Workgroup:	Network Working Group
Internet-Draft:	draft-przygienda-rift-dragonfly-00
Published:	20 October 2023
Intended Status:	Experimental
Expires:	22 April 2024
Author:	A. Przygienda, Ed. *Juniper*

RIFT in Dragonfly++ Topologies

Abstract

RIFT support for dragonfly topologies as ToF interconnect.

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 22 April 2024.

Copyright Notice

Table of Contents

1. Introduction 2
2. Glossary 7
3. Horizontal Link Behavior at ToF Level 7
4. First, Simplest Route Computation Change 7
4.1. Additional Bi-Sectional Bandwidth Route Computation Change 7
5. Schema Modifications 8
6. Special Considerations 8
6.1. Partitioning of Inter-Fabric Planes 8
7. IANA Considerations 8
8. Security Considerations 8
9. Acknowledgements 8
10. References 8
10.1. Informative References 8
10.2. Normative References 9
Author's Address 9

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 variant 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 be advised that due to complexity of figures involved the ASCII version of the document leaves those out. To start, 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. Hence, leaf LA1 will not have any

possibility to forward to LB1 using standard RIFT. 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.

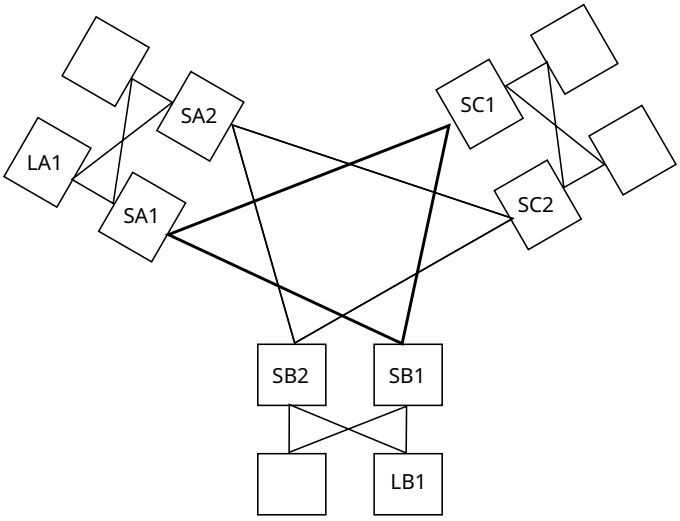
SA2
SC1
LA1
SA1
SC2
SB2
SB1
LB1

Figure 1: Topologically Connected Planes

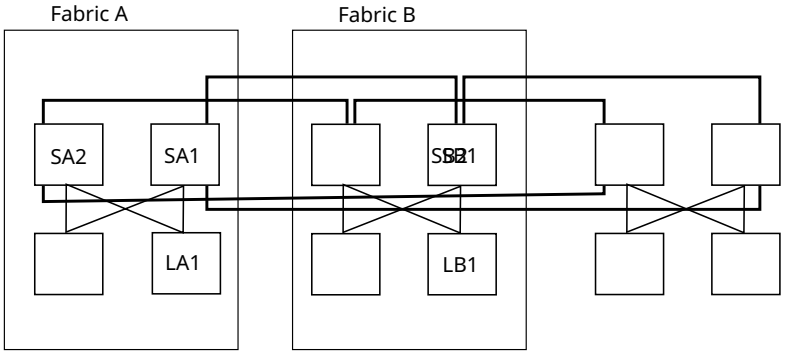
Fabric A
Fabric B
SA2
SA1
LA1
LB1

Figure 2: Topologically Connected Planes

2. Glossary

The following terms are used in this document.

Horizon:

3. Horizontal Link Behavior at ToF Level

Dragonfly+, being basically a multi-plane CLOS with many broken links, which we will call inter-fabric planes or IF planes to distinguish them from multi-plane within a fabric will need to change the behavior of RIFT to basically interconnect multiple distinct CLOS fabrics via horizontal links at ToF level that allow forwarding. Moreover, to deal with new mis-cabling concepts we will not consider this configuration a single fabric but a multi-fabric setup with "inter-fabric planes" and allow ToF horizontal links only to another fabric and moreover introduce forwarding through such links to distinguish those from normal "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 two "inter-fabric DF+" links, in short IF-links. Those links, just like all other horizontal ToF links are considered southbound from both sides and northbound flooding rules apply, an ideal thing since with that all ToFs will see full topology of their inter-fabric plane.

RIFT used in DF+ configuration will require on ToF not only a ToF flag but a fabric ID now which has to be distinct in each of the CLOS. In case of non-DF+ mode the ToF will declare such links miscabled, once enabled to be operate in DF+ it will mark those links as IF-links. Given `fabric_id` is an optional schema element a ToF operating in DF+ mode will reject all links which have a ToF on the other side but no `fabric_id` value set or do not indicate DF+ mode as mis-cabled to prevent a mixture of non-DF+ and DF+ ToFs in a setup.

4. First, Simplest Route Computation Change

Now that we can detect way IF-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 IF links and the nodes behind them as candidates. This computation will result in a "direct inter-fabric forwarding database" with basically shortest path to prefixes in fabric B.

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 additional, alternate next-hop to increase the bi-sectional bandwidth. In our example SB1 forwarding to LA1 can take instead of SA1 directly a path through SC1 relying on it forwarding to SA1. To support this we introduce an

additional SPF computation which takes in first 2 hops only the IF links is needed and generates a "indirect inter-fabric forwarding database". We will see later how those FIBs are used.

5. Schema Modifications

To be provided in future version of the draft.

6. Special Considerations

6.1. Partitioning of Inter-Fabric Planes

A special case where a plane within a fabric breaks down is not noticeable in another fabric and hence the traffic can black hole. To detect the condition reliably a ToF has to compute the inter-fabric view of all the other ToFs in its own fabric and take the resulting difference as "inter-fabric negative disaggregation".

7. IANA Considerations

This document requests allocation for the following RIFT codepoints.

TBD

8. Security Considerations

TBD

9. Acknowledgements

Dmitry Afanasiev based on his work with BGP and DF+ provided the crucial split horizon forwarding idea.

10. References

10.1. Informative References

[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>.

[RFC4456] Bates, T., Chen, E., and R. Chandra, "BGP Route Reflection: An Alternative to Full Mesh Internal BGP (IBGP)", RFC 4456, DOI 10.17487/RFC4456, April 2006, <https://www.rfc-editor.org/info/rfc4456>.

[RFC8099] Chen, H., Li, R., Retana, A., Yang, Y., and Z. Liu, "OSPF Topology-Transparent Zone", RFC 8099, DOI 10.17487/RFC8099, February 2017, <https://www.rfc-editor.org/info/rfc8099>.

10.2. 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>.

[RFC5302] Li, T., Smit, H., and T. Przygienda, "Domain-Wide Prefix Distribution with Two-Level IS-IS", RFC 5302, DOI 10.17487/RFC5302, October 2008, <https://www.rfc-editor.org/info/rfc5302>.

[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>.

[RFC7775] Ginsberg, L., Litkowski, S., and S. Previdi, "IS-IS Route Preference for Extended IP and IPv6 Reachability", RFC 7775, DOI 10.17487/RFC7775, February 2016, <https://www.rfc-editor.org/info/rfc7775>.

[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>.

[RFC9012] Patel, K., Van de Velde, G., Sangli, S., and J. Scudder, "The BGP Tunnel Encapsulation Attribute", RFC 9012, DOI 10.17487/RFC9012, April 2021, <https://www.rfc-editor.org/info/rfc9012>.

Author's Address

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