Network Working Group R. White
Internet-Draft S. Zandi
Intended status: Informational LinkedIn
Expires: September 4, 2017 March 3, 2017

# OpenFabric
# draft-white-openfabric-00

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 Sytem to Intermediate System (IS-IS) that is designed to:

- Provide a full view of the topology from a single point in the network to simplify operations
- Minimize configuration of each router (or switch) in the network
- 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.

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 4, 2017.



Table of Contents



## 1. Introduction

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 predicitable forwarding and convergence properties. This document descibes 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:

- Provide a full view of the topology from a single point in the network to simplify operations

- Minimize configuration of each router (or switch) in the network

- Optimize the operation of IS-IS within a spine and leaf fabric to enable scaling

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:

- Multilevel flooding domain support. The modifications described in this document will not work across multiple flooding domains. It is assumed that multiple fabrics will be connected through an Exterior Gateway Protocol (EGP), specifically BGP [RFC4271].

- All mutliaccess link processing, including Designated Intermediate Systems (DIS). Spine and leaf fabrics are normally built using only point-to-point links, so multiaccess link processing is not required in OpenFabric.

- 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.

- 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 router 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].

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

- A slightly modified adjacency formation process. This is largely a matter of forming adjacencies in a specific order, rather than forming an adjacency with every discovered neighbor at the same time.

- A mechanism for determining which tier within a spine and leaf fabric in which the router is located.

- A mechanism that reduces flooding to the minimum possible, while still ensuring complete database synchronization among the routers within the fabric.

- New sub-TLVs to carry OpenFabric specific information; specifically a new IS reachability tier sub-TLV.

OpenFabric implementations:

- MUST support [RFC5301] and enable hostname advertisement by default if a hostname is configured on the intermediate system.

- MUST support [RFC5311], simplified extension of the link state PDU space for IS-IS.

- MUST support [RFC5303] and enable three-way handshakes by default.

- MUST use TLV type 135 for carrying IPv4 reachability information, as defined in [RFC5305].

- MUST use TLV type 236 for carrying IPv6 reachability information, as defined in [RFC5308].

- MUST use TLV type 22 for carrying IS reachability information, as defined in [RFC5305].

- SHOULD support [RFC6232], purge originator identification for IS-IS.

- SHOULD support Segment Routing (SR). [I-D.ietf-spring-segment-routing]

- SHOULD support [I-D.ietf-isis-segment-routing-extensions].

- 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.

OpenFabric implementations 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.

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

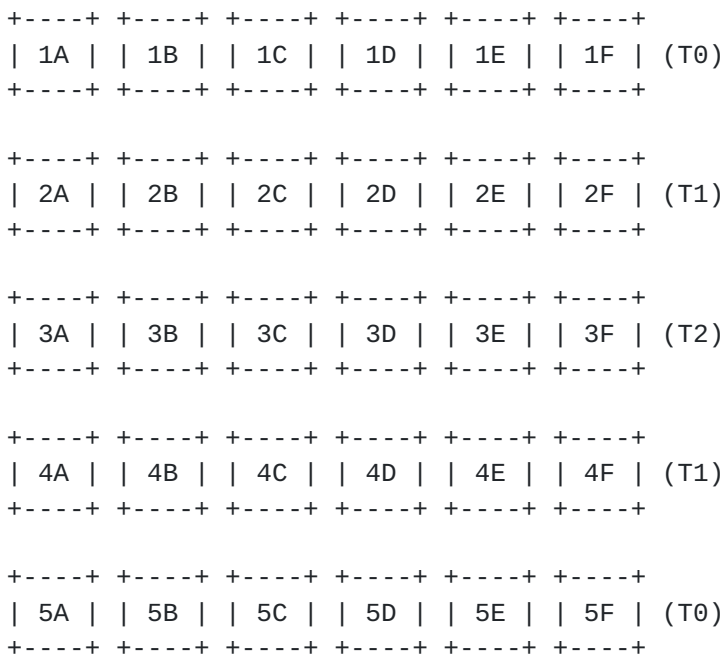


Figure 1

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 diven layer is connected to every device in the layer above it. For instance:

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

The authors would like to thank Nick Russo, Nikos Triantafillis, Rodny Molina, and Ivan Pepelnjak for their comments and review of the concepts and text of this document.

## 2. Modified Adjacency Formation

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

- An IS connected to the fabric will send hellos on all links.
- The IS will only complete the threeway 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.
- The IS will complete its databse exchange with this one newly adjacent neighbor.
- Once this process is completed, the IS will continue processing the remaining neighbors as normal.

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.

## 3. Determining Location on the Fabric

The tier to which a router is connected is useful to enable autoconfiguration of routers connected to the fabric, and to reduce flooding. This section describes mechanisms for determining the tier at which a router is connected in the fabric in several steps. The first step is to find the Farthest Distance (FD) and the Total Distance (TD), which are useful in this process. To find the FD and TD:

- Calculate a Shortest Path Tree (SPT) for the entire network with all link metrics set to 1; this has the effect of calculating a tree based only on hop count

- Find one node that is the farthest from the local node in the resulting tree; call this node F, and the distance to this node FD

- Calculate an SPT for the entire network with all link metrics set to 1 from the perspective of F; call this TD

### 3.1. Determining T0

If FD == TD == 2, this is a three stage fabric; it is not possible to determine the tier at which the local node is located based on any calculation, because the topology is perfectly symmetric. In this case:

- The T0 routers MAY be manually configured to advertise 0x00 in their IS reachability tier sub-TLV, indicating they are at the edge of the fabric (a ToR router).

- The T0 routers MAY detect that they are T0 through the the presence connected hosts (i.e. through a request for address assignment or some other means). This means of detection may not be reliable in all operational environments, and SHOULD be used with care. If such detection is used, and the router determines it is located at T0, it should advertise 0x00 in its IS reachability tier sub-TLV.

- The router MAY examine the IS reachability tier sub-TLV of directly connected neighbors and determine one or more is advertising 0x1 in its IS reachability tier sub-TLVs. This would be the case if the spine routers in a three stage spine and leaf fabric are manually configured to advertise their tier as 0x1.

- If there is no way to determine whether or not the local device is in T0 or T1, it MUST advertise 0xFF in its IS reachability tier sub-TLV.

If FD == TD, and TD >= 4, this is a greater than three stage fabric; the local device SHOULD advertise 0x00 in its IS reachability tier sub-TLV.

For instance, in the diagram above, 1A would:

- Calculate an SPT with all link metrics set to 1; on this SPT, 5A through 5F would all have a distance of 4

- Select one of these nodes as F; assume 5F is chosen as F

- Set FD to 4, the distance to 5F

- Run SPF from the perspective of 5F with all link metrics set to 1
- Set TD to 4, the cost from 5F to 1A
- TD - FD == 0, so 1A is at T0, and is a ToR

### 3.2. Determining T1 and above

If FD == TD == 2, this is a three stage fabric; it is not possible to determine the tier at which the local node is located based on any calculation, because the topology is perfectly symmetric. In this case:

- The T1 routers MAY be manually configured to advertise 0x01 in their IS reachability tier sub-TLV.
- The router MAY examine the IS reachability tier sub-TLV of directly connected neighbors and determine that one or more is advertising 0x00 in its IS reachability tier sub-TLVs. This would be the case if the ToR routers in a three stage spine and leaf fabric are manually configured to advertise their tier as 0x00.
- If there is no way to determine whether or not the local device is in T0 or T1, it should advertise 0xFF in its IS reachability tier sub-TLV.

If TD != FD, this is a greater than three stage fabric; the local device SHOULD advertise (TD - FD) in its IS reachability tier sub-TLV.

For example, in the above five stage fabric, 3B would:

- Calculate an SPT with all link metrics set to 1; on this SPT, 5A through 5F and 1A through 1F would all have a cost of 2
- Select one of these nodes as F; assume 5F is chosen as F
- Set FD to 2, the distance to 5F
- Run SPF from the perspective of 5F with all link metrics set to 1
- Set TD to 4, the cost from 5F to 1A
- TD - FD == 2, so 1A is at T2, and is a spine switch

## 4. Flooding Optimization

Flooding is perhaps the most challenging scaling issue for a link state protocol running on a dense, large scale fabric. To reduce flooding, 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:

- NL list: The set of neighbors
- NN list: The set of neighbor's neighbors; this can be calculated by running SPF truncated to two hops
- DNR list: The set of neighbors who should have LSPs (or fragments) marked Do Not Reflood (DNR)
- 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 router ID 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:

- Move any IS in NL with its tier (or fabric location) set to T0 to DNR
- If the LSP was received from an IS at a higher tier than the local IS, remove all intermediate systems from NL that are in the same tier as the IS the new LSP was received from

Then, for every IS in NL:

- 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
- Else move the IS to DNR

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, will have the DNR bit the optional flooding sub-TLV (see the packet format modifications and TLVs below).

Any IS receiving an LSP with the DNR bit set will not set the Send Route Message (SRM) flag on any interface for this LSP; hence the LSP will not be reflooded by this IS to any adjacent neighbor. This reduces flooding to the minimum possible while retaining full Link State Database (LSDB) synchronization.

## 5. OpenFabric and Route Aggregation

In data center fabrics, ToR routers SHOULD NOT be used to transit between two T1 (or above) spine routers. The simplest way to prevent this is to set the overload bit [RFC3277] for all the LSPs originated from T0 routers. However, this solution would have the unfortunate side effect of causing all reachability beyond any T0 router 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 routers.

Specifically, all T0 routers SHOULD advertise their metric to reach any T1 adjacent neighbor with a cost of 0XFFE. T1 routers, on the other hand, will advertise T0 routers with the actual interface cost used to reach the T0 router. Hence, links connecting T0 and T1 routers will be advertised with an assymetric cost that discourages transiting T0 routers, while leaving reachability to the destinations attached to T0 devices the same.

## 6. OpenFabric and Route Aggregation

While aggregation is not recommended in OpenFabric deployments, aggregation MAY take place when routing information is being transmitted from higher level tiers to lower level tiers. For instance, in the example network, 2A through 2F could advertise a single default route to 1A through 1F. 2A through 2F would simply advertise the default as if it were an attached to each router locally using either a type 135 or 236 TLV, and then block TLVs that contain reachability information (such as types 135 and 236). Type 22 TLVs, however, MUST be flooded through this boundary, so that every router in the network shares a common view of the topology.

Note that aggregation in a DC fabric can result in routing black holes in some cases, and also possibly reduce the efficiency of traffic engineering in the network.

## 7. OpenFabric Modifications to the IS-IS protocol

### 7.1. The Tier Level sub-TLV

A new sub-TLV is added to the type 22 TLV to indicate tier level, as follows:

- sub-TLV number (one octet): TBA
- Tier identifier (one octet)

The tier identifier field contains the tier number of the local router as calculated using the process above. If the tier number is unknown, the sub-TLV MUST be included with a tier ID of 0xFF, which indicates the advertising router does not have enough information to calculate its tier number, or there is some error in calculating a tier number.

### 7.2. The Do Not Reflood (DNR) bit

For OpenFabric implementations, the Partition Repair in the LSP PDU header SHALL be treated as the Do Not Reflood (DNR) bit. Any IS receiving an LSP with the DNR bit set SHOULD NOT set the SRM flag for the LSP, so the LSP will not be flooded to adjacent routers.

## 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 implementions 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.

## 9. References

### 9.1. Normative References

[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>.

[RFC2629] Rose, M., "Writing I-Ds and RFCs using XML", RFC 2629, DOI 10.17487/RFC2629, June 1999, <http://www.rfc-editor.org/info/rfc2629>.

[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>.

[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, <http://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, <http://www.rfc-editor.org/info/rfc5305>.

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

[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, <http://www.rfc-editor.org/info/rfc5311>.

## 9.2. Informative References

[I-D.ietf-isis-segment-routing-extensions] Previdi, S., Filsfils, C., Bashandy, A., Gredler, H., Litkowski, S., Decraene, B., and j. jefftant@gmail.com, "IS-IS Extensions for Segment Routing", draft-ietf-isis-segment-routing-extensions-10 (work in progress), February 2017.

[I-D.ietf-spring-segment-routing] Filsfils, C., Previdi, S., Decraene, B., Litkowski, S., and R. Shakir, "Segment Routing Architecture", draft-ietf-spring-segment-routing-11 (work in progress), February 2017.

[RFC3277] McPherson, D., "Intermediate System to Intermediate System (IS-IS) Transient Blackhole Avoidance", RFC 3277, DOI 10.17487/RFC3277, April 2002, <http://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, <http://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, <http://www.rfc-editor.org/info/rfc4271>.

[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>.

[RFC5440] Vasseur, JP., Ed. and JL. Le Roux, Ed., "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, DOI 10.17487/RFC5440, March 2009, <http://www.rfc-editor.org/info/rfc5440>.

[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, <http://www.rfc-editor.org/info/rfc6232>.

[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, <http://www.rfc-editor.org/info/rfc7921>.

## Authors' Addresses

Russ White
LinkedIn
Oak Island, NC 28465
USA

Email: russ@riw.us

Shawn Zandi
LinkedIn
San Francisco, CA XXXXX
USA

Email: szandi@linkedin.com