

Network Working Group  
Internet-Draft  
Intended status: Informational  
Expires: May 9, 2019

R. White, Ed.  
S. Zandi, Ed.  
LinkedIn  
November 5, 2018

IS-IS Support for Openfabric  
draft-white-openfabric-07

Abstract

Spine and leaf topologies are widely used in hyperscale and cloud scale networks. In most of these networks, configuration is automated, but difficult, and topology information is extracted through broad based connections. Policy is often integrated into the control plane, as well, making configuration, management, and troubleshooting difficult. Openfabric is an adaptation of an existing, widely deployed link state protocol, Intermediate System to Intermediate System (IS-IS) that is designed to:

- o Provide a full view of the topology from a single point in the network to simplify operations
- o Minimize configuration of each Intermediate System (IS) (also called a router or switch) in the network
- o Optimize the operation of IS-IS within a spine and leaf fabric to enable scaling

This document begins with an overview of openfabric, including a description of what may be removed from IS-IS to enable scaling. The document then describes an optimized adjacency formation process; an optimized flooding scheme; some thoughts on the operation of openfabric, metrics, and aggregation; and finally a description of the changes to the IS-IS protocol required for openfabric.

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 May 9, 2019.

#### Copyright Notice

Copyright (c) 2018 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 . . . . .	3
1.1.	Goals . . . . .	3
1.2.	Contributors . . . . .	3
1.3.	Simplification . . . . .	3
1.4.	Additions and Requirements . . . . .	4
1.5.	Sample Network . . . . .	4
2.	Modified Adjacency Formation . . . . .	6
2.1.	Level 2 Adjacencies Only . . . . .	6
2.2.	Point-to-point Adjacencies . . . . .	6
2.3.	Three Way Handshake Support . . . . .	7
2.4.	Adjacency Formation Optimization . . . . .	7
3.	Advertisement of Reachability Information . . . . .	8
4.	Determining and Advertising Location on the Fabric . . . . .	9
5.	Flooding Optimization . . . . .	10
5.1.	Flooding Failures . . . . .	11
6.	Other Optimizations . . . . .	12
6.1.	Transit Link Reachability . . . . .	12
6.2.	Transiting T0 Intermediate Systems . . . . .	12
7.	Openfabric and Route Aggregation . . . . .	13
8.	Security Considerations . . . . .	13
9.	References . . . . .	13
9.1.	Normative References . . . . .	13
9.2.	Informative References . . . . .	15
	Appendix A. Flooding Optimization Operation . . . . .	17
	Appendix B. Fabric Location Calculation . . . . .	19
	Authors' Addresses . . . . .	20

## 1. Introduction

### 1.1. Goals

Spine and leaf fabrics are often used in large scale data centers; in this application, they are commonly called a fabric because of their regular structure and predictable forwarding and convergence properties. This document describes modifications to the IS-IS protocol to enable it to run efficiently on a large scale spine and leaf fabric, openfabric. The goals of this control plane are:

- o Provide a full view of the topology from a single point in the network to simplify operations
- o Minimize configuration of each IS in the network
- o Optimize the operation of IS-IS within a spine and leaf fabric to enable scaling

### 1.2. Contributors

The following people have contributed to this draft: Nikos Triantafyllis (reflected flooding optimization), Ivan Pepelnjak (fabric locality calculation modifications), Christian Franke (fabric localigy calculation modification), Hannes Gredler (do not reflood optimizations), Les Ginsberg (capabilities encoding, circuit local reflooding), Naiming Shen (capabilities encoding, circuit local reflooding), Uma Chunduri (failure mode suggestions, flooding), Nick Russo, and Rodny Molina.

See [RFC5449], [RFC5614], and [RFC7182] for similar solutions in the Mobile Ad Hoc Networking (MANET) solution space.

### 1.3. Simplification

In building any scalable system, it is often best to begin by removing what is not needed. In this spirit, openfabric implementations MAY remove the following from IS-IS:

- o External metrics. There is no need for external metrics in large scale spine and leaf fabrics; it is assumed that metrics will be properly configured by the operator to account for the correct order of route preference at any route redistribution point.
- o Tags and traffic engineering processing. Openfabric is only designed to provide topology and reachability information. It is not designed to provide for traffic engineering, route preference through tags, or other policy mechanisms. It is assumed that all

routing policy will be provided through an overlay system which communicates directly with each IS in the fabric, such as PCEP [RFC5440] or I2RS [RFC7921]. Traffic engineering is assumed to be provided through Segment Routing (SR) [I-D.ietf-spring-segment-routing].

#### 1.4. Additions and Requirements

To create a scalable link state fabric, openfabric includes the following:

- o A slightly modified adjacency formation process.
- o Mechanisms for determining which tier within a spine and leaf fabric in which the IS is located.
- o A mechanism that reduces flooding to the minimum possible, while still ensuring complete database synchronization among the intermediate systems within the fabric.

Three general requirements are placed here; more specific requirements are considered in the following sections. Openfabric implementations:

- o MUST support [RFC5301] and enable hostname advertisement by default if a hostname is configured on the intermediate system.
- o SHOULD support [RFC6232], purge originator identification for IS-IS.
- o MUST NOT be mixed with standard IS-IS implementations in operational deployments. Openfabric and standard IS-IS implementations SHOULD be treated as two separate protocols.

#### 1.5. Sample Network

The following spine and leaf fabric will be used to describe these modifications.

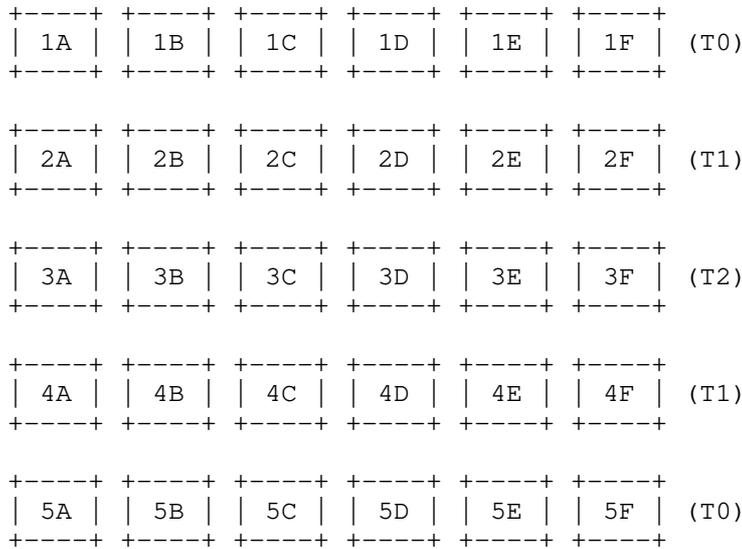


Figure 1

To reduce confusion (spine and leaf fabrics are difficult to draw in plain text art), this diagram does not contain the connections between devices. The reader should assume that each device in a given layer is connected to every device in the layer above it. For instance:

- o 5A is connected to 4A, 4B, 4C, 4D, 4E, and 4F
- o 5B is connected to 4A, 4B, 4C, 4D, 4E, and 4F
- o 4A is connected to 3A, 3B, 3C, 3D, 3E, 3F, 5A, 5B, 5C, 5D, 5E, and 5F
- o 4B is connected to 3A, 3B, 3C, 3D, 3E, 3F, 5A, 5B, 5C, 5D, 5E, and 5F
- o etc.

The tiers or stages of the fabric are also marked for easier reference. T0 is assumed to be connected to application servers, or rather they are Top of Rack (ToR) intermediate systems. The remaining tiers, T1 and T2, are connected only to the fabric itself. Note there are no "cross links," or "east west" links in the illustrated fabric. The fabric locality detection mechanism described here will not work if there are cross links running east/

west through the fabric. Locality detection may be possible in such a fabric; this is an area for further study.

## 2. Modified Adjacency Formation

Because Openfabric operates in a tightly controlled data center environment, various modifications can be made to the IS-IS neighbor formation process to increase efficiency and simplify the protocol. Specifically, Openfabric implementations SHOULD support [RFC3719], section 4, hello padding for IS-IS. Variable hello padding SHOULD NOT be used, as data center fabrics are built using high speed links on which padded hellos will have little performance impact. Further modifications to the neighbor formation process are considered in the following sections.

### 2.1. Level 2 Adjacencies Only

Openfabric is designed to work in a single flooding domain over a single data center fabric at the scale of thousands of routers with hundreds of thousands of routes (so a moderate scale in router and route count terms). Because of the way Openfabric optimizes operation in this environment, it is not necessary nor desirable to build multiple flooding domains. For instance, the flooding optimizations described later in this document require a full view of the topology, as does any proposed overlay to inject policy into the forwarding plane. In light of this, the following changes SHOULD BE to IS-IS implementations to support Openfabric:

- o IIH PDU 17 (level 2 point-to-point circuit hello) should be the only IIH PDU type transmitted (see section 9.7 of ISO 10589)
- o In IIH PDU 17 (level 2 point-to-point circuit hello), the Circuit Type field should be set to 2 (see section 9.7 of ISO 10589)
- o Support for IIH PDU 15 (level 1 broadcast hello) should be removed (see section 9.5 of ISO 10589)
- o Support for IIH PDU 16 (level 2 broadcast hello) should be removed (see section 9.6 of ISO 10589)

### 2.2. Point-to-point Adjacencies

Data center network fabrics only contain point-to-point links; because of this, there is no reason to support any broadcast link types, nor to support the Designated Intermediate System processing, including pseudonode creation. In light of this, processing related to sections 7.2.3 (broadcast networks), 7.3.8 (generation of level 1 pseudonode LSPs), 7.3.10 (generation of level 2 pseudonode LSPs), and

section 8.4.5 (LAN designated intermediate systems) in [ISO10589] SHOULD BE removed.

### 2.3. Three Way Handshake Support

It is important that two way connectivity be established before synchronizing the link state database, or routing through a link in a data center fabric. To reject optical failures that cause a one way connection between two routers, fabricDC must support the three way handshake mechanism described in [RFC5303].

### 2.4. Adjacency Formation Optimization

While adjacency formation is not considered particularly burdensome in IS-IS, it may still be useful to reduce the amount of state transferred across the network when connecting a new IS to the fabric. In its simplest form, the process is:

- o An IS connected to the fabric will send hellos on all links.
- o The IS will only complete the three-way handshake with one newly discovered neighbor; this would normally be the first neighbor which sends the newly connected intermediate system's ID back in the three-way handshake process.
- o The IS will complete its database exchange with this one newly adjacent neighbor.
- o Once this process is completed, the IS will continue processing the remaining neighbors as normal.
- o If synchronization is not achieved within twice the dead timer on the local interface, the newly connected IS will repeat this process with the second neighbor with which it forms a three-way adjacency.

This process allows each IS newly added to the fabric to exchange a full table once; a very minimal amount of information will be transferred with the remaining neighbors to reach full synchronization.

Any such optimization is bound to present a tradeoff between several factors; the mechanism described here increases the amount of time required to form adjacencies slightly in order to reduce the total state carried across the network. An alternative mechanism could provide a better balance of the amount of information carried across the network for initial synchronization and the time required to synchronize a new IS. For instance, an IS could choose to

synchronize its database with two or three adjacent intermediate systems, which could speed the synchronization process up at the cost of carrying additional data on the network. A locally determined balance between the speed of synchronization and the amount of data carried on the network can be achieved by adjusting the number of adjacent intermediate systems the newly attached IS synchronizes with.

### 3. Advertisement of Reachability Information

IS-IS describes the topology in two different sets of TLVs; the first describes the set of neighbors connected to an IS, the second describes the set of reachable destination connected to an IS. There are two different forms of both of these descriptions, one of which carries what are widely called narrow metrics, the other of which carries what are widely called wide metrics. In a tightly controlled data center fabric implementation, such as the ones Openfabric is designed to support, no IS that supports narrow metrics will ever be deployed or supported; hence there is no reason to support any metric type other than wide metrics.

- o The Level 2 Link State PDU (type 20 in section 9.9 of [ISO10589]) and the scoped flooding PDU (type 10 in section 3.1 of [RFC7356]) SHOULD BE the only PDU types used to carry link state information in a Openfabric implementation
- o Processing related to the Level 1 Link State PDU (type 18) MAY BE removed from Openfabric implementations (see section 9.8 of [ISO10589])
- o Neighbor reachability MUST BE carried in TLV type 22 (see section 3 of [RFC5305])
- o IPv4 reachability SHOULD BE carried in TLV type 135 (see section 4 of [RFC5305]), or TLV type 235 for multitopology implementations (see [RFC5120])
- o IPv6 reachability SHOULD BE carried in TLV type 236 (see [RFC5308]), or TLV type 237 for multitopology implementations (see [RFC5120])
- o Processing related to the neighbor reachability TLV (type 2, see sections 9.8 and 9.9 of [ISO10589]) SHOULD BE removed
- o Processing related to the narrow metric IP reachability TLV (types 128 and 130) SHOULD BE removed

Further, if segment routing support is desired, Openfabric MAY support the Prefix Segment Identifier sub-TLV and other TLVs as required in [I-D.ietf-isis-segment-routing-extensions].

#### 4. Determining and Advertising Location on the Fabric

The tier to which a IS is connected is useful to enable autoconfiguration of intermediate systems connected to the fabric and to reduce flooding. Once the tier of an intermediate system within the fabric has been determined, it MUST be advertised using the 4 bit Tier field described in section 3.3 of [I-D.shen-isis-spine-leaf-ext]. This section describes a method of calculating the tier number, assuming the tier numbers rise in value from the edge of the fabric.

This method begins with two of the T0 intermediate systems advertising their location in the fabric. This information can either be obtained through:

- o Two T0 intermediate systems are manually configured to advertise 0x00 in their IS reachability tier sub-TLV, indicating they are at the edge of the fabric (a ToR IS).
- o The T0 intermediate systems detect they are T0 through the presence connected hosts (i.e. through a request for address assignment or some other means). If such detection is used, and the IS determines it is located at T0, it should advertise 0x00 in its IS reachability tier sub-TLV.

If the first method is used, the two T0 routers MUST be "maximally separated" on the fabric. They must be a maximal number of hops apart, or rather they MUST NOT be connected to the same T1 device as their "upstream" towards the superspines in a 5 ary fabric.

The second method above SHOULD be used with care, as it may not be secure, and it may not work in all data center environments. For instance, if a host is mistakenly (or intentionally, as a form of attack) attached to a spine IS, or a request for address assignment is transmitted to a spine IS during the bootup phase of the device or fabric, it is possible to cause a spine IS to advertise itself as a T0. Unless the autodetection of the T0 devices is secured, the manual mechanism SHOULD BE used (configuring at least one T0 device manually).

Given the correct configuration of two T0 devices, maximally spaced on the fabric, the remaining intermediate systems calculate their tier number as follows:

- o The local IS calculates an SPT (using SPF) setting the cost of every link to 1; this effectively calculates a topology only view of the network, without considering any configured link costs
- o Ensure that at least two T0 are in the calculated SPT; otherwise abort
- o Find the furthest T0; call this node A and set LD to the cost; the "furthest T0" is the T0 with the largest metric, or the furthest distance from the local calculating node
- o Calculate an SPT (using SPF) from the perspective of A (above) setting the cost of every link to 1
- o Find the furthest IS in A's SPT; call this node B and set RD to the cost from A to B
- o Calculate the tier number of the local IS by subtracting LD from RD

In the example network, assume 5A and 1C are manually configured as a T0, and are advertising their tier numbers. From here:

- o From 1A the path to 5A is 4 hops; this is LD
- o Run SPF from the perspective of 5A with all link metrics set to 1
- o From 5A the path length to 1C is 4; this is RD
- o  $RD - LD$  is 0 at 1A, so 1A is T0, or a ToR

This process will work for any spine and leaf fabric without "cross links."

## 5. Flooding Optimization

Flooding is perhaps the most challenging scaling issue for a link state protocol running on a dense, large scale fabric. To reduce the flooding of link state information in the form of Link State Protocol Data Units (LSPs), Openfabric takes advantage of information already available in the link state protocol, the list of the local intermediate system's neighbor's neighbors, and the fabric locality computed above. The following tables are required to compute a set of reflooders:

- o Neighbor List (NL) list: The set of neighbors

- o Neighbor's Neighbors (NN) list: The set of neighbor's neighbors; this can be calculated by running SPF truncated to two hops
- o Do Not Reflood (DNR) list: The set of neighbors who should have LSPs (or fragments) who should not reflood LSPs
- o Reflood (RF) list: The set of neighbors who should flood LSPs (or fragments) to their adjacent neighbors to ensure synchronization

NL is set to contain all neighbors, and sorted deterministically (for instance, from the highest IS identifier to the lowest). All intermediate systems within a single fabric SHOULD use the same mechanism for sorting the NL list. NN is set to contain all neighbor's neighbors, or all intermediate systems that are two hops away, as determined by performing a truncated SPF. The DNR and RF tables are initially empty. To begin, the following steps are taken to reduce the size of NN and NL:

- o Move any IS in NL with its tier (or fabric location) set to T0 to DNR
- o Remove all intermediate systems from NL and NN that in the shortest path to the IS that originated the LSP

Then, for every IS in NL:

- o If the current entry in NL is connected to any entries in NN:
  - \* Move the IS to RF
  - \* Remove the intermediate systems connected to the IS from NN
- o Else move the IS to DNR

The calculation terminates when the NL is empty.

When flooding, LSPs transmitted to adjacent neighbors on the RF list will be transmitted normally. Adjacent intermediate systems on this list will reflood received LSPs into the next stage of the topology, ensuring database synchronization. LSPs transmitted to adjacent neighbors on the DNR list, however, MUST be transmitted using a circuit scope PDU as described in [RFC7356].

### 5.1. Flooding Failures

It is possible in some failure modes for flooding to be incomplete because of the flooding optimizations outlined. Specifically, if a reflooder fails, or is somehow disconnected from all the links across

which it should be reflooding, it is possible an LSP is only partially flooded through the fabric. To prevent such situations, any IS receiving an LSP transmitted using DNR SHOULD:

- o Set a short timer; the default should be less than one second
- o When the timer expires, send a Complete Sequence Number Packet (CSNP) to all neighbors
- o Process any Partial Sequence Number Packets (PSNPs) as required to resynchronize
- o If a resynchronization is required, notify the network operator through a network management system

## 6. Other Optimizations

### 6.1. Transit Link Reachability

In order to reduce the amount of control plane state carried on large scale spine and leaf fabrics, openfabric implementations SHOULD NOT advertise reachability for transit links. These links MAY remain unnumbered, as IS-IS does not require layer 3 IP addresses to operate. Each IS SHOULD be configured with a single loopback address, which is assigned an IPv6 address, to provide reachability to intermediate systems which make up the fabric.

[RFC3277] SHOULD be supported on devices supporting openfabric with unnumbered interface in order to support traceability and network management.

### 6.2. Transiting T0 Intermediate Systems

In data center fabrics, ToR intermediate systems SHOULD NOT be used to transit between two T1 (or above) spine intermediate systems. The simplest way to prevent this is to set the overload bit [RFC3277] for all the LSPs originated from T0 intermediate systems. However, this solution would have the unfortunate side effect of causing all reachability beyond any T0 IS to have the same metric, and many implementations treat a set overload bit as a metric of 0xFFFF in calculating the Shortest Path Tree (SPT). This document proposes an alternate solution which preserves the leaf node metric, while still avoiding transiting T0 intermediate systems.

Specifically, all T0 intermediate systems SHOULD advertise their metric to reach any T1 adjacent neighbor with a cost of 0XFFE. T1 intermediate systems, on the other hand, will advertise T0 intermediate systems with the actual interface cost used to reach the

T0 IS. Hence, links connecting T0 and T1 intermediate systems will be advertised with an asymmetric cost that discourages transiting T0 intermediate systems, while leaving reachability to the destinations attached to T0 devices the same.

## 7. Openfabric and Route Aggregation

While schemes may be designed so reachability information can be aggregated in Openfabric deployments, this is not a recommended configuration.

## 8. Security Considerations

This document outlines modifications to the IS-IS protocol for operation on large scale data center fabrics. While it does add new TLVs, and some local processing changes, it does not add any new security vulnerabilities to the operation of IS-IS. However, openfabric implementations SHOULD implement IS-IS cryptographic authentication, as described in [RFC5304], and should enable other security measures in accordance with best common practices for the IS-IS protocol.

If T0 intermediate systems are auto-detected using information outside Openfabric, it is possible to attack the calculations used for flooding reduction and auto-configuration of intermediate systems. For instance, if a request for an address pool is used as an indicator of an attached host, and hence receiving such a request causes an intermediate system to advertise itself as T0, it is possible for an attacker (or a simple mistake) to cause auto-configuration to fail. Any such auto-detection mechanisms SHOULD BE secured using appropriate techniques, as described by any protocols or mechanisms used.

## 9. References

### 9.1. Normative References

[I-D.shen-isis-spine-leaf-ext]

Shen, N., Ginsberg, L., and S. Thyamagundalu, "IS-IS Routing for Spine-Leaf Topology", draft-shen-isis-spine-leaf-ext-07 (work in progress), October 2018.

- [ISO10589] International Organization for Standardization, "Intermediate system to Intermediate system intra-domain routing 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.
- [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>>.
- [RFC2629] Rose, M., "Writing I-Ds and RFCs using XML", RFC 2629, DOI 10.17487/RFC2629, June 1999, <<https://www.rfc-editor.org/info/rfc2629>>.
- [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>>.
- [RFC5303] Katz, D., Saluja, R., and D. Eastlake 3rd, "Three-Way Handshake for IS-IS Point-to-Point Adjacencies", RFC 5303, DOI 10.17487/RFC5303, October 2008, <<https://www.rfc-editor.org/info/rfc5303>>.
- [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>>.
- [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, <<https://www.rfc-editor.org/info/rfc5309>>.

- [RFC5311] McPherson, D., Ed., Ginsberg, L., Previdi, S., and M. Shand, "Simplified Extension of Link State PDU (LSP) Space for IS-IS", RFC 5311, DOI 10.17487/RFC5311, February 2009, <<https://www.rfc-editor.org/info/rfc5311>>.
- [RFC5316] Chen, M., Zhang, R., and X. Duan, "ISIS Extensions in Support of Inter-Autonomous System (AS) MPLS and GMPLS Traffic Engineering", RFC 5316, DOI 10.17487/RFC5316, December 2008, <<https://www.rfc-editor.org/info/rfc5316>>.
- [RFC7356] Ginsberg, L., Previdi, S., and Y. Yang, "IS-IS Flooding Scope Link State PDUs (LSPs)", RFC 7356, DOI 10.17487/RFC7356, September 2014, <<https://www.rfc-editor.org/info/rfc7356>>.
- [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

- [I-D.ietf-isis-segment-routing-extensions]  
Previdi, S., Ginsberg, L., Filsfils, C., Bashandy, A., Gredler, H., Litkowski, S., Decraene, B., and J. Tantsura, "IS-IS Extensions for Segment Routing", draft-ietf-isis-segment-routing-extensions-19 (work in progress), July 2018.
- [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.
- [RFC3277] McPherson, D., "Intermediate System to Intermediate System (IS-IS) Transient Blackhole Avoidance", RFC 3277, DOI 10.17487/RFC3277, April 2002, <<https://www.rfc-editor.org/info/rfc3277>>.
- [RFC3719] Parker, J., Ed., "Recommendations for Interoperable Networks using Intermediate System to Intermediate System (IS-IS)", RFC 3719, DOI 10.17487/RFC3719, February 2004, <<https://www.rfc-editor.org/info/rfc3719>>.

- [RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, DOI 10.17487/RFC4271, January 2006, <<https://www.rfc-editor.org/info/rfc4271>>.
- [RFC5304] Li, T. and R. Atkinson, "IS-IS Cryptographic Authentication", RFC 5304, DOI 10.17487/RFC5304, October 2008, <<https://www.rfc-editor.org/info/rfc5304>>.
- [RFC5440] Vasseur, JP., Ed. and JL. Le Roux, Ed., "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, DOI 10.17487/RFC5440, March 2009, <<https://www.rfc-editor.org/info/rfc5440>>.
- [RFC5449] Baccelli, E., Jacquet, P., Nguyen, D., and T. Clausen, "OSPF Multipoint Relay (MPR) Extension for Ad Hoc Networks", RFC 5449, DOI 10.17487/RFC5449, February 2009, <<https://www.rfc-editor.org/info/rfc5449>>.
- [RFC5614] Ogier, R. and P. Spagnolo, "Mobile Ad Hoc Network (MANET) Extension of OSPF Using Connected Dominating Set (CDS) Flooding", RFC 5614, DOI 10.17487/RFC5614, August 2009, <<https://www.rfc-editor.org/info/rfc5614>>.
- [RFC5820] Roy, A., Ed. and M. Chandra, Ed., "Extensions to OSPF to Support Mobile Ad Hoc Networking", RFC 5820, DOI 10.17487/RFC5820, March 2010, <<https://www.rfc-editor.org/info/rfc5820>>.
- [RFC5837] Atlas, A., Ed., Bonica, R., Ed., Pignataro, C., Ed., Shen, N., and JR. Rivers, "Extending ICMP for Interface and Next-Hop Identification", RFC 5837, DOI 10.17487/RFC5837, April 2010, <<https://www.rfc-editor.org/info/rfc5837>>.
- [RFC6232] Wei, F., Qin, Y., Li, Z., Li, T., and J. Dong, "Purge Originator Identification TLV for IS-IS", RFC 6232, DOI 10.17487/RFC6232, May 2011, <<https://www.rfc-editor.org/info/rfc6232>>.
- [RFC7182] Herberg, U., Clausen, T., and C. Dearlove, "Integrity Check Value and Timestamp TLV Definitions for Mobile Ad Hoc Networks (MANETs)", RFC 7182, DOI 10.17487/RFC7182, April 2014, <<https://www.rfc-editor.org/info/rfc7182>>.
- [RFC7921] Atlas, A., Halpern, J., Hares, S., Ward, D., and T. Nadeau, "An Architecture for the Interface to the Routing System", RFC 7921, DOI 10.17487/RFC7921, June 2016, <<https://www.rfc-editor.org/info/rfc7921>>.

## Appendix A. Flooding Optimization Operation

Recent testing has shown that flooding is largely a "non-issue" in terms of scaling when using high speed links connecting intermediate systems with reasonable processing power and memory. However, testing has also shown that flooding will impact convergence speed even in such environments, and flooding optimization has a major impact on the performance of a link state protocol in resource constrained environments. Some thoughts on flooding optimization in general, and the flooding optimization contained in this document, follow.

There are two general classes of flooding optimization available for link state protocols. The first class of optimization relies on a centralized service or server to gather the link state information and redistribute it back into the intermediate systems making up the fabric. Such solutions are attractive in many, but not all, environments; hence these systems compliment, rather than compete with, the system described here. Systems relying on a service or server necessarily also rely on connectivity to that service or server, either through an out-of-band network or connectivity through the fabric itself. Because of this, these mechanisms do not apply to all deployments; some deployments require underlying reachability regardless of connectivity to an outside service or server.

The second possibility is to create a fully distributed system that floods the minimal amount of information possible to every intermediate system. The system described in this draft is an example of such a system. Again, there are many ways to accomplish this goal, but simplicity is a primary goal of the system described in this draft.

The system described here divides the work into two different parts; forward and reverse optimization. The forward optimization begins by finding the set of intermediate systems two hops away from the flooding device, and choosing a subset of connected neighbors that will successfully reach this entire set of intermediate systems, as shown in the diagram below.

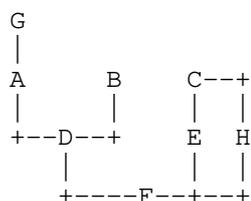


Figure 2

If F is flooding some piece of information, then it will find the entire set of intermediate systems within two hops by discovering its neighbors and their neighbors from the local LSDB. This will include A, B, C, D, and E--but not G. From this set, F can determine that D can reach A and B, while a single flood to either E or H will reach C. Hence F can flood to D and either E or H to reach C. F can choose to flood to D and E normally. Because H still needs to receive this new LSP (or fragment!), but does not need to reflood to C, F can send the LSP using link local signaling. In this case, H will receive and process the new LSP, but not reflood it.

Rather than carrying the information necessary through hello extensions, as is done in [RFC5820], the neighbors are allowed to complete initial synchronization, and then a truncated shortest path tree is built to determine the "two hop neighborhood." This has the advantage of using mechanisms already used in IS-IS, rather than adding new processes. The risk with this process is any LSPs flooded through the network before this initial calculation takes place will be suboptimal. This "two hop neighborhood" process has been used in OSPF deployments for a number of years, and has proven stable in practice.

Rather than setting a timer for reflooding, the implementation described here uses IS-IS' ability to describe the entire database using a CSNP to ensure flooding is successful. This adds some small amount of overhead, so there is some balance between optimal flooding and ensuring flooding is complete.

The reverse optimization is simpler. It relies on the observation that any intermediate system between the local IS and the origin of the LSP, other than in the case of floods removing an LSP from the shared LSDB, should have already received a copy of the LSP. For instance, if F originates an LSP in the figure above, and E refloods the LSP to C, C does not need to reflood back to F if F is on its shortest path tree towards F. It is obvious this is not a "perfect" optimization. A perfect optimization would block flooding back along a directed acyclic graph towards the originator. Using the SPT, however, is a quick way to reduce flooding without performing more calculations.

The combination of these two optimizations have been seen, in testing, to reduce the number of copies any IS receives from the tens to precisely one.

## Appendix B. Fabric Location Calculation

Determining the location of a device in a symmetric topology is quite challenging. The authors of this draft worked through a number of possible solutions to this problem, each of which was found to either not work in some topology, or was found to be liable to unacceptable errors. For instance:

## o Method 1:

- \* Calculate the maximum distance through the fabric, and the distance from one of those points to the local intermediate system
- \* This works in a five stage Clos spine and leaf, but not in a three stage, nor in some other five stage spine and leaf fabrics, such as the common butterfly or Benes fabric

## o Method 2:

- \* Manually mark one edge leaf node in the fabric as T0
- \* Calculate maximum distance through the fabric from this point
- \* Calculate local position based on this maximum distance the distance to the single marked device
- \* This works in three and five stage Clod fabrics, but does not work from every location in other spine and leaf fabrics, such as the common butterfly or Benes fabric

In the end, marking two devices located as far from one another topologically as possible provides the anchor points necessary to calculate the total distance through the fabric, and then from those points to the location of the calculating device.

The information obtained in this way can also be combined with other forms of location calculation, such as whether a device requesting an address through some mechanism is attached to the local device, or other indications of fabric locality. It generally true that having more than one method to determine fabric location will be better than any single method to account for errors, failures, and other problems that can arise with any mechanism.

Authors' Addresses

Russ White (editor)  
LinkedIn

Email: russ@riw.us

Shawn Zandi (editor)  
LinkedIn

Email: szandi@linkedin.com