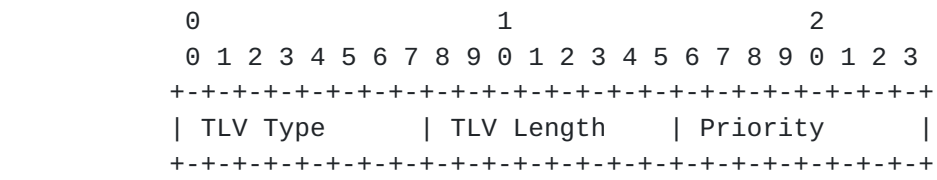
The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

2. Area Leader TLV

The Area Leader TLV allows a system to indicate its eligibility and priority for becoming Area Leader. Intermediate Systems (routers) not advertising this TLV are not eligible to become Area Leader.

The Area Leader is the router with the numerically highest Area Leader priority in the area. In the event of ties, the router with the numerically highest system ID is the Area Leader. Due to transients during database flooding, different routers may not agree on the Area Leader.

The format of the Area Leader TLV is:



TLV Type: XXX

TLV Length: 1

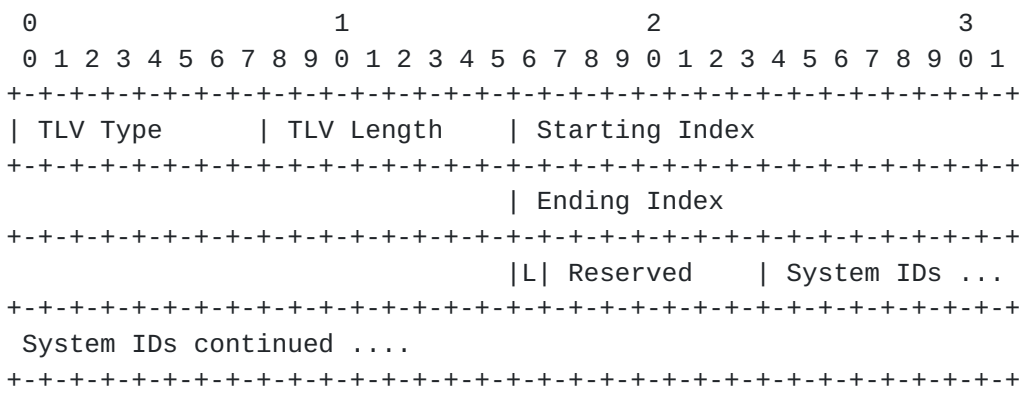
Priority: 0-255, unsigned integer

3. Area System IDs TLV

The Area System IDs TLV is used by the Area Leader to enumerate the system IDs that it has used in computing the flooding topology. Conceptually, the Area Leader creates a list of system IDs for all routers in the area, assigning indices to each system, starting with index 0.

Because the space in a single TLV is small, it may require more than one TLV to encode all of the system IDs in the area. This TLV may recur in multiple LSP segments so that all system IDs can be advertised.

The format of the Area System IDs TLV is:



TLV Type: YYY

TLV Length: 9 + (ID length * N)

Starting index: The index of the first system ID that appears in this TLV.

Ending index: The index of the last system ID that appears in this TLV.

L (Last): This bit is set if the ending index of this TLV is the last index in the full list of system IDs for the area.

System IDs: A concatenated list of system IDs for the area.

4. Flooding Adjacency Matrix TLV

The Flooding Adjacency Matrix TLV is used to describe the flooding topology for the area. This is computed and advertised by the Area Leader. Routers in the area that receive a full set of Area System IDs TLVs and a full adjacency matrix from the Area Leader should use them to compute the flooding topology and restrict their flooding to this portion of the topology. Updates that arrive from outside of the flooding topology should be flooded on the flooding topology.

The flooding topology is encoded in the adjacency matrix. The routers in the area form both the rows and columns of this matrix. A set bit in a given row and column indicates that there is connectivity between the router on the row and the router on the column. For our purposes, the links of the area can be taken to be symmetric, so the matrix is symmetric about the diagonal. The diagonal itself represents a router being connected to itself, which is not interesting and thus can be taken to always be zero. Thus, it is only necessary to encode half of the matrix. Without loss of generality, we choose to encode the upper right portion of the matrix.

The bits of the matrix are encoded by concatenating the rows of the matrix, ignoring the lower left half and diagonal of the matrix. The result is padded at the end with 0 bits to end on an octet boundary. Thus, the bit that indicates an adjacency between the router at index 0 and the router at index 1 will be the most significant bit in the first octet. The bit indicating an adjacency between the router at index 0 and the router at index 1 will be the next bit, and so forth.

As an example, consider the adjacency matrix:

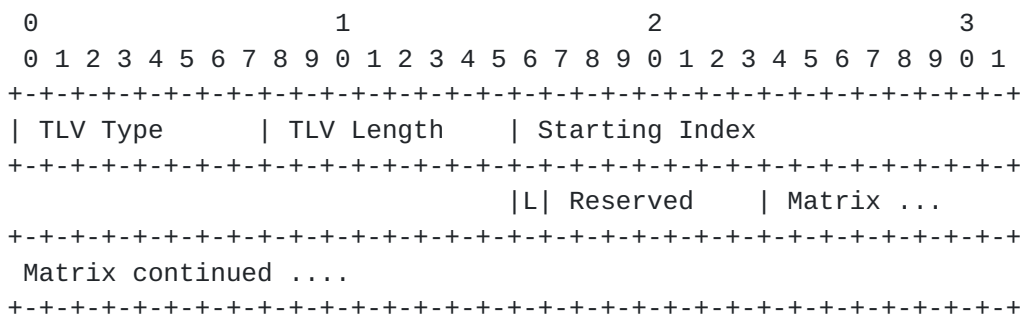
```
    0 1 2 3 4 5 6 7
   +-+-+-+-+-+-+-+-+
  0|  1 0 0 1 0 0 1|
   +-+-+-+-+-+-+-+-+
  1|    1 1 0 1 1 1|
   +-+-+-+-+-+-+-+-+
  2|      0 1 1 0 1|
   +-+-+-+-+-+-+-+-+
  3|        0 1 1 0|
   +-+-+-+-+-+-+-+-+
  4|          1 0 1|
   +-+-+-+-+-+-+-+-+
  5|            0 1|
   +-+-+-+-+-+-+-+-+
  6|              0|
   +-+-+-+-+-+-+-+-+
  7|                |
   +-+-+-+-+-+-+-+-+
```

Bits that need not be encoded are not shown. The rows of the matrix are then concatenated and padded to form octets:

```
    Row 0         Row 1       Row 2     Row 3   4     5   6 Pad
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |1 0 0 1 0 0 1|1 1 0 1 1 1|0 1 1 0 1|0 1 1 0|1 0 1|0 1|0|0 0 0 0|
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+

    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
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |1 0 0 1 0 0 1 1|1 0 1 1 1 0 1 1|0 1 0 1 1 0 1 0|1 0 1 0 0 0 0 0|
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

The Flooding Adjacency Matrix TLV then has the format:



TLV Type: 999

TLV Length: 9 + Length of Matrix octet contents

Starting index: The index of the first octet of the Matrix contents relative to the entire matrix.

L (Last): This bit is set if the last octet of the adjacency matrix is encoded in this TLV.

Matrix: The concatenated rows of the upper right triangular portion of the adjacency matrix for the flooding topology, padded with 0 bits to an octet boundary.

5. Acknowledgements

To be written.

6. IANA Considerations

This memo requests that IANA allocate and assign three code points from the IS-IS TLV Codepoints registry. One for each of the following TLVs:

1. Area Leader TLV

2. Area System IDs TLV

3. Flooding Adjacency Matrix TLV

7. 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.

8. References

8.1. Normative References

[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>.

8.2. Informative References

[Architecture] Li, T., "An Architecture for Dynamic Flooding on Dense Graphs", Internet draft draft-li-dynamic-flooding, Jan. 2018.

[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>.

[Leiserson] Leiserson, C., "Fat-Trees: Universal Networks for Hardware-Efficient Supercomputing", IEEE Transactions on Computers 34(10):892-901, 1985.

[RFC5340] Coltun, R., Ferguson, D., Moy, J., and A. Lindem, "OSPF for IPv6", RFC 5340, DOI 10.17487/RFC5340, July 2008, <https://www.rfc-editor.org/info/rfc5340>.

[RFC7938] Lapukhov, P., Premji, A., and J. Mitchell, Ed., "Use of BGP for Routing in Large-Scale Data Centers", RFC 7938, DOI 10.17487/RFC7938, August 2016, <https://www.rfc-editor.org/info/rfc7938>.

Author's Address

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

Email: tony.li@tony.li