Workgroup: RIFT WG
Published: 13 October 2020
Intended Status: Informational
Expires: 16 April 2021
Authors: Yuehua. Wei, Ed. Zheng. Zhang
ZTE Corporation ZTE Corporation
Dmitry. Afanasiev Tom. Verhaeg
Yandex Juniper Networks
Jaroslaw. Kowalczyk P. Thubert
Orange Polska Cisco Systems

RIFT Applicability

Abstract

This document discusses the properties, applicability and operational considerations of RIFT in different network scenarios. It intends to provide a rough guide how RIFT can be deployed to simplify routing operations in Clos topologies and their variations.

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 16 April 2021.

Copyright Notice

Table of Contents

1. Introduction

This document intends to explain the properties and applicability of "Routing in Fat Trees" [RIFT] in different deployment scenarios and

highlight the operational simplicity of the technology compared to traditional routing solutions. It also documents special considerations when RIFT is used with or without overlays, controllers and corrects topology miscablings and/or node and link failures.

2. Problem Statement of Routing in Modern IP Fabric Fat Tree Networks

Clos and Fat-Tree topologies have gained prominence in today's networking, primarily as result of the paradigm shift towards a centralized data-center based architecture that is poised to deliver a majority of computation and storage services in the future.

Today's current routing protocols were geared towards a network with an irregular topology and low degree of connectivity originally. When they are applied to Fat-Tree topologies:

* they tend to need extensive configuration or provisioning during bring up and re-dimensioning.

* spine and leaf nodes have the entire network topology and routing information, which is in fact, not needed on the leaf nodes during normal operation.

* significant Link State PDUs (LSPs) flooding duplication between spine nodes and leaf nodes occurs during network bring up and topology updates. It consumes both spine and leaf nodes' CPU and link bandwidth resources and with that limits protocol scalability.

3. Applicability of RIFT to Clos IP Fabrics

Further content of this document assumes that the reader is familiar with the terms and concepts used in OSPF [RFC2328] and IS-IS [ISO10589-Second-Edition] link-state protocols and at least the sections of [RIFT] outlining the requirement of routing in IP fabrics and RIFT protocol concepts.

3.1. Overview of RIFT

RIFT is a dynamic routing protocol for Clos and fat-tree network topologies. It defines a link-state protocol when "pointing north" and path-vector protocol when "pointing south".

It floods flat link-state information northbound only so that each level obtains the full topology of levels south of it. That information is never flooded east-west or back South again. So a top tier node has full set of prefixes from the SPF calculation.

In the southbound direction the protocol operates like a "fully summarizing, unidirectional" path vector protocol or rather a distance vector with implicit split horizon whereas the information propagates one hop south and is 're-advertised' by nodes at next lower level, normally just the default route.

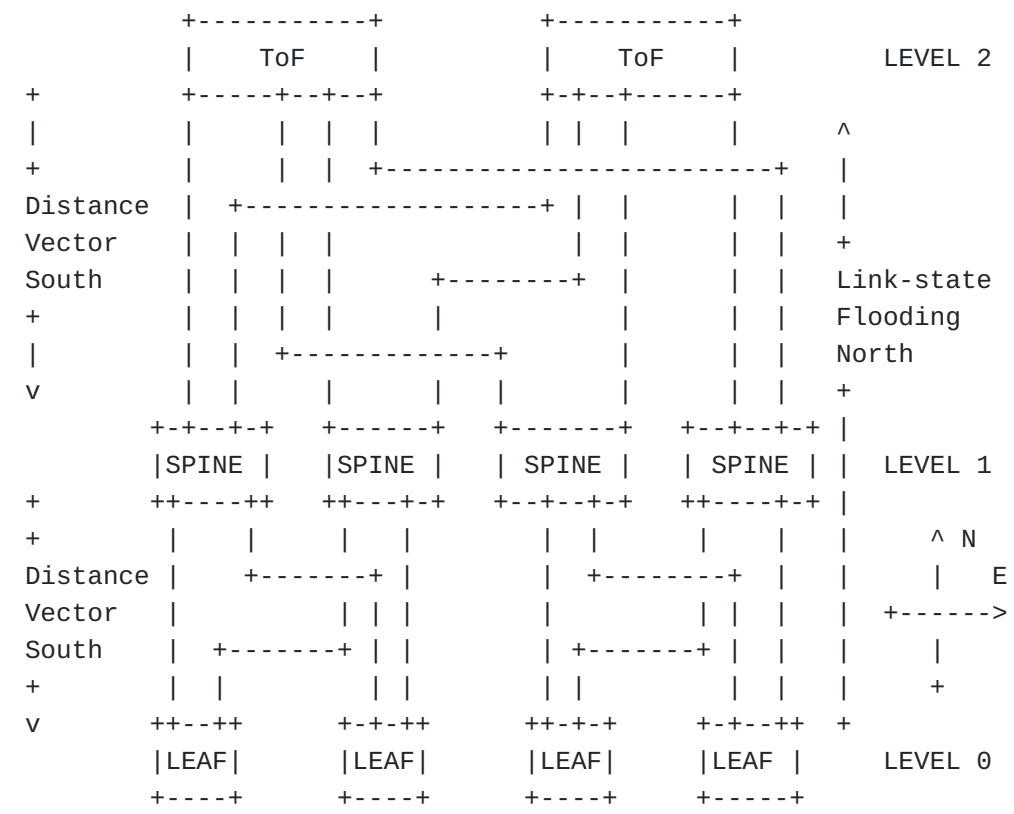

Figure 1: Rift overview

A middle tier node has only information necessary for its level, which are all destinations south of the node based on SPF calculation, default route and potential disaggregated routes.

RIFT combines the advantage of both link-state and distance vector:

- *Fastest Possible Convergence
- *Automatic Detection of Topology
- *Minimal Routes/Info on TORs
- *High Degree of ECMP
- *Fast De-commissioning of Nodes
- *Maximum Propagation Speed with Flexible Prefixes in an Update

And RIFT eliminates the disadvantages of link-state or distance vector:

*Reduced and Balanced Flooding

*Automatic Neighbor Detection

So there are two types of link-state database which are "north representation" N-TIEs and "south representation" S-TIEs. The N-TIEs contain a link-state topology description of lower levels and S-TIEs carry simply default routes for the lower levels.

There are a bunch of more advantages unique to RIFT listed below which could be understood if you read the details of [RIFT].

*True ZTP

*Minimal Blast Radius on Failures

*Can Utilize All Paths Through Fabric Without Looping

*Automatic Disaggregation on Failures

*Simple Leaf Implementation that Can Scale Down to Servers

*Key-Value Store

*Horizontal Links Used for Protection Only

*Supports Non-Equal Cost Multipath and Can Replace MC-LAG

*Optimal Flooding Reduction and Load-Balancing

3.2. Applicable Topologies

Albeit RIFT is specified primarily for "proper" Clos or "fat-tree" structures, it already supports PoD concepts which are strictly speaking not found in original Clos concepts.

Further, the specification explains and supports operations of multi-plane Clos variants where the protocol relies on set of rings to allow the reconciliation of topology view of different planes as most desirable solution making proper disaggregation viable in case of failures. These observations hold not only in case of RIFT but in the generic case of dynamic routing on Clos variants with multiple planes and failures in bi-sectional bandwidth, especially on the leafs.

3.2.1. Horizontal Links

RIFT is not limited to pure Clos divided into PoD and multi-planes but supports horizontal links below the top of fabric level. Those links are used however only as routes of last resort northbound when a spine loses all northbound links or cannot compute a default route through them.

A possible configuration is a "ring" of horizontal links at a level. In presence of such a "ring" in any level (except ToF level) neither N-SPF nor S-SPF will provide a "ring-based protection" scheme since such a computation would have to deal necessarily with breaking of "loops" in Dijkstra sense; an application for which RIFT is not intended.

A full-mesh connectivity between nodes on the same level can be employed and that allows N-SPF to provide for any node loosing all its northbound adjacencies (as long as any of the other nodes in the level are northbound connected) to still participate in northbound forwarding.

3.2.2. Vertical Shortcuts

Through relaxations of the specified adjacency forming rules RIFT implementations can be extended to support vertical "shortcuts" as proposed by e.g. [I-D.white-distoptflood]. The RIFT specification itself does not provide the exact details since the resulting solution suffers from either much larger blast radius with increased flooding volumes or in case of maximum aggregation routing bow-tie problems.

3.2.3. Generalizing to any Directed Acyclic Graph

RIFT is an anisotropic routing protocol, meaning that it has a sense of direction (northbound, southbound, east-west) and that it operates differently depending on the direction.

* Northbound, RIFT operates as a link-state IGP, whereby the control packets are reflooded first all the way North and only interpreted later. All the individual fine grained routes are advertised.

* Southbound, RIFT operates as a distance vector IGP, whereby the control packets are flooded only one hop, interpreted, and the consequence of that computation is what gets flooded on more hop South. In the most common use-cases, a ToF node can reach most of the prefixes in the fabric. If that is the case, the ToF node advertises the fabric default and disaggregates the prefixes that it cannot reach. On the other hand, a ToF Node that can reach only a small subset of the prefixes in the fabric will preferably advertise those prefixes and refrain from aggregating.

In the general case, what gets advertised South is in more details:

1. A fabric default that aggregates all the prefixes that are reachable within the fabric, and that could be a default route or a prefix that is dedicated to this particular fabric.

2. The loopback addresses of the northbound nodes, e.g., for inband management.

3. The disaggregated prefixes for the dynamic exceptions to the fabric Default, advertised to route around the black hole that may form

*east-west routing can optionally be used, with specific restrictions. It is useful in particular when a sibling has access to the fabric default but this node does not.

A Directed Acyclic Graph (DAG) provides a sense of North (the direction of the DAG) and of South (the reverse), which can be used to apply RIFT. For the purpose of RIFT, an edge in the DAG that has only incoming vertices is a ToF node.

There are a number of caveats though:

*The DAG structure must exist before RIFT starts, so there is a need for a companion protocol to establish the logical DAG structure.

*A generic DAG does not have a sense of east and west. The operation specified for east-west links and the southbound reflection between nodes are not applicable.

*In order to aggregate and disaggregate routes, RIFT requires that all the ToF nodes share the full knowledge of the prefixes in the fabric. This can be achieved with a ring as suggested by the RIFT main specification, by some preconfiguration, or using a synchronization with a common repository where all the active prefixes are registered.

3.3. Use Cases

3.3.1. DC Fabrics

RIFT is largely driven by demands and hence ideally suited for application in underlay of data center IP fabrics, vast majority of which seem to be currently (and for the foreseeable future) Clos architectures. It significantly simplifies operation and deployment

of such fabrics as described in Section 4 for environments compared to extensive proprietary provisioning and operational solutions.

3.3.2. Metro Fabrics

The demand for bandwidth is increasing steadily, driven primarily by environments close to content producers (server farms connection via DC fabrics) but in proximity to content consumers as well. Consumers are often clustered in metro areas with their own network architectures that can benefit from simplified, regular Clos structures and hence RIFT.

3.3.3. Building Cabling

Commercial edifices are often cabled in topologies that are either Clos or its isomorphic equivalents. With many floors the Clos can grow rather high and with that present a challenge for traditional routing protocols (except BGP and by now largely phased-out PNNI) which do not support an arbitrary number of levels which RIFT does naturally. Moreover, due to limited sizes of forwarding tables in active elements of building cabling the minimum FIB size RIFT maintains under normal conditions can prove particularly cost-effective in terms of hardware and operational costs.

3.3.4. Internal Router Switching Fabrics

It is common in high-speed communications switching and routing devices to use fabrics when a crossbar is not feasible due to cost, head-of-line blocking or size trade-offs. Normally such fabrics are not self-healing or rely on 1:/+1 protection schemes but it is conceivable to use RIFT to operate Clos fabrics that can deal effectively with interconnections or subsystem failures in such module. RIFT is neither IP specific and hence any link addressing connecting internal device subnets is conceivable.

3.3.5. CloudCO

The Cloud Central Office (CloudCO) is a new stage of telecom Central Office. It takes the advantage of Software Defined Networking (SDN) and Network Function Virtualization (NFV) in conjunction with general purpose hardware to optimize current networks. The following figure illustrates this architecture at a high level. It describes a single instance or macro-node of cloud CO. An Access I/O module faces a Cloud CO Access Node, and the CPEs behind it. A Network I/O module is facing the core network. The two I/O modules are interconnected by a leaf and spine fabric. [TR-384]

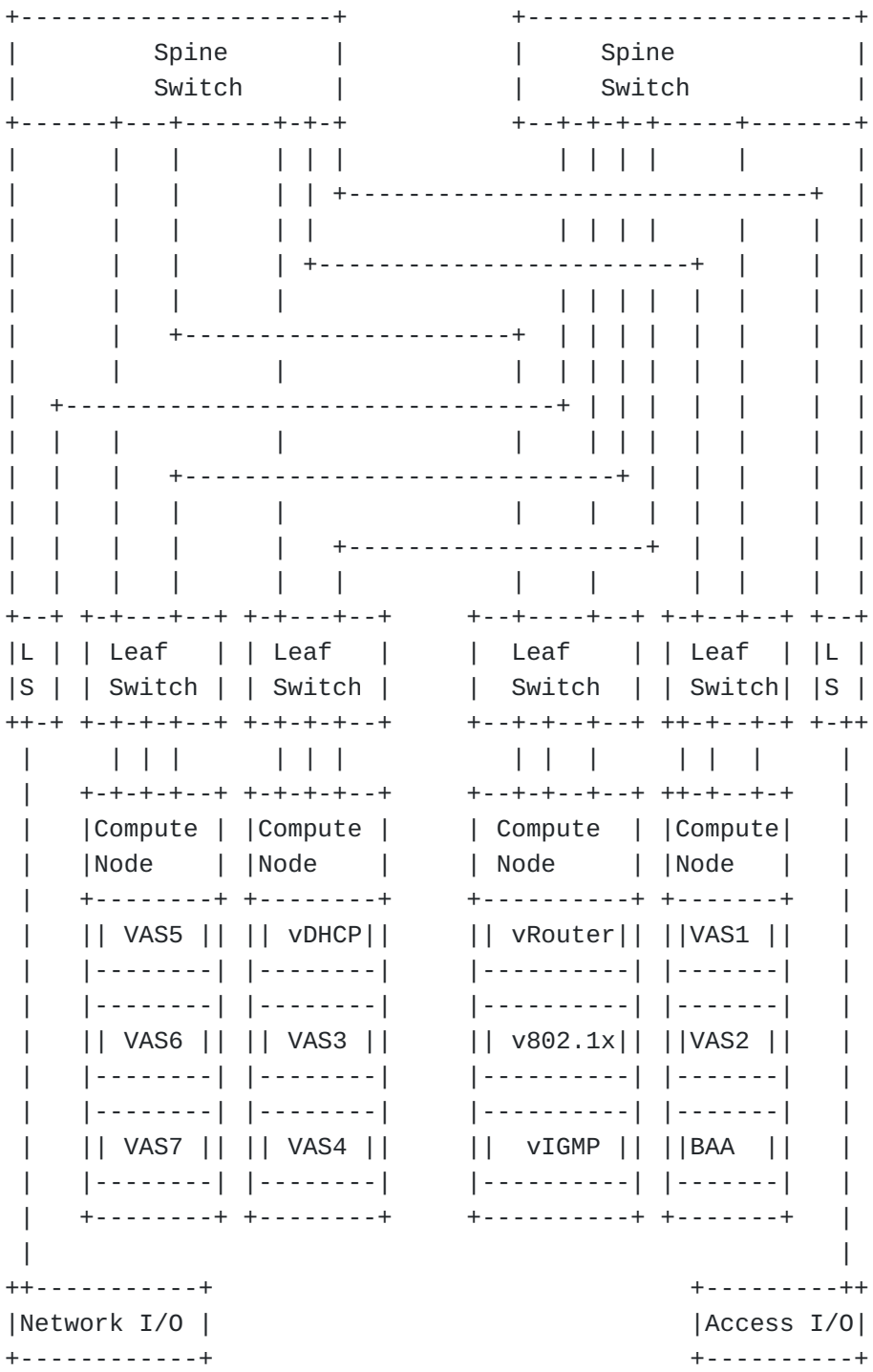

Figure 2: An example of CloudCO architecture

The Spine-Leaf architecture deployed inside CloudCO meets the network requirements of adaptable, agile, scalable and dynamic.

4. Deployment Considerations

RIFT presents the opportunity for organizations building and operating IP fabrics to simplify their operation and deployments while achieving many desirable properties of a dynamic routing on such a substrate:

* RIFT design follows minimum blast radius and minimum necessary epistemological scope philosophy which leads to very good scaling properties while delivering maximum reactiveness.

* RIFT allows for extensive Zero Touch Provisioning within the protocol. In its most extreme version RIFT does not rely on any specific addressing and for IP fabric can operate using IPv6 ND [RFC4861] only.

* RIFT has provisions to detect common IP fabric mis-cabling scenarios.

* RIFT negotiates automatically BFD per link allowing this way for IP and micro-BFD [RFC7130] to replace LAGs which do hide bandwidth imbalances in case of constituent failures. Further automatic link validation techniques similar to [RFC5357] could be supported as well.

* RIFT inherently solves many difficult problems associated with the use of traditional routing topologies with dense meshes and high degrees of ECMP by including automatic bandwidth balancing, flood reduction and automatic disaggregation on failures while providing maximum aggregation of prefixes in default scenarios.

* RIFT reduces FIB size towards the bottom of the IP fabric where most nodes reside and allows with that for cheaper hardware on the edges and introduction of modern IP fabric architectures that encompass e.g. server multi-homing.

* RIFT provides valley-free routing and with that is loop free. This allows the use of any such valley-free path in bi-sectional fabric bandwidth between two destination irrespective of their metrics which can be used to balance load on the fabric in different ways.

* RIFT includes a key-value distribution mechanism which allows for many future applications such as automatic provisioning of basic overlay services or automatic key roll-overs over whole fabrics.

* RIFT is designed for minimum delay in case of prefix mobility on the fabric.

* Many further operational and design points collected over many years of routing protocol deployments have been incorporated in

RIFT such as fast flooding rates, protection of information lifetimes and operationally easily recognizable remote ends of links and node names.

4.1. South Reflection

South reflection is a mechanism that South Node TIEs are "reflected" back up north to allow nodes in same level without E-W links to "see" each other.

For example, Spine111\Spine112\Spine121\Spine122 reflects Node S-TIEs from ToF21 to ToF22 separately. Respectively, Spine111\Spine112\Spine121\Spine122 reflects Node S-TIEs from ToF22 to ToF21 separately. So ToF22 and ToF21 see each other's node information as level 2 nodes.

In an equivalent fashion, as the result of the south reflection between Spine121-Leaf121-Spine122 and Spine121-Leaf122-Spine122, Spine121 and Spine 122 knows each other at level 1.

4.2. Suboptimal Routing on Link Failures

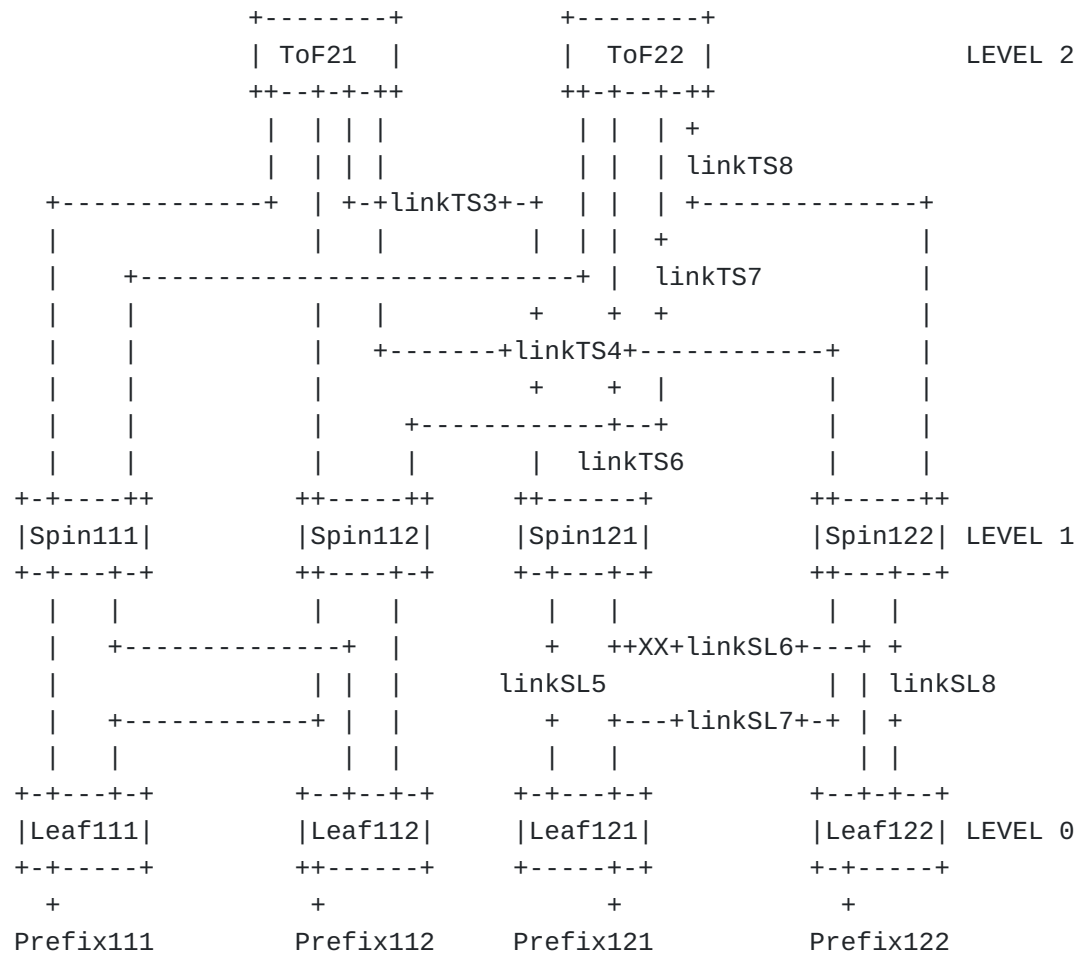

Figure 3: Suboptimal routing upon link failure use case

As shown in Figure 3, as the result of the south reflection between Spine121-Leaf121-Spine122 and Spine121-Leaf122-Spine122, Spine121 and Spine 122 knows each other at level 1.

Without disaggregation mechanism, when linkSL6 fails, the packet from leaf121 to prefix122 will probably go up through linkSL5 to linkTS3 then go down through linkTS4 to linkSL8 to Leaf122 or go up through linkSL5 to linkTS6 then go down through linkTS4 and linkSL8 to Leaf122 based on pure default route. It's the case of suboptimal routing or bow-tieing.

With disaggregation mechanism, when linkSL6 fails, Spine122 will detect the failure according to the reflected node S-TIE from Spine121. Based on the disaggregation algorithm provided by RIFT, Spine122 will explicitly advertise prefix122 in Disaggregated Prefix S-TIE PrefixesElement(prefix122, cost 1). The packet from leaf121 to prefix122 will only be sent to linkSL7 following a longest-prefix match to prefix 122 directly then go down through linkSL8 to Leaf122 .

4.3. Black-Holing on Link Failures

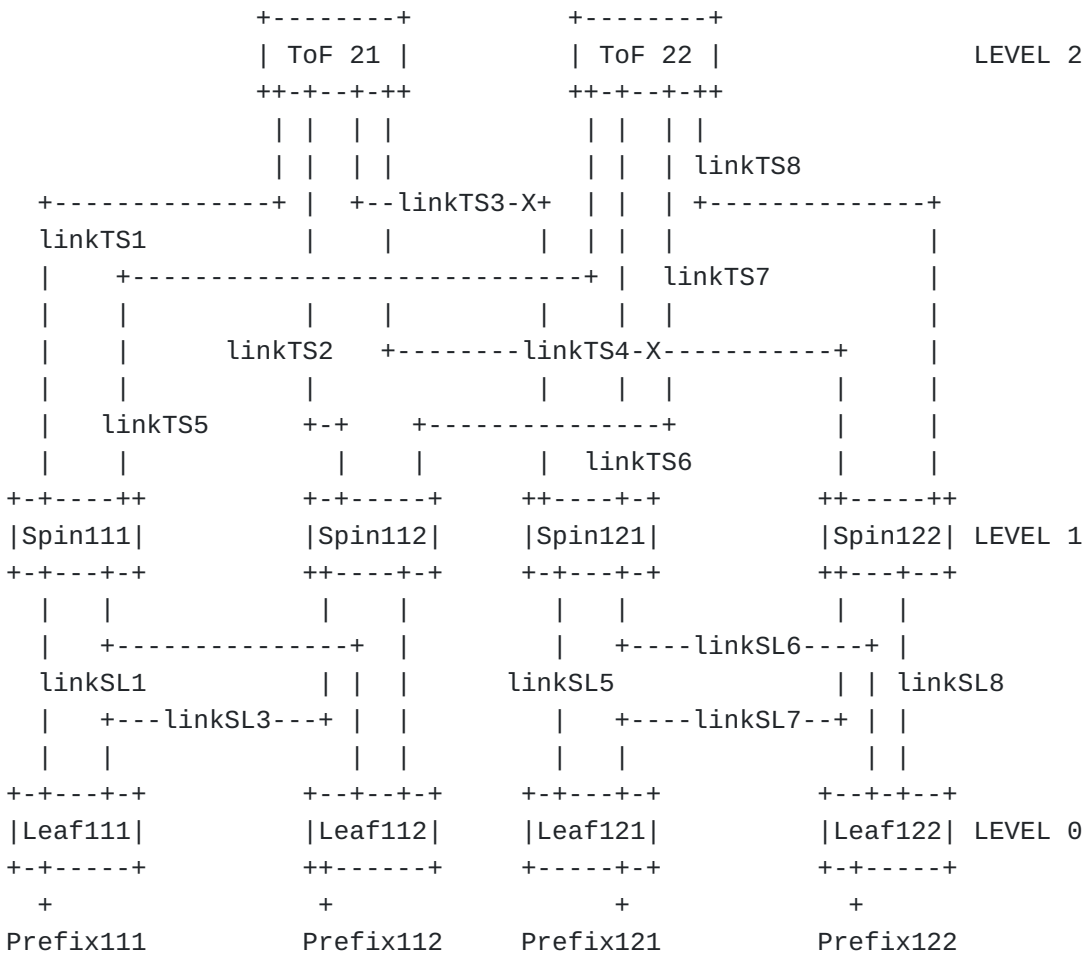

Figure 4: Black-holing upon link failure use case

This scenario illustrates a case when double link failure occurs and with that black-holing can happen.

Without disaggregation mechanism, when linkTS3 and linkTS4 both fail, the packet from leaf111 to prefix122 would suffer 50% black-holing based on pure default route. The packet supposed to go up through linkSL1 to linkTS1 then go down through linkTS3 or linkTS4 will be dropped. The packet supposed to go up through linkSL3 to linkTS2 then go down through linkTS3 or linkTS4 will be dropped as well. It's the case of black-holing.

With disaggregation mechanism, when linkTS3 and linkTS4 both fail, ToF22 will detect the failure according to the reflected node S-TIE of ToF21 from Spine111\Spine112. Based on the disaggregation algorithm provided by RITF, ToF22 will explicitly originate an S-TIE with prefix 121 and prefix 122, that is flooded to spines 111, 112, 121 and 122.

The packet from leaf111 to prefix122 will not be routed to linkTS1 or linkTS2. The packet from leaf111 to prefix122 will only be routed to linkTS5 or linkTS7 following a longest-prefix match to prefix122.

4.4. Zero Touch Provisioning (ZTP)

Each RIFT node may operate in zero touch provisioning (ZTP) mode. It has no configuration (unless it is a Top-of-Fabric at the top of the topology or it is desired to confine it to leaf role w/o leaf-2-leaf procedures). In such case RIFT will fully configure the node's level after it is attached to the topology.

The most import component for ZTP is the automatic level derivation procedure. All the Top-of-Fabric nodes are explicitly marked with TOP_OF_FABRIC flag which are initial 'seeds' needed for other ZTP nodes to derive their level in the topology. The derivation of the level of each node happens then based on LIEs received from its neighbors whereas each node (with possibly exceptions of configured leafs) tries to attach at the highest possible point in the fabric. This guarantees that even if the diffusion front reaches a node from "below" faster than from "above", it will greedily abandon already negotiated level derived from nodes topologically below it and properly peer with nodes above.

4.5. Miscabling Examples

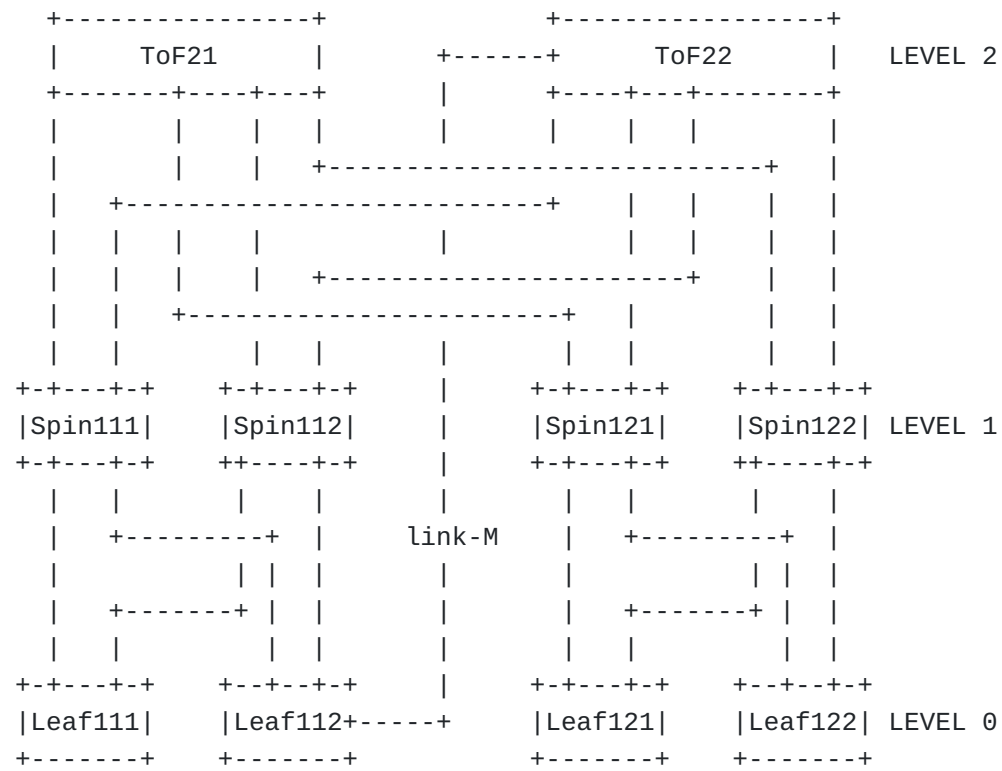

Figure 5: A single plane miscabling example

Figure 5 shows a single plane miscabling example. It's a perfect fat-tree fabric except link-M connecting Leaf112 to ToF22.

The RIFT control protocol can discover the physical links automatically and be able to detect cabling that violates fat-tree topology constraints. It react accordingly to such mis-cabling attempts, at a minimum preventing adjacencies between nodes from being formed and traffic from being forwarded on those mis-cabled links. Leaf112 will in such scenario use link-M to derive its level (unless it is leaf) and can report links to spines 111 and 112 as miscabled unless the implementations allows horizontal links.

Figure 6 shows a multiple plane miscabling example. Since Leaf112 and Spine121 belong to two different PoDs, the adjacency between Leaf112 and Spine121 can not be formed. link-W would be detected and prevented.

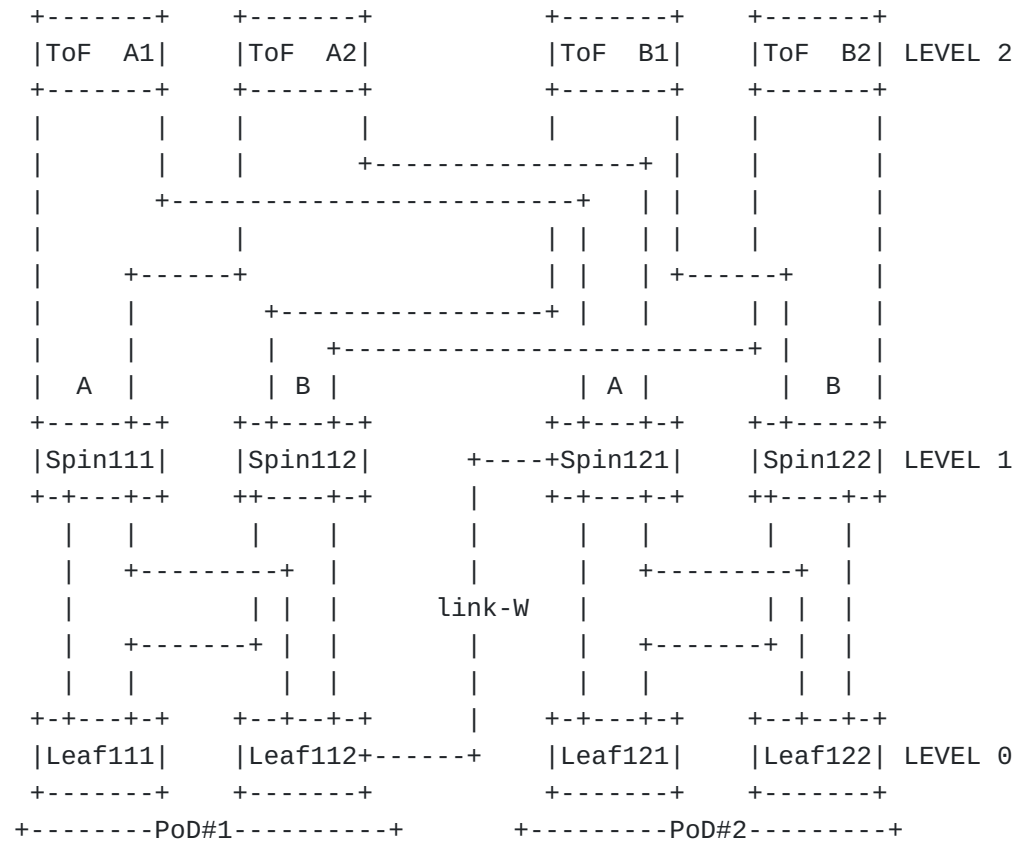

Figure 6: A multiple plane miscabling example

RIFT provides an optional level determination procedure in its Zero Touch Provisioning mode. Nodes in the fabric without their level configured determine it automatically. This can have possibly

counter-intuitive consequences however. One extreme failure scenario is depicted in Figure 7 and it shows that if all northbound links of spine11 fail at the same time, spine11 negotiates a lower level than Leaf11 and Leaf12.

To prevent such scenario where leafs are expected to act as switches, LEAF_ONLY flag can be set for Leaf111 and Leaf112. Since level -1 is invalid, Spine11 would not derive a valid level from the topology in Figure 7. It will be isolated from the whole fabric and it would be up to the leafs to declare the links towards such spine as miscabled.

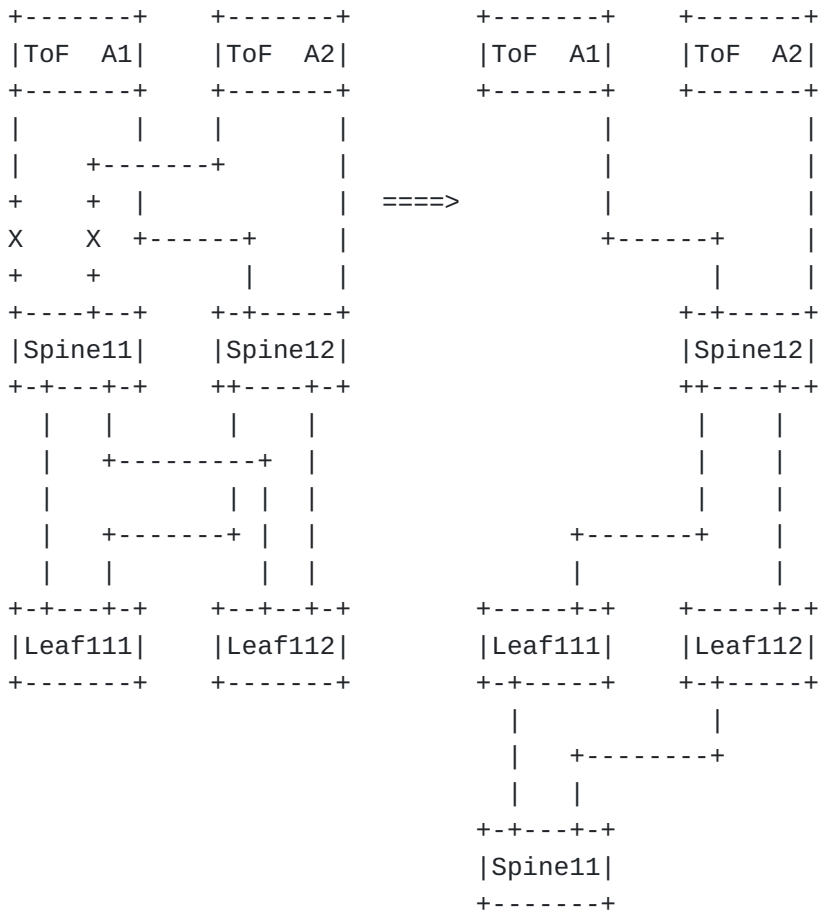

Figure 7: Fallen spine

4.6. Positive vs. Negative Disaggregation

Disaggregation is the procedure whereby [RIFT] advertises a more specific route Southwards as an exception to the aggregated fabric-default North. Disaggregation is useful when a prefix within the aggregation is reachable via some of the parents but not the others at the same level of the fabric. It is mandatory when the level is the ToF since a ToF node that cannot reach a prefix becomes a black

hole for that prefix. The hard problem is to know which prefixes are reachable by whom.

In the general case, [RIFT] solves that problem by interconnecting the ToF nodes so they can exchange the full list of prefixes that exist in the fabric and figure when a ToF node lacks reachability and to existing prefix. This requires additional ports at the ToF, typically 2 ports per ToF node to form a ToF-spanning ring. [RIFT] also defines the southbound reflection procedure that enables a parent to explore the direct connectivity of its peers, meaning their own parents and children; based on the advertisements received from the shared parents and children, it may enable the parent to infer the prefixes its peers can reach.

When a parent lacks reachability to a prefix, it may disaggregate the prefix negatively, i.e., advertise that this parent can be used to reach any prefix in the aggregation except that one. The Negative Disaggregation signaling is simple and functions transitively from ToF to ToP and then from Top to Leaf. But it is hard for a parent to figure which prefix it needs to disaggregate, because it does not know what it does not know; it results that the use of a spanning ring at the ToF is required to operate the Negative Disaggregation. Also, though it is only an implementation problem, the programmation of the FIB is complex compared to normal routes, and may incur recursions.

The more classical alternative is, for the parents that can reach a prefix that peers at the same level cannot, to advertise a more specific route to that prefix. This leverages the normal longest prefix match in the FIB, and does not require a special implementation. But as opposed to the Negative Disaggregation, the Positive Disaggregation is difficult and inefficient to operate transitively.

Transitivity is not needed to a grandchild if all its parents received the Positive Disaggregation, meaning that they shall all avoid the black hole; when that is the case, they collectively build a ceiling that protects the grandchild. But until then, a parent that received a Positive Disaggregation may believe that some peers are lacking the reachability and readvertise too early, or defer and maintain a black hole situation longer than necessary.

In a non-partitioned fabric, all the ToF nodes see one another through the reflection and can figure if one is missing a child. In that case it is possible to compute the prefixes that the peer cannot reach and disaggregate positively without a ToF-spanning ring. The ToF nodes can also ascertain that the ToP nodes are connected each to at least a ToF node that can still reach the prefix, meaning that the transitive operation is not required.

The bottom line is that in a fabric that is partitioned (e.g., using multiple planes) and/or where the ToP nodes are not guaranteed to always form a ceiling for their children, it is mandatory to use the Negative Disaggregation. On the other hand, in a highly symmetrical and fully connected fabric, (e.g., a canonical Clos Network), the Positive Disaggregation methods allows to save the complexity and cost associated to the ToF-spanning ring.

Note that in the case of Positive Disaggregation, the first ToF node(s) that announces a more-specific route attracts all the traffic for that route and may suffer from a transient incast. A ToP node that defers injecting the longer prefix in the FIB, in order to receive more advertisements and spread the packets better, also keeps on sending a portion of the traffic to the black hole in the meantime. In the case of Negative Disaggregation, the last ToF node(s) that injects the route may also incur an incast issue; this problem would occur if a prefix that becomes totally unreachable is disaggregated, but doing so is mostly useless and is not recommended.

4.7. Mobile Edge and Anycast

When a physical or a virtual node changes its point of attachement in the fabric from a previous-leaf to a next-leaf, new routes must be installed that supersede the old ones. Since the flooding flows Northwards, the nodes (if any) between the previous-leaf and the common parent are not immediately aware that the path via previous-leaf is obsolete, and a stale route may exist for a while. The common parent needs to select the freshest route advertisement in order to install the correct route via the next-leaf. This requires that the fabric determines the sequence of the movements of the mobile node.

On the one hand, a classical sequence counter provides a total order for a while but it will eventually wrap. On the other hand, a timestamp provides a permanent order but it may miss a movement that happens too quickly vs. the granularity of the timing information. It is not envisioned in the short term that the average fabric supports a Precision Time Protocol, and the precision that may be available with the Network Time Protocol [RFC5905], in the order of 100 to 200ms, may not be necessarily enough to cover, e.g., the fast mobility of a Virtual Machine.

Section 4.3.3. "Mobility" of [RIFT] specifies an hybrid method that combines a sequence counter from the mobile node and a timestamp from the network taken at the leaf when the route is injected. If the timestamps of the concurrent advertisements are comparable (i.e., more distant than the precision of the timing protocol), then the timestamp alone is used to determine the relative freshness of the routes. Otherwise, the sequence counter from the mobile node, if available, is used. One caveat is that the sequence counter must not

wrap within the precision of the timing protocol. Another is that the mobile node may not even provide a sequence counter, in which case the mobility itself must be slower than the precision of the timing.

Mobility must not be confused with Anycast. In both cases, a same address is injected in RIFT at different leaves. In the case of mobility, only the freshest route must be conserved, since mobile node changed its point of attachment for a leaf to the next. In the case of anycast, the node may be either multihomed (attached to multiple leaves in parallel) or reachable beyond the fabric via multiple routes that are redistributed to different leaves; either way, in the case of anycast, the multiple routes are equally valid and should be conserved. Without further information from the redistributed routing protocol, it is impossible to sort out a movement from a redistribution that happens asynchronously on different leaves. [RIFT] expects that anycast addresses are advertised within the timing precision, which is typically the case with a low-precision timing and a multihomed node. Beyond that time interval, RIFT interprets the lag as a mobility and only the freshest route is retained.

When using IPv6 [RFC8200], RIFT suggests to leverage "Registration Extensions for IPv6 over Low-Power Wireless Personal Area Network (6LoWPAN) Neighbor Discovery (ND)" [RFC8505] as the IPv6 ND interaction between the mobile node and the leaf. This provides not only a sequence counter but also a lifetime and a security token that may be used to protect the ownership of an address. When using [RFC8505], the parallel registration of an anycast address to multiple leaves is done with the same sequence counter, whereas the sequence counter is incremented when the point of attachement changes. This way, it is possible to differentiate a mobile node from a multihomed node, even when the mobility happens within the timing precision. It is also possible for a mobile node to be multihomed as well, e.g., to change only one of its points of attachement.

4.8. IPv4 over IPv6

RIFT allows advertising IPv4 prefixes over IPv6 RIFT network. IPv6 AF configures via the usual ND mechanisms and then V4 can use V6 nexthops analogous to RFC5549. It is expected that the whole fabric supports the same type of forwarding of address families on all the links. RIFT provides an indication whether a node is v4 forwarding capable and implementations are possible where different routing tables are computed per address family as long as the computation remains loop-free.

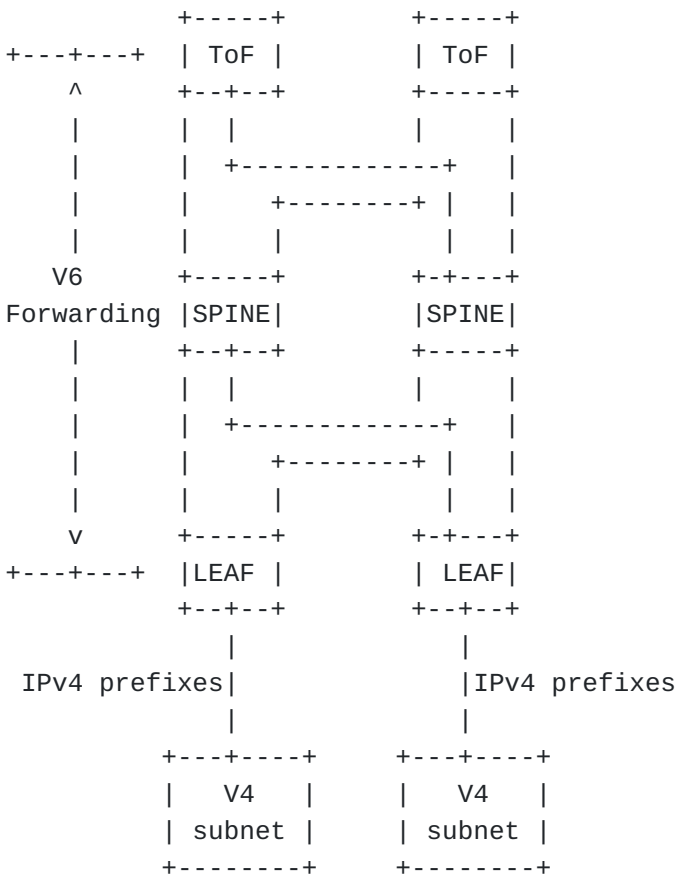

Figure 8: IPv4 over IPv6

4.9. In-Band Reachability of Nodes

RIFT doesn't precondition that nodes of the fabric have reachable addresses. But the operational purposes to reach the internal nodes may exist. Figure 9 shows an example that the NMS attaches to LEAF1.

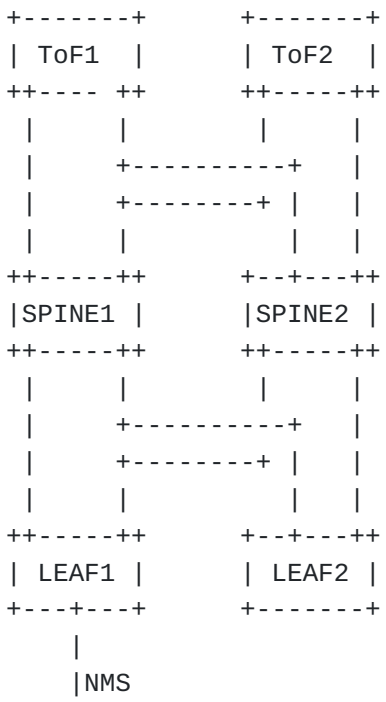

Figure 9: In-Band reachability of node

If NMS wants to access LEAF2, it simply works. Because loopback address of LEAF2 is flooded in its Prefix North TIE.

If NMS wants to access SPINE2, it simply works too. Because spine node always advertises its loopback address in the Prefix North TIE. NMS may reach SPINE2 from LEAF1-SPINE2 or LEAF1-SPINE1-ToF1/ToF2-SPINE2.

If NMS wants to access ToF2, ToF2's loopback address needs to be injected into its Prefix South TIE. Otherwise, the traffic from NMS may be sent to ToF1.

And in case of failure between ToF2 and spine nodes, ToF2's loopback address must be sent all the way down to the leaves.

4.10. Dual Homing Servers

Each RIFT node may operate in zero touch provisioning (ZTP) mode. It has no configuration (unless it is a Top-of-Fabric at the top of the topology or the must operate in the topology as leaf and/or support leaf-2-leaf procedures) and it will fully configure itself after being attached to the topology.

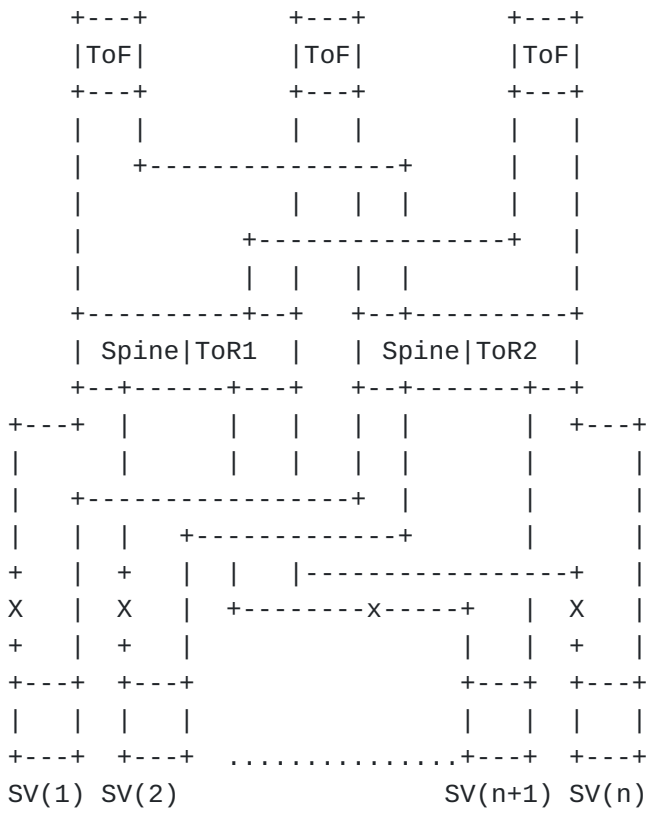

Figure 10: Dual-homing servers

In the single plane, the worst condition is disaggregation of every other servers at the same level. Suppose the links from ToR1 to all the leaves become not available. All the servers' routes are disaggregated and the FIB of the servers will be expanded with n-1 more specific routes.

Sometimes, people may prefer to disaggregate from ToR to servers from start on, i.e. the servers have couple tens of routes in FIB from start on beside default routes to avoid breakages at rack level. Full disaggregation of the fabric could be achieved by configuration supported by RIFT.

4.11. Fabric With A Controller

There are many different ways to deploy the controller. One possibility is attaching a controller to the RIFT domain from ToF and another possibility is attaching a controller from the leaf.

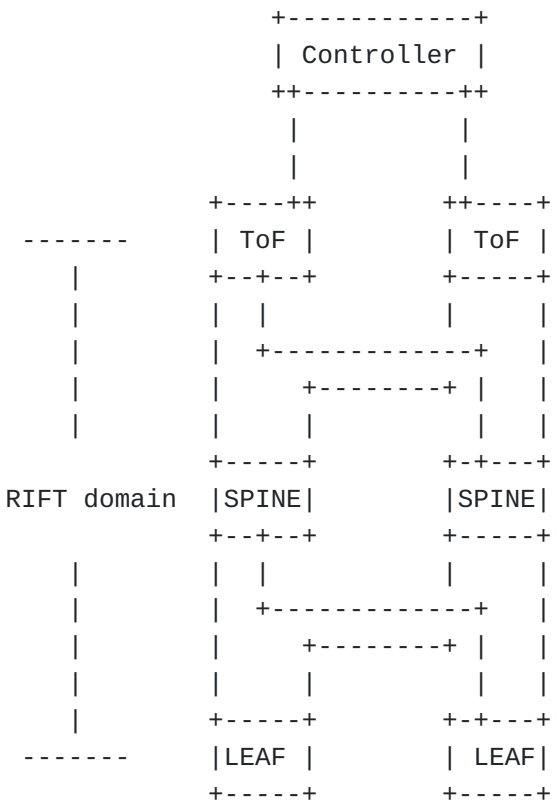

Figure 11: Fabric with a controller

4.11.1. Controller Attached to ToFs

If a controller is attaching to the RIFT domain from ToF, it usually uses dual-homing connections. The loopback prefix of the controller should be advertised down by the ToF and spine to leaves. If the controller loses link to ToF, make sure the ToF withdraw the prefix of the controller(use different mechanisms).

4.11.2. Controller Attached to Leaf

If the controller is attaching from a leaf to the fabric, no special provisions are needed.

4.12. Internet Connectivity With Underlay

If global addressing is running without overlay, an external default route needs to be advertised through rift fabric to achieve internet connectivity. For the purpose of forwarding of the entire rift fabric, an internal fabric prefix needs to be advertised in the South Prefix TIE by ToF and spine nodes.

4.12.1. Internet Default on the Leaf

In case that an internet access request comes from a leaf and the internet gateway is another leaf, the leaf node as the internet gateway needs to advertise a default route in its Prefix North TIE.

4.12.2. Internet Default on the ToFs

In case that an internet access request comes from a leaf and the internet gateway is a ToF, the ToF and spine nodes need to advertise a default route in the Prefix South TIE.

4.13. Subnet Mismatch and Address Families

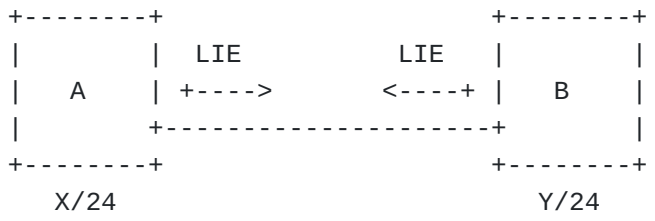

Figure 12: subnet mismatch

LIEs are exchanged over all links running RIFT to perform Link (Neighbor) Discovery. A node MUST NOT originate LIEs on an address family if it does not process received LIEs on that family. LIEs on same link are considered part of the same negotiation independent on the address family they arrive on. An implementation MUST be ready to accept TIEs on all addresses it used as source of LIE frames.

As shown in the above figure, without further checks adjacency of node A and B may form, but the forwarding between node A and node B may fail because subnet X mismatches with subnet Y.

To prevent this a RIFT implementation should check for subnet mismatch just like e.g. ISIS does. This can lead to scenarios where an adjacency, despite exchange of LIEs in both address families may end up having an adjacency in a single AF only. This is a consideration especially in Section 4.8 scenarios.

4.14. Anycast Considerations

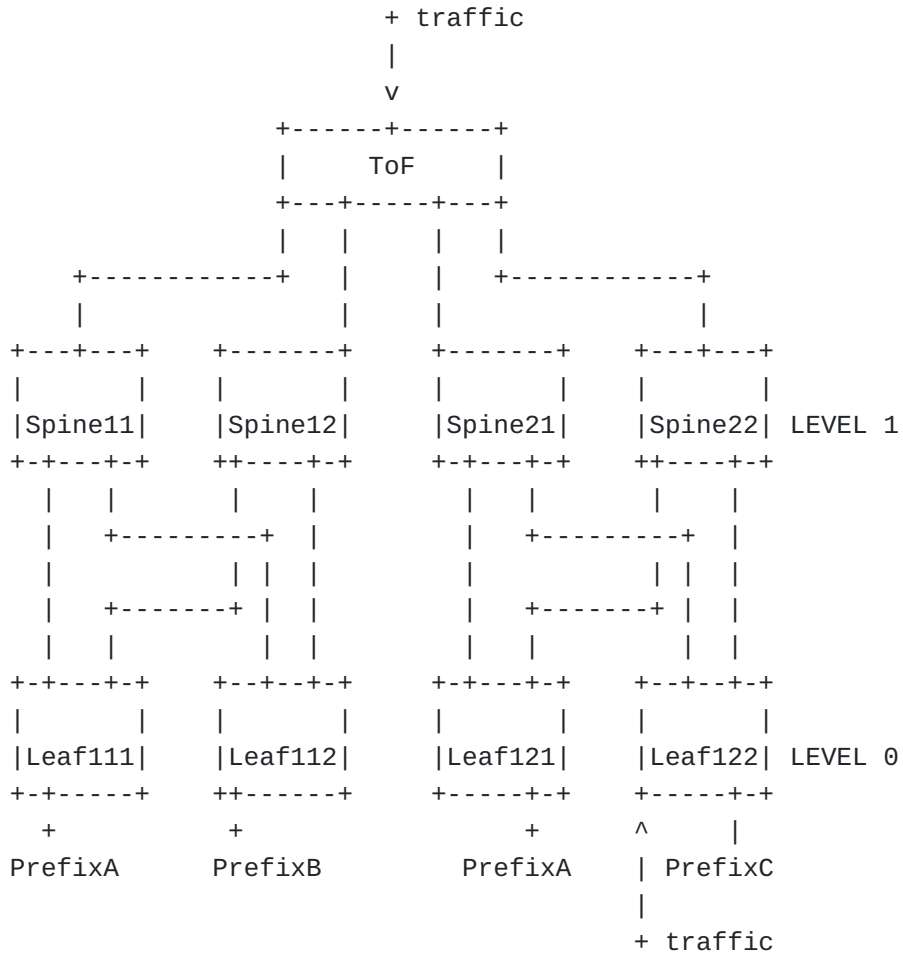

Figure 13: Anycast

If the traffic comes from ToF to Leaf111 or Leaf121 which has anycast prefix PrefixA. RIFT can deal with this case well. But if the traffic comes from Leaf122, it arrives Spine21 or Spine22 at level 1. But Spine21 or Spine22 doesn't know another PrefixA attaching Leaf111. So it will always get to Leaf121 and never get to Leaf111. If the intension is that the traffic should been offloaded to Leaf111, then use policy guided prefixes [PGP reference].

5. Acknowledgements

6. Contributors

The following people (listed in alphabetical order) contributed significantly to the content of this document and should be considered co-authors:

Tony Przygienda

Juniper Networks

1194 N. Mathilda Ave

Sunnyvale, CA 94089

US

Email: prz@juniper.net

7. Normative References

[ISO10589-Second-Edition] International Organization for Standardization, "Intermediate system to Intermediate system intra-domain routeing information exchange protocol for use in conjunction with the protocol for providing the connectionless-mode Network Service (ISO 8473)", November 2002.

[TR-384] Broadband Forum Technical Report, "TR-384 Cloud Central Office Reference Architectural Framework", January 2018.

[RFC2328] Moy, J., "OSPF Version 2", STD 54, RFC 2328, DOI 10.17487/RFC2328, April 1998, <https://www.rfc-editor.org/info/rfc2328>.

[RFC4861] Narten, T., Nordmark, E., Simpson, W., and H. Soliman, "Neighbor Discovery for IP version 6 (IPv6)", RFC 4861, DOI 10.17487/RFC4861, September 2007, <https://www.rfc-editor.org/info/rfc4861>.

[RFC5357] Hedayat, K., Krzanowski, R., Morton, A., Yum, K., and J. Babiarz, "A Two-Way Active Measurement Protocol (TWAMP)", RFC 5357, DOI 10.17487/RFC5357, October 2008, <https://www.rfc-editor.org/info/rfc5357>.

[RFC7130] Bhatia, M., Ed., Chen, M., Ed., Boutros, S., Ed., Binderberger, M., Ed., and J. Haas, Ed., "Bidirectional Forwarding Detection (BFD) on Link Aggregation Group (LAG) Interfaces", RFC 7130, DOI 10.17487/RFC7130, February 2014, <https://www.rfc-editor.org/info/rfc7130>.

[RIFT] Przygienda, T., Sharma, A., Thubert, P., Rijsman, B., and D. Afanasiev, "RIFT: Routing in Fat Trees", Work in Progress, Internet-Draft, draft-ietf-rift-rift-12, 26 May 2020, <https://tools.ietf.org/html/draft-ietf-rift-rift-12>.

[I-D.white-distoptflood] White, R., Hegde, S., and S. Zandi, "IS-IS Optimal Distributed Flooding for Dense Topologies", Work in Progress, Internet-Draft, draft-white-distoptflood-04, 27 July 2020, <https://tools.ietf.org/html/draft-white-distoptflood-04>.

8. Informative References

[RFC5905] Mills, D., Martin, J., Ed., Burbank, J., and W. Kasch, "Network Time Protocol Version 4: Protocol and Algorithms Specification", RFC 5905, DOI 10.17487/RFC5905, June 2010, <https://www.rfc-editor.org/info/rfc5905>.

[RFC8200] Deering, S. and R. Hinden, "Internet Protocol, Version 6 (IPv6) Specification", STD 86, RFC 8200, DOI 10.17487/RFC8200, July 2017, <https://www.rfc-editor.org/info/rfc8200>.

[RFC8505] Thubert, P., Ed., Nordmark, E., Chakrabarti, S., and C. Perkins, "Registration Extensions for IPv6 over Low-Power Wireless Personal Area Network (6LoWPAN) Neighbor Discovery", RFC 8505, DOI 10.17487/RFC8505, November 2018, <https://www.rfc-editor.org/info/rfc8505>.

Authors' Addresses

Yuehua Wei (editor)
ZTE Corporation
No.50, Software Avenue
Nanjing
210012
China

Email: wei.yuehua@zte.com.cn

Zheng Zhang
ZTE Corporation
No.50, Software Avenue
Nanjing
210012
China

Email: zhang.zheng@zte.com.cn

Dmitry Afanasiev
Yandex

Email: fl0w@yandex-team.ru

Tom Verhaeg
Juniper Networks

Email: tverhaeg@juniper.net

Jaroslaw Kowalczyk
Orange Polska

Email: jaroslaw.kowalczyk2@orange.com

Pascal Thubert
Cisco Systems, Inc
Building D
45 Allee des Ormes - BP1200
06254 MOUGINS - Sophia Antipolis
France

Phone: +33 497 23 26 34
Email: pthubert@cisco.com