

Network Working Group  
Internet-Draft  
Intended status: Experimental  
Expires: 8 October 2026

A. Przygienda, Ed.  
HPE Juniper Networking  
6 April 2026

RIFT in Dragonfly Topologies  
draft-przygienda-rift-dragonfly-02

Abstract

RIFT extensions for dragonfly topologies.

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

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 8 October 2026.

Copyright Notice

Copyright (c) 2026 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 Revised BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Revised BSD License.

## Table of Contents

1. Introduction . . . . .	2
2. Glossary . . . . .	4
3. Horizontal Link Behavior at ToF Level . . . . .	5
4. First Route Computation Change . . . . .	6
4.1. Additional Bi-Sectional Bandwidth Route Computation Change . . . . .	6
4.2. Dragonfly with Multi-Plane CLOS Fabrics . . . . .	7
5. Forwarding Considerations . . . . .	9
6. Partitioning of inter Fabric Planes . . . . .	9
7. Specification . . . . .	10
8. Summary Overview . . . . .	10
9. IANA Considerations . . . . .	10
10. Security Considerations . . . . .	10
11. Acknowledgements . . . . .	11
12. References . . . . .	11
12.1. Informative References . . . . .	11
12.2. Normative References . . . . .	11
Author's Address . . . . .	11

## 1. Introduction

RIFT today is standardized to deal with CLOS variant fabrics with some horizontal link exceptions. Given that interconnecting multiple CLOS via a dragonfly and its variants is an interesting topology (whether it's a full mesh or some kind of non-completely meshed regular lattice) this document addresses the resulting changes necessary to base RIFT specification to support dragonfly interconnected CLOS fabrics. The reader is advised that due to complexity of figures involved the ASCII version of the document may present those in simplified fashion.

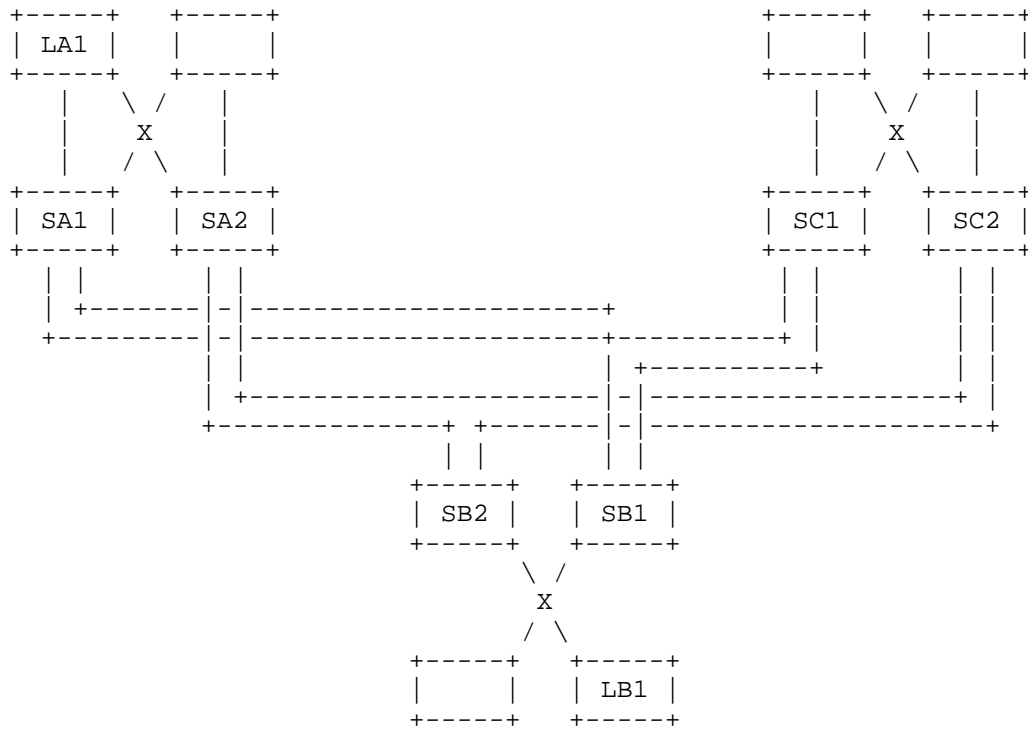


Figure 1: Sparse Dragonfly of CLOS Fabrics

To start with, Figure 1 visualizes three simple single plane fabrics interconnected via a DragonFly+ backbone. The behavior of standard RIFT is better understood if we look at the homomorphic version of the same topology in Figure 2. We can see that it is nothing else but a multi-plane CLOS with a lot of broken links for standard RIFT. The planes consist of  $S_{x_1}$  and  $S_{x_2}$  ToFs in each CLOS. Given this, leaf LB1 should be connected to SA1 to be in the plane and since it is not, SA1 will deduct that leaf LB1 fell off the plane 1 and negatively disaggregate it. Unfortunately the same is true for leaf LB1 from the view the SA2 in 2nd plane and it will negatively disaggregate it as well. Hence, leaf LA1 will not have any possibility to forward to LB1 using standard RIFT computed forwarding. This points us already to the first modification needed; we have to relax RIFT to forward through the horizontal links on ToFs and this will be the starting point of the next section.

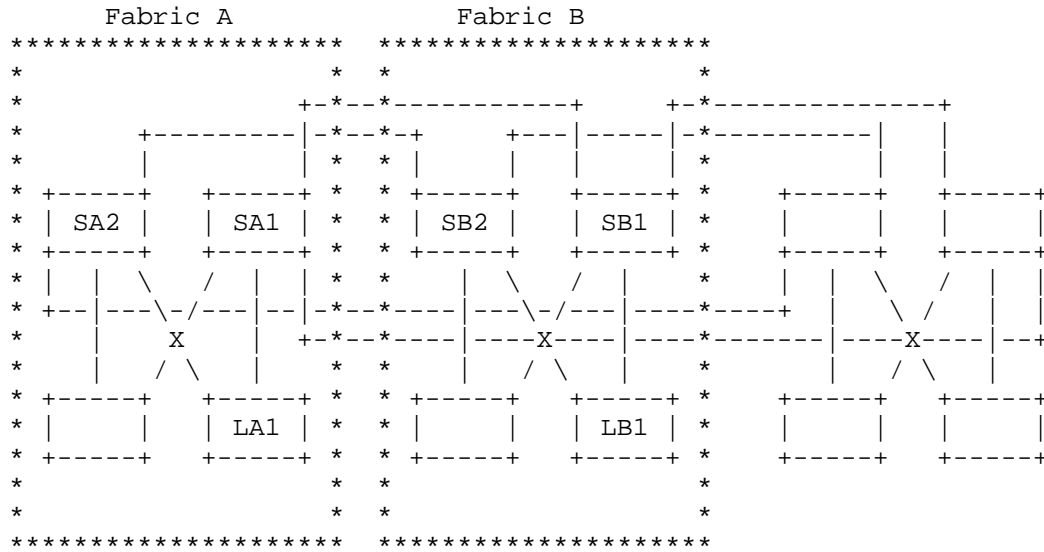


Figure 2: Homomorphic View of Sparse Dragonfly as a Multi-Plane CLOS

## 2. Glossary

The following terms are used in this document.

DF+ capable ToF:

ToF that provides DF+ extensions, both in recognizing the inter fabric links and computation procedures necessary to support those. The resulting combination allows the use of RIFT with dragonfly topologies overall.

Horizon:

We define horizon as a concept differentiating between inter fabric links and southbound pointing standard RIFT intra fabric links on a ToF. Both type of links need a different FIB to support alternate next hop when routing between fabrics. A link can only be on one side of the horizon but not both sides.

Inter Fabric Planes or IFL-planes:

Multi-Plane that spans multiple fabrics.

Inter Fabric link or IFL links:

A horizontal ToF link between two fabrics.

#### K-Alternate Next Hops or KNHs:

Next hop used by the originating fabric on a k-shortest path [EPPSTEIN] to the destination fabric. The receiver on the other side must subsequently use as next hop shortest path next hop to the destination fabric.

### 3. Horizontal Link Behavior at ToF Level

Dragonfly+, being basically, when seen a single fabric, a multi-plane CLOS with many broken links (which we will call inter fabric planes or IFL planes to distinguish them from multi-plane within a fabric later) will somehow need to change the behavior of RIFT to allow forwarding via horizontal links at ToF level lest we end up inverting the fabric and force leaves to deal with transit traffic. Moreover, the necessity to deal with new mis-cabling concepts leads us to change the solution framework and consider this configuration not as a single fabric but as a multi-fabric setup with dragonfly links building inter fabric planes now. Additionally we will have to allow adjacencies on ToF horizontal links to another fabric and permit those to forward through such inter fabric planes while distinguishing such inter fabric (or IF) links from normal horizontal ToF "multi-plane ringing". Hence in Figure 2 instead of the first assumption of a single fabric we break out fabric A and fabric B and consider the links SA2-SB2 and SA1-SB1 as two "inter fabric DF+" links, or in short, as already introduced, IFL links. And fortunately enough, IFL links, just like all other horizontal ToF links, are considered northbound from both sides and northbound flooding rules apply, an ideal thing since with that ToFs will see full topology of their inter fabric plane.

RIFT used in such DF+ configuration will require on ToF not only a DF+ capability flag but a fabric ID now which has to be distinct in each of the CLOS or dragonfly cliques. In case of non-DF+ mode a ToF will declare such links miscabled, once enabled to operate in DF+ it will mark those links as IFL links. Given 'fabric\_id' is an optional schema element a ToF operating in DF+ mode will reject all links to other ToFs without 'fabric\_id' value set or not indicating DF+ mode as mis-cabled to prevent a mixture of non-DF+ and DF+ ToFs in a setup. On the other hand, a ToF indicating DF+ capability and showing matching fabric id is clearly a normal horizontal multi plane ring in the same fabric.

#### 4. First Route Computation Change

Now that we can detect IFL links reliably we can also remove those from the computations used in negative disaggregation as first step. This will prevent ToFs in fabric A negatively disaggregating Fabric B prefixes, a desirable behavior. Not being able to forward from Fabric A to fabric B is obviously a far less desirable behavior and hence a ToF in DF+ mode needs to extend its route computation by a special southbound DF+ computation where we use SPF taking in first step all IFL links and the nodes behind them as candidates. This computation will result in a "direct inter fabric forwarding database" containing amongst others shortest path to prefixes in fabric B or in other words, direct inter fabric next hops.

##### 4.1. Additional Bi-Sectional Bandwidth Route Computation Change

One of the DF+ properties is that it not only provides a direct path to a destination but guarantees that destinations are reachable via KNHs to increase the bi-sectional bandwidth. In our first simple example SB1 forwarding to LA1 can take instead of SA1 directly a path through SC1 relying on it forwarding to SA1. And in less dense DF+ backbones we can even encounter longer indirect paths. To support this we introduce an additional SPF computation which computes for each IFL interface k-shortest-paths whereas K should be obviously reasonably constrained. To illustrate this by an example Figure 3 introduces a 17 fabrics rather sparse dragonfly mesh where a 3 hops alternate paths are viable (if so chosen). FID 0 forwarding to FID 2 has obviously the shortest path via FID1 but a k-shortest path computation will yield FID 13 and FID 16 as viable 3 hops k-shortest-paths. However, only the originating fabric can choose a k-shortest path and every subsequent non originator MUST follow equal cost shortest paths to prevent looping or excessive bow-tying through the fabric. Not only MUST it follow the shortest path to the destination fabric, the shortest path MUST be actually computed on a topology where the sender node is excluded to prevent looping.

For the sake of clarity Figure 3 does not visualize the full path diversity since, e.g. FID 13 can viably choose FIDs 3, 6, 9 beside FID 12 as shortest path next hops providing very significant path diversity on a 3 hops path. Further "steering" of chosen nexthops can be achieved by e.g. DSCP marking or congestion notification schemes but such approaches are outside the scope of this document.

Computing such alternate next hops will have the other beneficial effect of actually providing a backup path in case the direct IFL plane link to another fabric becomes unavailable.

missing TDB

Figure 3: Multi-Plane CLOS Fabrics Connected in Sparse Dragonfly

#### 4.2. Dragonfly with Multi-Plane CLOS Fabrics

Most complex case of RIFT deployment would be a dragonfly topology of CLOS fabrics which are in themselves already multi-plane fabrics. To present it as homomorphic graph Figure 4 is included. The symmetry is obvious, we end up with the normal RIFT ringing within the fabric, e.g. `r_A` for fabric A and then for the inter fabric planes dragonfly is basically the according ringing itself, here `IR_1` and `IR_2`. Observe that the northbound flooding occurring on all those links will present each ToF with the full topology of the dragonfly, a necessary condition for proper disaggregation and further reachability computations. If the intra fabric ToF ringing should be avoided a tunnel between the ToFs within a fabric are necessary and may go all the way down to the leaves. How such tunnels are provisioned is outside the specification here but it will necessitate basically flat distribution of the loopbacks of the ToFs across whole fabric via e.g. redistribution of some RIFT routes in northbound and southbound direction or an equivalent scheme.

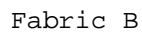




Figure 4: Multi-Plane CLOS Fabrics Connected in Sparse Dragonfly

## 5. Forwarding Considerations

Since RIFT is being extended with the concept of KNH and IP packets do not carry any marking as the path they have taken indiscriminate forwarding using non-shortest paths at ToF level may loop in inter fabric case. To prevent this the ToFs have to maintain the concept of a "split horizon" on the arriving traffic. Any traffic arriving at the ToF that is targeted at the prefix within its fabric can be forwarded without any further considerations. On the other hand, traffic not targeted at the fabric of the ToF arriving at a inter fabric link MUST use a FIB generated by Section 4.1. The solution will naturally limit any non-shortest inter fabric path in ToF case to shortest paths computed without including the sender. As last case, traffic arriving from within its own fabric but targeted at another fabric can use any of the k-shortest paths as next hop as described in Section 4.1. Observe that per interface specific FIB is nothing particularly special, any technology supporting VPN or trunking today is already capable of provisioning interface specific forwarding behavior.

## 6. Partitioning of inter Fabric Planes

A special case where a plane within a remote fabric breaks down is not noticeable in another fabric and hence the traffic can black hole since we do suppress the IFL links during negative disaggregation normally. To detect the condition reliably a ToF has to compute the inter fabric view of all the other ToFs in its own fabric while including IFL links and consider the resulting difference as "inter fabric negative disaggregation". This is possible but at scale can present significant computational load and is left therefore as optional behavior. Additionally, even when the fabric is a single plane fabric it must be then ringed at ToF level since otherwise the ToFs do not see the inter fabric planes they are not part of as an IFL ring.

The same computation will deal with an even stranger case of a double failure on the IF links where a ToF becomes completely separated from the other fabrics. It will detect this and initiate negative disaggregation for the according prefixes.

## 7. Specification

Precise schema changes and computation algorithms are to be provided in future version of the draft in detail. Basically the LIEs and Node TIEs need to be extended by fabric\_id and DF+ mode indication and computations described conceptually in former chapters tightly specified.

## 8. Summary Overview

A final Figure 5 is provided to map things back to the usual dragonfly sparse topology and show the concepts in action.

We see three fabrics, each of them multi-plane (though mixes are absolutely possible as long the number of ToFs connected to dragonfly are kept the same). The fat links represent the "IFL horizon", i.e. any traffic coming from those links cannot use alternate next hops to the destination. In this example traffic from LA11 going through PA11 and SA2 towards LC11 is given two choices of next hops, either SC2 or SB2. Now that it entered the IFL horizon in case SB2 receives it no further alternate next hops will be used but traffic will be handed off to SC2 which applies the same rule and in this case actually forwards the traffic into the fabric.

[[ Overview of the solution,  
Refer to PDF for Picture ]]

Figure 5: Multi-Plane CLOS Fabrics Connected in Sparse Dragonfly

For the more complex case of alternate paths longer than 2 hops and the resulting forwarding behavior and path diversity Figure 3 has been described in the document.

## 9. IANA Considerations

This document requests allocation for the following RIFT codepoints.

TBD

## 10. Security Considerations

TBD

## 11. Acknowledgements

Dmitry Afanasiev's ideas around his work with BGP and dragonfly started interesting discussions, and he provided the crucial split horizon forwarding idea. Jeff Tantsura encouraged the work from its initial conception. Many thanks to Benson Muite for ASCII figures.

## 12. References

### 12.1. Informative References

[EPPSTEIN] Eppstein, D., "Finding the k-Shortest Paths", 1997.

### 12.2. 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, <<https://www.rfc-editor.org/info/rfc2119>>.

[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>>.

## Author's Address

Tony Przygienda (editor)  
HPE Juniper Networking  
1137 Innovation Way  
Sunnyvale, CA  
United States of America  
Email: [prz@juniper.net](mailto:prz@juniper.net)