Networking Working Group
Internet-Draft
Intended status: Standards Track
Expires: April 30, 2017

N. Shen
S. Thyamagundalu
Cisco Systems

October 27, 2016

IS-IS Routing for Spine-Leaf Topology
draft-shen-isis-spine-leaf-ext-02

Abstract

This document describes a mechanism for routers and switches in Spine-Leaf type topology to have non-reciprocal Intermediate System to Intermediate System (IS-IS) routing relationships between the leafs and spines. The leaf nodes do not need to have the topology information of other nodes and exact prefixes in the network. This extension also has application in the Internet of Things (IoT).

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 http://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 30, 2017.

Copyright Notice

Table of Contents

1. Introduction

The IS-IS routing protocol defined by [ISO10589] has been widely deployed in provider networks, data centers and enterprise campus environments. In the data center and enterprise switching networks, Spine-Leaf topology is commonly used. This document describes the mechanism where IS-IS routing can be optimized to take the advantage of the unique Spine-Leaf topology.

When the network is in Spine-Leaf topology, normally a leaf node connects to a number of spine nodes. Data traffic going from one leaf node to another leaf node needs to pass through one of the spine nodes. Also, the decision to choose one of the spine nodes is usually part of the equal cost multi-path (ECMP) load sharing. The

spine nodes can be considered as gateway devices to reach the destination leaf nodes. In this type of topologies, the spine nodes have to know the topology and routing information of the entire network, but the leaf nodes only need to know how to reach the gateway devices which are the spine nodes they are uplinked to.

This document describes the IS-IS Spine-Leaf extension that allows the spine nodes to have all the topology and routing information, while keeping the leaf nodes free of topology information other than the default gateway routing information. The leaf nodes do not even need to run their Shortest Path First (SPF) since there is no network topology to run for.

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. Motivations

- The leaf nodes in Spine-Leaf topology do not benefit much to have the complete topology and routing information of the entire domain while the forwarding actions are only to use ECMP with spine nodes as nexthops.

- The spine nodes in Spine-Leaf topology are richly connected to leaf nodes, and they need to flood every Link State PDUs (LSPs) to all the leaf nodes. It saves the spine nodes' CPU and link bandwidth resources if the flooding is blocked to those leaf nodes.

- During the time a spine node has a network problem, every leaf node connected to it will generate its LSP update to report the problem to all the other spine nodes, and those spine nodes will further flood them to all the leaf nodes, causing a O(n^2) flooding storm unnecessarily since every leaf node already knows that spine node having problem.

- Small devices and appliances of Internet of Things (IoT) can be considered as leafs in the routing topology sense. They have CPU and memory constrains in design, and those IoT devices do not have to know the exact network topology and prefixes as long as there are ways to reach the cloud servers or other devices and they want to be part of the dynamic routing.

3. Spine-Leaf (SL) Extension

3.1. Topology Example

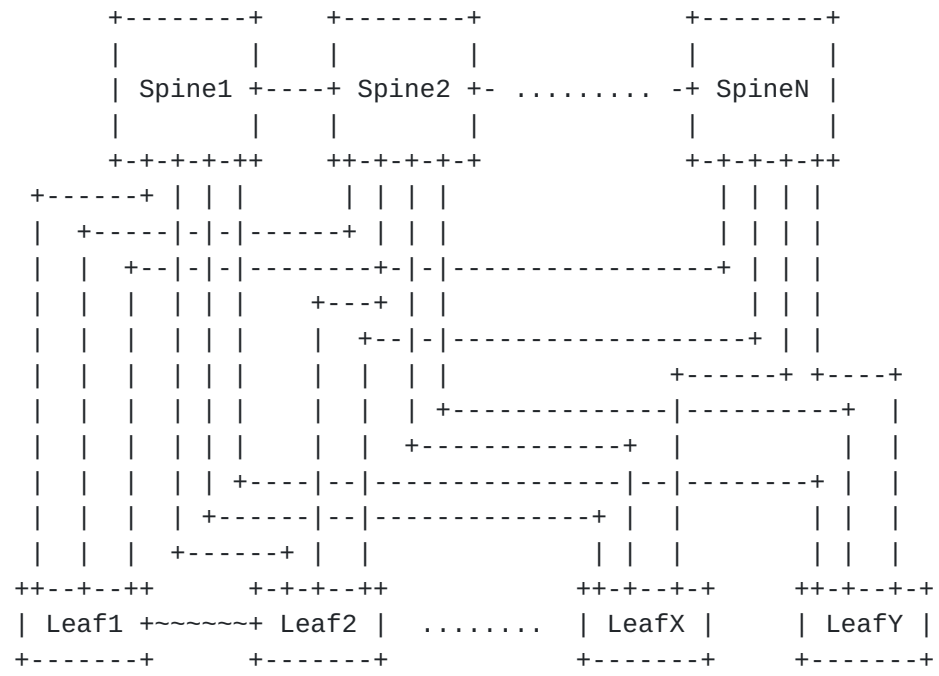


Figure 1: A Spine-Leaf Topology

3.2. Applicability Statement

This extension assumes the network is a basic Spine-Leaf topology, and it will not work in an arbitrary network setup. The spine nodes can be viewed as the aggregation layer of the network, and the leaf nodes as the access layer of the network. The leaf nodes use load sharing algorithm with spine nodes as nexthops in routing and forwarding.

This extension assumes the spine nodes are inter-connected. Spine nodes exchanges normal IS-IS topology and routing information among themselves. This extension does not apply in the case where spine nodes only have links to leaf nodes but not to themselves.

Although the example diagram in Figure 1 shows a fully meshed Spine-Leaf topology, but this extension also works in the case where they are partially meshed. For instance, the leaf1 through leaf10 are fully meshed with spine1 through spine5; and leaf11 through leaf20 are fully meshed with spine4 through spine8, and all the spines are inter-connected in a redundant fashion.

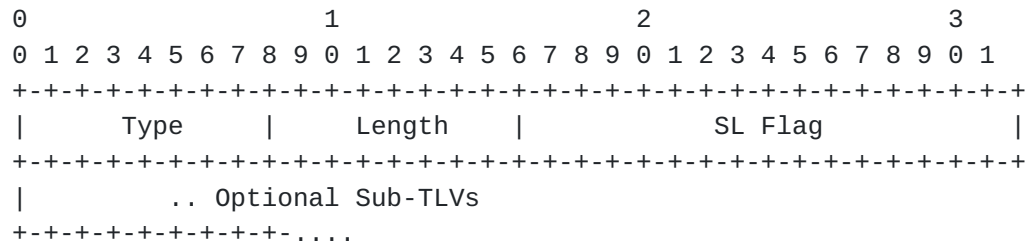
This extension also works with the topology with more than the typical two layers of spine and leaf. For instance, in example diagram Figure 1, there can be another Core layer of routers/switches on top of the aggregation layer. From an IS-IS routing point of view, the Core nodes are not affected by this extension and will have the complete topology and routing information just like the spine nodes. To make the network even more scalable, the Core layer can be run at the level-2 IS-IS domain while the Spine layer and the Leaf layer staying at the level-1 IS-IS domain.

This extension also supports the leaf nodes having local connections to other leaf nodes, in the example diagram Figure 1 there is a connection between 'Leaf1' node and 'Leaf2' node, and an external host can be dual homed into both of the leaf nodes.

This extension assumes the link between the spine and leaf nodes are point-to-point, or point-to-point over LAN [RFC5309]. The links connecting the spine nodes, or the links between the leaf nodes can be any type.

3.3. Extension Encoding

This extension introduces one TLV for IS-IS Hello (IIH) PDU and it is used by both spine and leaf nodes in the Spine-Leaf mechanism.

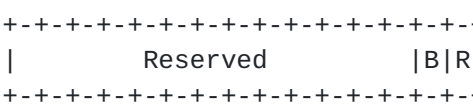


The fields of this TLV are defined as follows:

Type: TBD. 8 bits value, suggested value 150.

Length: Variable. 8 bits value. The mandatory part is 6 octets.

SL Flag: 16 bits value field of following flags:



L bit (0x01): Only leaf node sets this bit. If the L bit is set in the SL flag, the node indicates it is in 'Leaf-Mode'.

R bit (0x02): Only Spine node sets this bit. If the R bit is set, the node indicates to the leaf neighbor that it can be used as the default route gateway.

B bit (0x04): Only leaf node sets this bit on Leaf-Leaf link, in additional to the 'L' bit setting. If the B bit is set, the node indicates to its leaf neighbor that it can be used as the backup default route gateway.

Optional Sub-TLV: Not defined in this document, for future extension on SL.

3.4. Mechanism

Each leaf node is provisioned by network operators as in IS-IS 'Leaf-Mode'. A spine node does not need explicit configuration. A leaf node inserts the Spine-Leaf TLV and sets the 'L' bit in the SL flag field when sending out its IIH PDU over all its links.

The spine node when receiving the IIH with the SL TLV and 'L' bit set, it labels the point-to-point interface and adjacency to be a 'Leaf-Peer'. When the spine node sending out IIH PDU to the 'Leaf-Peer', it will also insert the Spine-Leaf TLV and set the 'R' bit in the SL flag field. This 'R' bit indicates to the 'Leaf-Peer' neighbor that the spine node can be used as a default routing nexthop.

There is no change to the IS-IS adjacency bring-up mechanism for the point-to-point interface.

For the spine node with 'Leaf-Peer' adjacencies, the IS-IS LSP flooding is blocked to the 'Leaf-Peer' interface, except for the LSP PDUs in which the IS-IS System-ID matches the System-ID of the 'Leaf-Peer' adjacency. This exception is needed since when the leaf node reboots, the spine node needs to forward to the leaf node its previous generation of LSP. No other LSP PDU needs to be flooded over this 'Leaf-Peer' interface.

The leaf node will perform IS-IS LSP flooding as normal over all of its IS-IS adjacencies, this means the leaf node will flood its own LSPs over to spine nodes since those are all the LSPs in its LSP database.

The spine node will receive all the LSP PDUs in the network, including all the spine nodes and leaf nodes. It will perform Shortest Path First (SPF) as normal IS-IS node does. There is no change to the route calculation and forwarding on the spine nodes.

But the leaf node does not have any LSP in the network except for its own, and there is no need to perform SPF algorithm on the system. It only needs to download the default route with the nexthops of those 'Spine-Peer' which has the 'R' bit set in the Spine-Leaf TLV in IIH PDUs. IS-IS can perform equal cost or unequal cost load sharing while using the spine nodes as nexthops. The aggregated metric of the outbound interface and the 'Reverse Metric' [REVERSE-METRIC] can be used for this purpose.

In summary, this extension requires leaf node to insert Spine-Leaf TLV in IIH, and set the 'L' bit in the SL flag, and download IS-IS default route using the spine nodes as nexthops where the 'Spine-Peer' set the 'R' bit in its IIH PDU; It requires spine node to respond from 'Leaf-Peer' by inserting Spine-Leaf TLV in its IIH, setting the 'R' bit in the SL flag, and blocking the LSP flooding with the exception that it will set SRMflag on the LSPs that belong to the 'Leaf-Peer' over that interface.

3.5. Implementation and Operation

3.5.1. CSNP PDU

In Spine-Leaf extension, Complete Sequence Number PDU (CSNP) does not need to be transmitted over the Spine-Leaf link. Some IS-IS implementation sends CSNPs after the initial adjacency bring-up over point-to-point interface. There is no need for this optimization here since the Leaf does not need to receive any other LSPs from the network, and the only LSPs transmitted across the Spine-Leaf link is the leaf node LSP.

Also in the graceful restart case[RFC5306], for the same reason, there is no need to send the CSNPs over the Spine-Leaf interface. It only needs to set the SRMflag on the LSPs belonging to the 'Leaf-Peer' on the spine node, and set the SRMflag on its own LSPs on the leaf node.

3.5.2. Leaf to Leaf connection

Leaf to leaf node links are useful in host redundancy cases in switching networks, and normally there is no special requirement of mechanism is needed for this case. Each leaf node will set the 'L' bit in its IIH of the Spine-Leaf flag. LSP will be exchanged over this link. In the example diagram Figure 1, the Leaf1 will get

Leaf2's LSP and Leaf2 will get Leaf1's LSP. They will install more specific routes towards each other using this local Leaf-Leaf link. SPF will be performed in this case just like when the entire network only involves with those two IS-IS nodes. This does not affect the normal Spine-Leaf mechanism they perform toward the spine nodes.

Besides the local leaf-to-leaf traffic, the leaf node can serve as a backup gateway for its leaf neighbor. It needs to remove the 'Overload-Bit' setting in its LSP, and it sets both the 'L' bit and the 'B' bit in the SL-flag with a high 'Reverse Metric' value.

3.5.3. Overload Bit

The leaf node SHOULD set the 'overload' bit on its LSP PDU, since if the spine nodes were to forward traffic not meant for the local node, the leaf node does not have the topology information to prevent a routing/forwarding loop.

3.5.4. Spine Node Hostname

This extension creates a non-reciprocal relationship between the spine node and leaf node. The spine node will receive leaf's LSP and will know the leaf's hostname, but the leaf does not have spine's LSP. This extension allows the Dynamic Hostname TLV [RFC5301] to be optionally included in spine's IIH PDU when sending to a 'Leaf-Peer'. This is useful in troubleshooting cases.

3.5.5. IS-IS Reverse Metric

This metric is part of the aggregated metric for leaf's default route installation with load sharing among the spine nodes. When a spine node is in 'overload' condition, it should use the IS-IS Reverse Metric TLV in IIH [REVERSE-METRIC] to set this metric to maximum to discourage the leaf using it as part of the loadsharing.

In some cases, certain spine nodes may have less bandwidth in link provisioning or in real-time condition, and it can use this metric to signal to the leaf nodes dynamically.

In other cases, such as when the spine node loses a link to a particular leaf node, although it can redirect the traffic to other spine nodes to reach that destination leaf node, but it MAY want to increase this metric value if the inter-spine connection becomes over utilized, or the latency becomes an issue.

In the leaf-leaf link as a backup gateway use case, the 'Reverse Metric' SHOULD always be set to very high value.

3.5.6. Other End-to-End Services

Losing the topology information will have an impact on some of the end-to-end network services, for instance, MPLS TE or end-to-end segment routing. Some other mechanisms such as those described in PCE [RFC4655] based solution may be used. In this Spine-Leaf extension, the role of the leaf node is not too much different from the multi-level IS-IS routing while the level-1 IS-IS nodes only have the default route information towards the node which has the Attach Bit (ATT) set, and the level-2 backbone does not have any topology information of the level-1 areas. The exact mechanism to enable certain end-to-end network services in Spine-Leaf network is outside the scope of this document.

3.5.7. Address Family and Topology

IPv6 Address families[RFC5308], Multi-Topology (MT)[RFC5120] and Multi-Instance (MI)[RFC6822] information is carried over the IIH PDU. Since the goal is to simplify the operation of IS-IS network, for the simplicity of this extension, the Spine-Leaf mechanism is applied the same way to all the address families, MTs and MIs.

3.5.8. Migration

For this extension to be deployed in existing networks, a simple migration scheme is needed. To support any leaf node in the network, all the involved spine nodes have to be upgraded first. So the first step is to migrate all the involved spine nodes to support this extension, then the leaf nodes can be enabled with 'Leaf-Mode' one by one. No flag day is needed for the extension migration.

4. IANA Considerations

A new TLV codepoint is defined in this document and needs to be assigned by IANA from the "IS-IS TLV Codepoints" registry. It is referred to as the Spine-Leaf TLV and the suggested value is 150. This TLV is only to be optionally inserted in the IIH PDU. This document does not propose any sub-TLV out of this Spine-Leaf TLV. IANA is also requested to maintain the SL-flag bit values in this TLV, and 0x01, 0x02 and 0x04 bits are defined in this document.

Value	Name	IIH	LSP	SNP	Purge
150	Spine-Leaf	y	n	n	n

This extension also proposes to have the Dynamic Hostname TLV, already assigned as code 137, to be allowed in IIH PDU.

Value	Name	IIH	LSP	SNP	Purge
137	Dynamic Name	y	y	n	y

5. Security Considerations

Security concerns for IS-IS are addressed in [ISO10589], [RFC5304], [RFC5310], and [RFC7602]. This extension does not raise additional security issues.

6. Acknowledgments

TBD.

7. Document Change Log

7.1. Changes to draft-shen-isis-spine-leaf-ext-02.txt

- Submitted October 2016.
- Removed the 'Default Route Metric' field in the Spine-Leaf TLV and changed to using the IS-IS Reverse Metric in IIH.

7.2. Changes to draft-shen-isis-spine-leaf-ext-01.txt

- Submitted April 2016.
- No change. Refresh the draft version.

7.3. Changes to draft-shen-isis-spine-leaf-ext-00.txt

- Initial version of the draft is published in November 2015.

8. References

8.1. Normative References

[ISO10589]
ISO "International Organization for Standardization", "Intermediate system to Intermediate system intra-domain routeing information exchange protocol for use in conjunction with the protocol for providing the connectionless-mode Network Service (ISO 8473), ISO/IEC 10589:2002, Second Edition.", Nov 2002.

[REVERSE-METRIC]
Shen, N., Amante, S., and M. Abrahamsson, "IS-IS Routing with Reverse Metric", draft-ietf-isis-reverse-metric-04 (work in progress), 2016.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <http://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, <http://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, <http://www.rfc-editor.org/info/rfc5301>.

[RFC5304] Li, T. and R. Atkinson, "IS-IS Cryptographic Authentication", RFC 5304, DOI 10.17487/RFC5304, October 2008, <http://www.rfc-editor.org/info/rfc5304>.

[RFC5306] Shand, M. and L. Ginsberg, "Restart Signaling for IS-IS", RFC 5306, DOI 10.17487/RFC5306, October 2008, <http://www.rfc-editor.org/info/rfc5306>.

[RFC5308] Hopps, C., "Routing IPv6 with IS-IS", RFC 5308, DOI 10.17487/RFC5308, October 2008, <http://www.rfc-editor.org/info/rfc5308>.

[RFC5310] Bhatia, M., Manral, V., Li, T., Atkinson, R., White, R., and M. Fanto, "IS-IS Generic Cryptographic Authentication", RFC 5310, DOI 10.17487/RFC5310, February 2009, <http://www.rfc-editor.org/info/rfc5310>.

[RFC6822] Previdi, S., Ed., Ginsberg, L., Shand, M., Roy, A., and D. Ward, "IS-IS Multi-Instance", RFC 6822, DOI 10.17487/RFC6822, December 2012, <http://www.rfc-editor.org/info/rfc6822>.

[RFC7602] Chunduri, U., Lu, W., Tian, A., and N. Shen, "IS-IS Extended Sequence Number TLV", RFC 7602, DOI 10.17487/RFC7602, July 2015, <http://www.rfc-editor.org/info/rfc7602>.

8.2. Informative References

[RFC4655] Farrel, A., Vasseur, J., and J. Ash, "A Path Computation Element (PCE)-Based Architecture", RFC 4655, DOI 10.17487/RFC4655, August 2006, <http://www.rfc-editor.org/info/rfc4655>.

[RFC5309] Shen, N., Ed. and A. Zinin, Ed., "Point-to-Point Operation over LAN in Link State Routing Protocols", RFC 5309, DOI 10.17487/RFC5309, October 2008, <http://www.rfc-editor.org/info/rfc5309>.

Authors' Addresses

Naiming Shen
Cisco Systems
560 McCarthy Blvd.
Milpitas, CA 95035
US

Email: naiming@cisco.com

Sanjay Thyamagundalu
Cisco Systems
3625 Cisco Way
San Jose, CA 95134
US

Email: sanjayt@cisco.com