Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: September 14, 2017

Z. Chen
X. Xu
Huawei Technologies

March 13, 2017

Avoiding Traffic Black-Holes for Route Aggregation in IS-IS
draft-chen-isis-black-hole-avoid-00

Abstract

When the Intermediate System to Intermediate System (IS-IS) routing protocol is adopted by a highly symmetric network such as the Leaf-Spine or Fat-Tree network, the Leaf nodes (e.g., Top of Rack switches in datacenters) are recommended to be prevented from receiving other nodes' explicit routes in order to achieve scalability. However, such a setup would cause traffic black-holes or suboptimal routing if link failure happens in the network. This document extends IS-IS to solve this problem.

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 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 September 14, 2017.

Copyright Notice

Table of Contents

1. Introduction

When running the Intermediate System to Intermediate System (IS-IS) routing protocol in a highly symmetric network such as the Leaf-Spine or Fat-Tree network, the Leaf nodes (e.g., Top of Rack switches in datacenters) are recommended to be prevented from receiving other nodes' explicit routes in order to achieve scalability, as proposed in [IS-IS-SL-Extension], [IS-IS-Overhead-Reduction], [RIFT], and [OpenFabric]. In particular, each Leaf node SHOULD simply maintain a default (or aggregated) route (e.g., 0.0.0.0/0) in its routing table, of which the next hop SHOULD be an Equal Cost Multi Path (ECMP) group including all Spines nodes that the Leaf node connects to. However, such a setup would cause traffic black-holes or suboptimal routing if link failure happens in the network, since the Leaf nodes are not aware of any topology information.

To solve this problem, this document extends IS-IS to advertise unreachable prefixes, which are defined as the prefixes that a default (or aggregated) route's next hop can no longer reach. When link failure happens between a Spine node and a Leaf node, the Spine node SHOULD advertise all prefixes attached to the Leaf node (i.e., the unreachable prefixes) to every other Leaf node it connects to. On receiving the unreachable prefixes, each Leaf node SHOULD add the

unreachable prefixes to its routing table, thus avoiding traffic black-holes and suboptimal routing.

2. Problem Description

This section illustrates why link failure would cause traffic black-hole or suboptimal routing when Leaf nodes only maintain default (or aggregated) routes.

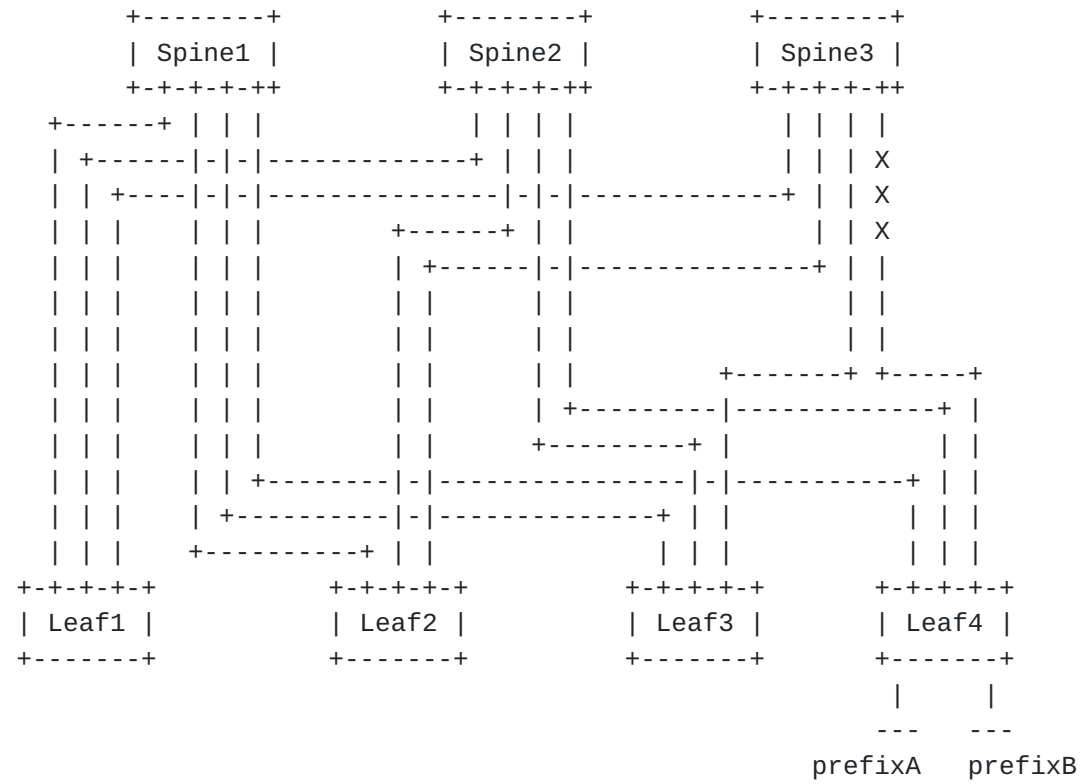


Figure 1: Topology Example

Figure 1 shows a Spine-Leaf topology example where Leaf1 to Leaf4 are connected to Spine1 to Spine3, and prefixA and prefixB are attached to Leaf4. To achieve scalability, as proposed in [IS-IS-SL-Extension], [IS-IS-Overhead-Reduction], [RIFT], and [OpenFabric], Leaf1 to Leaf4 SHOULD NOT receive explict routes from each other nor the Spine nodes. Instead, each of them maintains a default (or aggregated) route (e.g., 0.0.0.0/0) in the routing table, of which the next hop is an ECMP group including Spine1, Spine2, and Spine3. Flows from one Leaf node to another are shared among Spine1, Spine2, and Spine3 based on the well known 5-tuple hashing.

However, such a setup would cause traffic black-hole or suboptimal routing when link failure happens in the network. For example, if

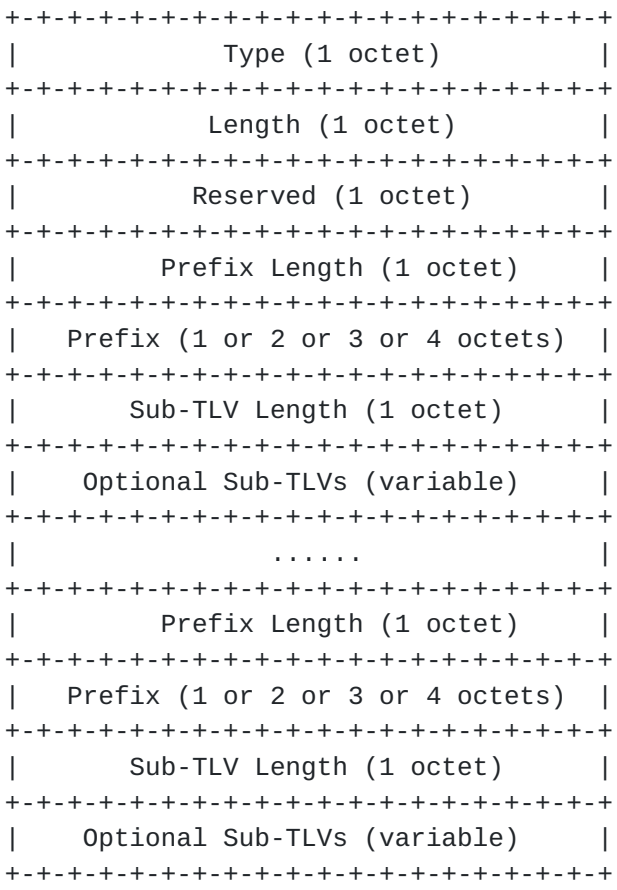
the link between Spine3 and Leaf4 is broken, Leaf1, Leaf2, and Leaf3 could not get aware of the failure. As a result, these Leaf nodes will still send a portion of traffic destined for prefixA or prefixB toward Spine3, which makes the traffic be discarded at Spine3, causing traffic black-hole. On the other hand, if there is a higher tier of switches interconnecting Spine1, Spine2, and Spine3, the traffic will be steered up to the higher-tier switches by Spine3, causing suboptimal routing.

Therefore, this document extends IS-IS to advertise unreachable prefixes thus solving this problem.

3. IS-IS Extensions

3.1. TLV Encoding

This document introduces one IS-IS TLV to advertise unreachable prefixes, called the IP Unreachability TLV, which SHOULD be carried in the IS-IS Link State Packet (LSP). The format of the IP Unreachability TLV is shown as follow:



The fields of this TLV are defined as follows:

Type: TBD.

Length: Length of the Value field of the TLV.

Reserved: Bits reserved for future usage.

Prefix Length: The value can be 0 to 32, indicating the number of effective bits in the Prefix field.

Prefix: Encoding the unreachable prefix in the minimal number of octets for the given number of effective bits (i.e., the Prefix Length field). The remaining bits of prefix SHOULD be set zero and ignored upon receipt.

Sub-TLV Length: Length of Sub-TLVs.

Sub-TLVs: Optional Sub-TLVs for future extension.

Note that the last four fields can appear repeatedly.

3.2. Unreachable Prefixes Advertisement

When link failure happens between a Spine node and a Leaf node, the Spine node SHOULD 1) encode all prefixes attached to the Leaf node (i.e., the unreachable prefixes) into the IP Unreachability TLV, 2) append the IP Unreachability TLV to the IS-IS LSP, and 3) send the LSP to every other Leaf node it connects to.

When a Leaf node receives unreachable prefixes (contained in a LSP) advertised by a Spine node, it SHOULD install each of the unreachable prefixes into its routing table, of which the next hop SHOULD be set an ECMP group including all Spine nodes it connects to except the one who advertises the unreachable prefix.

For example, if the link between Spine3 and Leaf4 in Figure 1 is broken, Spine3 SHOULD advertise prefixA and prefixB to Leaf1, Leaf2, and Leaf3, by sending them an IS-IS LSP containing the IP Unreachability TLV. On receiving the LSP, Leaf1, Leaf2, and Leaf3 SHOULD install prefixA and prefixB into their routing tables, and the next hop of prefixA or prefixB SHOULD be set an ECMP group including Spine1 and Spine2. For instance, the routing table of Leaf1 before and after the link failure is shown in Figure 2 and Figure 3, respectively.

Note that the mechanism described above could achieve minimal signaling latency, which helps to avoid black-hole or suboptimal routing rapidly when link failure happens.

Destination	Proto	Pre	Cost	Flags	NextHop	Interface
0.0.0.0/0	ISIS	15	20	D	Spine1	Ethernet0/0/0
	ISIS	15	20	D	Spine2	Ethernet0/0/1
	ISIS	15	20	D	Spine3	Ethernet0/0/2

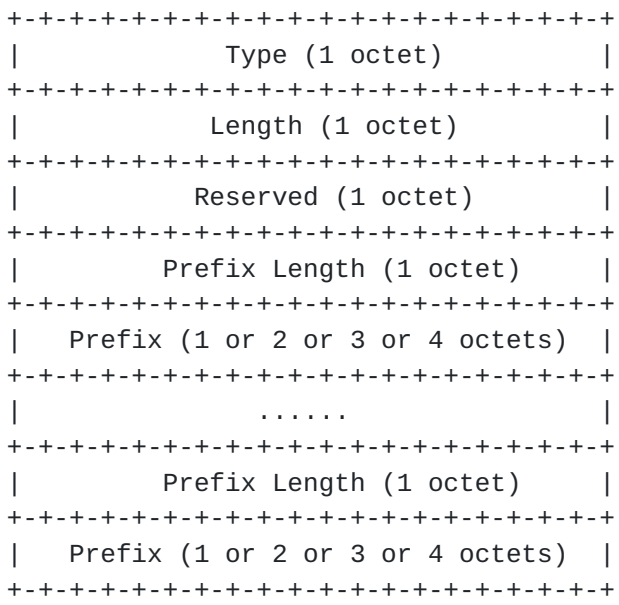
Figure 2: Routing Table of Leaf1 before link failure

Destination	Proto	Pre	Cost	Flags	NextHop	Interface
0.0.0.0/0	ISIS	15	20	D	Spine1	Ethernet0/0/0
	ISIS	15	20	D	Spine2	Ethernet0/0/1
	ISIS	15	20	D	Spine3	Ethernet0/0/2
prefixA	ISIS	15	20	D	Spine1	Ethernet0/0/0
	ISIS	15	20	D	Spine2	Ethernet0/0/1
prefixB	ISIS	15	20	D	Spine1	Ethernet0/0/0
	ISIS	15	20	D	Spine2	Ethernet0/0/1

Figure 3: Routing Table of Leaf1 after link failure

4. Alternative Solution

The unreachable prefixes can alternatively be encoded as a new Sub-TLV of the Extended IP Reachability TLV defined in [RFC 5305]. The format of the Sub-TLV is shown as follow:



The fields of this Sub-TLV are defined as follows:

Type: TBD.

Length: Length of the Value field of the Sub-TLV.

Reserved: Bits reserved for future usage.

Prefix Length: The value can be 0 to 32, indicating the number of effective bits in the Prefix field.

Prefix: Encoding the unreachable prefix in the minimal number of octets for the given number of effective bits (i.e., the Prefix Length field). The remaining bits of prefix SHOULD be set zero and ignored upon receipt.

Note that the last two fields can appear repeatedly.

When link failure happens between a Spine node and a Leaf node, the Spine node SHOULD 1) encode all prefixes attached to the Leaf node (i.e., the unreachable prefixes) into the Sub-TLV described above, 2) encode the Sub-TLV into the Extended IP Reachability TLV, 3) append the Extended IP Reachability TLV to the IS-IS LSP, and 4) send the LSP to every other Leaf node it connects to. The Prefix field of the Extended IP Reachability TLV SHOULD be set the default (or aggregated) route that each of the Leaf nodes already maintains.

When a Leaf node receives unreachable prefixes (contained in a LSP) advertised by a Spine node, it SHOULD install each of the unreachable prefixes into its routing table, of which the next hop SHOULD be set

an ECMP group including all Spine nodes it connects to except the one who advertises the unreachable prefix.

5. IPv6 Support

Will be completed in the next version of the document.

6. IANA Considerations

TBD.

7. Security Considerations

TBD.

8. Acknowledgements

TBD.

9. References

[IS-IS-Overhead-Reduction]
Chen, Z. and X. Xu, "Overheads Reduction for IS-IS Enabled Spine-Leaf Networks", draft-chen-isis-sl-overheads-reduction-00 (work in progress) , January 2017.

[IS-IS-SL-Extension]
Shen, N. and S. Thyamagundalu, "IS-IS Routing for Spine-Leaf Topology", draft-shen-isis-spine-leaf-ext-02 (work in progress) , October 2016.

[OpenFabric]
White, R. and S. Zandi, "OpenFabric", draft-white-openfabric-00 (work in progress) , March 2017.

[RFC1195] Callon, R., "Use of OSI IS-IS for Routing in TCP/IP and Dual Environments", RFC 1195 , December 1990.

[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>.

[RFC5305] Li, T. and H. Smit, "IS-IS Extensions for Traffic Engineering", RFC 5305 , October 2008.

[RIFT] Przygienda, T., Drake, J., and A. Atlas, "RIFT: Routing in Fat Trees", draft-przygienda-rift-01 (work in progress) , January 2017.

Authors' Addresses

Zhe Chen
Huawei Technologies
No. 156 Beiqing Rd
Beijing 100095
China

Email: chenzhe17@huawei.com

Xiaohu Xu
Huawei Technologies
No. 156 Beiqing Rd
Beijing 100095
China

Email: xuxiaohu@huawei.com