RIFT Working Group
Internet-Draft
Intended status: Informational
Expires: 19 October 2026

J. D. Gomez
Independent

17 April 2026

Operational Considerations for Multicast over RIFT-based Data Center Fabrics

draft-gomez-rift-multicast-operational-00

Abstract

RIFT (Routing in Fat Trees) is increasingly used as an underlay routing protocol in modern data center fabrics. However, RIFT does not natively define mechanisms for multicast traffic distribution.

This document provides operational guidance and best practices for deploying multicast in RIFT-based data center fabrics. It analyzes PIM, EVPN multicast, BIER, and head-end replication, highlighting trade-offs in scalability, efficiency, and operational complexity.

This document does not define new protocol mechanisms. It aims to assist network operators in making informed design decisions when deploying multicast services over RIFT-based fabrics.

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 19 October 2026.

Copyright Notice

Table of Contents

1. Introduction

Modern data center fabrics rely on Clos-based topologies to achieve scalability, high bandwidth, and fault tolerance. RIFT [RFC9692] provides efficient unicast routing with topology awareness and Zero Touch Provisioning (ZTP).

Multicast traffic remains relevant for telemetry distribution, financial data delivery, streaming, and EVPN-VXLAN BUM traffic handling. However, RIFT does not define native multicast capabilities.

Operators must rely on external mechanisms such as PIM [RFC4601], EVPN multicast [RFC7432], BIER [RFC8279], or head-end replication. This document provides guidance for selecting and operating these mechanisms in RIFT-based fabrics.

2. Terminology

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.

RIFT Fabric: A Clos or fat-tree data center network using RIFT as the underlay.

Leaf Node: A node connecting endpoints to the fabric.

Spine Node: A node interconnecting leaf nodes within the fabric.

BUM Traffic: Broadcast, Unknown Unicast, and Multicast traffic in EVPN-VXLAN.

VTEP: VXLAN Tunnel Endpoint.

Head-End Replication (HER): Ingress node replicates packets to each remote receiver via unicast.

RP (Rendezvous Point): PIM-SM node coordinating multicast group membership and tree construction.

RPF (Reverse Path Forwarding): Mechanism validating multicast packet ingress interface against the unicast routing table.

KV-TIE: Key-Value Topology Information Element. RIFT mechanism for distributing key-value data within the fabric.

ZTP: Zero Touch Provisioning. RIFT automatic node-level discovery and configuration.

3. Applicability Statement

This document applies to data center fabrics using RIFT [RFC9692] as the underlay routing protocol, specifically Clos and fat-tree topologies as described in [RFC9696]. It is not intended for general-purpose IP networks or WAN environments.

Numerical values and thresholds presented in this document are illustrative and may vary by platform. Operators SHOULD validate all thresholds against their platform documentation prior to production deployment.

4. Problem Statement

RIFT provides efficient unicast routing but does not define mechanisms for multicast group membership or tree construction. Existing solutions introduce the following trade-offs:

* PIM introduces control-plane complexity requiring RP placement design and specialized expertise.

* EVPN multicast increases signaling overhead and requires overlay-underlay coordination.

* Head-end replication does not scale efficiently; cost grows linearly with VTEP count.

* None of these approaches natively leverage RIFT topology awareness or KV-TIE distribution.

There is no standardized operational guidance for multicast in RIFT-based fabrics. This document addresses this gap.

5. Multicast Deployment Models

Operators should evaluate deployment models based on scale, platform support, and operational requirements. Table 1 summarizes the trade-offs.

Model	Config	Underlay Mcast	Scale	Convergence
RIFT Native	Auto	None	Excellent	Excellent
PIM	Medium	Required	Good	1-5 s
EVPN / IR	Low	None	Limited	Good
EVPN / Incl. Trees	High	Required	Medium	Moderate
BIER	Auto	None	Excellent	Sub-second
Head-End Replication	Low	None	Poor	Good

Table 1: Deployment Model Comparison

5.1. Native RIFT Multicast

Work is ongoing [RIFT-MULTICAST] to define native RIFT multicast support. Operators MAY consider this approach once standardized. It is expected to provide automatic tree construction via ZTP and KV-TIEs.

5.2. Underlay Multicast using PIM

PIM MAY be deployed over a RIFT underlay. The RP SHOULD be placed on spine nodes. PIM BSR SHOULD be used for automatic RP advertisement. A dedicated multicast group range SHOULD be assigned for fabric use. Convergence is typically 1 to 5 seconds after failure.

5.3. Overlay Multicast using EVPN

EVPN multicast [RFC7432] MAY be used over a RIFT underlay. Ingress Replication is RECOMMENDED for small deployments only (fewer than 100 VTEPs). Inclusive Multicast Trees scale better but require underlay multicast. Operators SHOULD carefully plan VNI-to-multicast group mappings.

5.4. BIER-based Multicast

BIER [RFC8279] MAY be used as a multicast forwarding mechanism. [RFC9624] defines how BIER optimizes EVPN BUM forwarding. BIER is stateless in intermediate nodes and provides sub-second convergence. Operators SHOULD evaluate BIER where platform support is available.

5.5. Head-End Replication

Head-end replication MAY be used for simplicity. It is RECOMMENDED only for small deployments (fewer than 50 VTEPs) and SHOULD NOT be used in large fabrics.

6. Operational Best Practices

The values presented are illustrative and may vary by platform.

6.1. Multicast Group Planning

A multicast group allocation policy SHOULD be established prior to deployment. Each VNI SHOULD be mapped to a unique multicast group for EVPN BUM traffic. All mappings SHOULD be documented centrally to prevent conflicts.

6.2. RP Redundancy Strategies

A primary RP and at least one backup RP SHOULD be configured on different spine nodes. PIM Anycast-RP [RFC4610] or BSR-based redundancy MAY be used for automatic failover.

6.3. Configuration Automation

Operators SHOULD use configuration templates and automation frameworks for consistent configuration. Manual per-node configuration is error-prone and does not scale beyond 100 nodes.

6.4. Convergence Validation

Before production deployment, operators SHOULD validate convergence by deliberately failing an RP or spine node and measuring traffic interruption duration.

7. Monitoring and Observability

Operators SHOULD monitor multicast traffic using platform telemetry. Table 2 provides recommended alert thresholds. Values are illustrative.

Metric	Alert Threshold	Recommended Action
Active Multicast Groups	> 80% of HW limit	Add capacity / optimize
Active Sources	Unexpected spike	Investigate source
RP CPU Usage	> 70%	Add RP / rate-limit
BUM Traffic Volume	> 150% of baseline	Check storm / misconfig
PIM Join/Prune Rate	Sustained high rate	Check topology stability

Table 2: Key Multicast Monitoring Metrics

8. Convergence Considerations

Multicast convergence depends on underlay convergence time, protocol behavior, and control-plane load. Some packet loss during convergence is expected. Operators SHOULD use fast failure detection on PIM-enabled links.

Metric	Small DC (<100)	Medium DC (100-500)	Large DC (500+)
Max Multicast Groups	~1,000	~5,000	~50,000+
Recommended RPs	1	2-3	4-6
Convergence Target	< 1 second	1-5 seconds	5-10 seconds
Recommended BUM Model	HER / PIM	Inclusive Trees	BIER / Dist. RPs
HW Threshold Alert	80% of 4K	80% of 16K	80% of 64K+

Table 3: Scalability and Convergence Guidelines

9. BUM Traffic Handling

9.1. Broadcast Traffic

EVPN ARP suppression SHOULD be enabled on all leaf nodes. DHCP relay SHOULD be configured on leaf nodes. Broadcast rate limiting SHOULD be configured to prevent storm propagation.

9.2. Unknown Unicast

Leaf nodes SHOULD properly learn and advertise MAC addresses via EVPN Type 2 routes. Unknown unicast rates per VNI SHOULD be monitored. Unknown unicast suppression SHOULD be enabled where supported.

9.3. Multicast Traffic

IGMP snooping SHOULD be enabled on all leaf nodes. An IGMP querier SHOULD be designated per broadcast domain. For IPv6 segments, MLD snooping and an MLD querier SHOULD be configured.

10. Use Cases and Deployment Examples

10.1. Small Data Center (fewer than 100 Nodes)

A small RIFT fabric (2-4 spines, 10-50 leaves) prioritizes simplicity. RIFT ZTP SHOULD be deployed. EVPN with Head-End Replication is RECOMMENDED for BUM. If native multicast is required, a single RP on one spine node is sufficient. ARP suppression SHOULD be enabled on all leaves.

10.2. Medium Data Center (100 to 500 Nodes)

A medium RIFT fabric (10-30 spines, 100-300 leaves) requires scalability and redundancy. Operators SHOULD deploy 2-3 RPs on different spine nodes with PIM BSR. EVPN inclusive multicast trees are RECOMMENDED over head-end replication. Separate multicast group ranges SHOULD be assigned for infrastructure and application traffic.

10.3. Large Data Center (500 or More Nodes)

A large RIFT fabric (50+ spines, 500+ leaves) requires high efficiency. BIER [RFC8279] [RFC9624] is RECOMMENDED where platform support is available. If BIER is unavailable, 4-6 distributed RPs with PIM Anycast-RP [RFC4610] SHOULD be deployed. Regional multicast domains SHOULD limit state propagation.

11. Troubleshooting

Diagnostic procedures vary by platform. Operators SHOULD consult platform documentation for specific commands and tools.

11.1. Multicast Traffic Not Reaching All VTEPs

Possible causes: incorrect IGMP/MLD group membership on leaf nodes; EVPN IMET routes not advertised; RPF check failing due to asymmetric routing. Steps: verify group membership state; verify EVPN IMET route advertisement; check RPF path; confirm RIFT unicast convergence is complete.

11.2. RP Node High CPU

Possible causes: excessive PIM Register messages; (S,G) state approaching platform limits. Steps: examine PIM message rates on the RP; check active multicast state count; consider additional RPs or BIER migration.

11.3. Slow Multicast Convergence

Possible causes: PIM waiting for RIFT unicast convergence; slow backup RP detection; fast failure detection not configured. Steps: measure RIFT convergence time independently; verify fast failure detection on PIM links; test RP failover in a lab.

11.4. Excessive Broadcast Traffic

Possible causes: ARP suppression not enabled; host generating broadcast storm; MAC learning failures. Steps: verify ARP suppression per VNI; monitor per-interface broadcast counters; identify source via MAC address tables; inspect offending host.

12. Security Considerations

Multicast deployments MAY introduce flooding, amplification, and unauthorized group access risks. Operators SHOULD enable RPF checking; use SSM with IGMPv3 where possible; filter 224.0.0.0/24 at fabric boundaries (this range MUST NOT be routed across the fabric); configure per-source rate limits; and apply control-plane policing on RP nodes.

13. IANA Considerations

This document has no IANA actions.

14. References

14.1. 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, <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>.

[RFC9692] Przygienda, T., Ed., Head, J., Ed., Sharma, A., and B. Rijsman, "RIFT: Routing in Fat Trees", RFC 9692, DOI 10.17487/RFC9692, April 2025, <https://www.rfc-editor.org/info/rfc9692>.

14.2. Informative References

[RFC4601] Fenner, B., Ed., "Protocol Independent Multicast - Sparse Mode (PIM-SM)", RFC 4601, DOI 10.17487/RFC4601, August 2006, <https://www.rfc-editor.org/info/rfc4601>.

[RFC4610] Farinacci, D. and Y. Cai, "Anycast-RP Using Protocol Independent Multicast (PIM)", RFC 4610, DOI 10.17487/RFC4610, August 2006, <https://www.rfc-editor.org/info/rfc4610>.

[RFC7432] Sajassi, A., Ed., "BGP MPLS-Based Ethernet VPN", RFC 7432, DOI 10.17487/RFC7432, February 2015, <https://www.rfc-editor.org/info/rfc7432>.

[RFC8279] Wijnands, I. J., Ed. and E. Rosen, Ed., "Multicast Using Bit Index Explicit Replication (BIER)", RFC 8279, DOI 10.17487/RFC8279, November 2017, <https://www.rfc-editor.org/info/rfc8279>.

[RFC8365] Sajassi, A., Ed., "A Network Virtualization Overlay Solution Using Ethernet VPN (EVPN)", RFC 8365, DOI 10.17487/RFC8365, March 2018, <https://www.rfc-editor.org/info/rfc8365>.

[RFC9251] Sajassi, A., Ed., "IGMP and MLD Proxies for Ethernet VPN (EVPN)", RFC 9251, DOI 10.17487/RFC9251, June 2022, <https://www.rfc-editor.org/info/rfc9251>.

[RFC9624] Zhang, Z., Przygienda, T., and A. Sajassi, "EVPN BUM Using Bit Index Explicit Replication (BIER)", RFC 9624, DOI 10.17487/RFC9624, August 2024, <https://www.rfc-editor.org/info/rfc9624>.

[RFC9696] Przygienda, T., Ed., "RIFT: Applicability and Operational Considerations", RFC 9696, DOI 10.17487/RFC9696, April 2025, <https://www.rfc-editor.org/info/rfc9696>.

[RIFT-MULTICAST] Zhang, Z., "Multicast Routing In Fat Trees", Work in Progress, Internet-Draft, draft-zzhang-rift-multicast, April 2025, <https://datatracker.ietf.org/doc/draft-zzhang-rift-multicast/>.

Appendix A: Two-Level RIFT Fabric (Small/Medium DC)

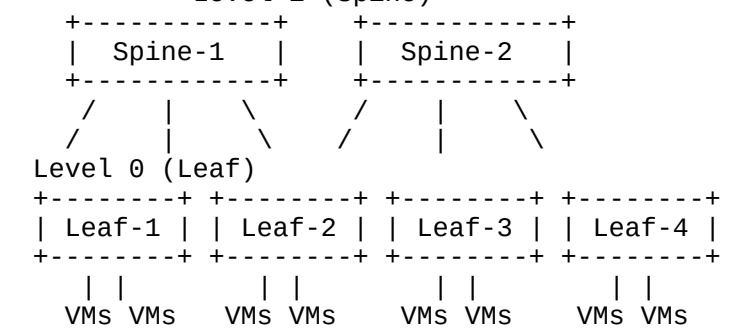

Figure 1: Two-Level RIFT Fabric

Appendix B: PIM Multicast Tree Example

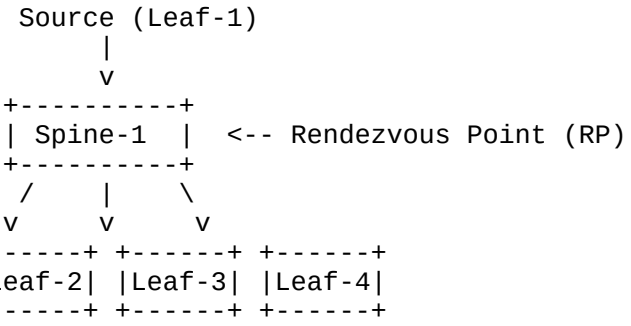

Figure 2: PIM Multicast Tree in a RIFT Fabric

Appendix C: EVPN BUM Head-End Replication

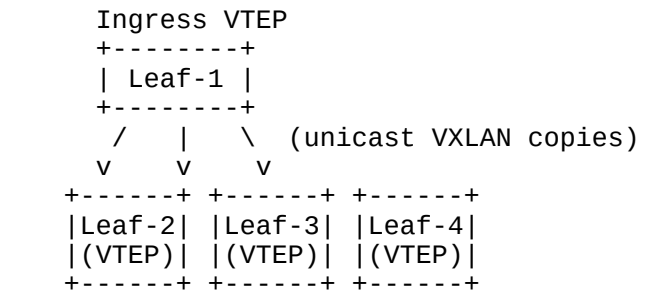

Figure 3: EVPN BUM Head-End Replication

Appendix D: Large DC Multi-Level Topology

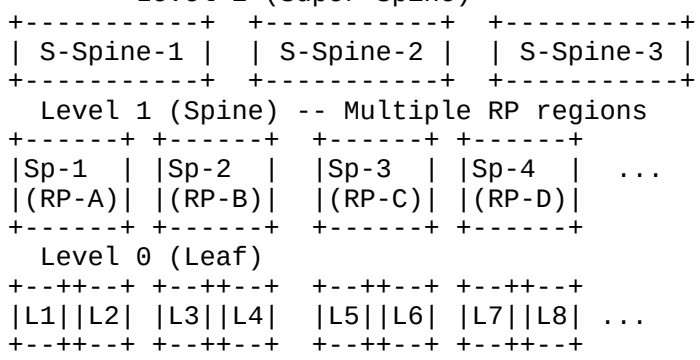

Figure 4: Three-Level RIFT Fabric (Large DC)

Author’s Address

Jesus David Gomez Zavala
Independent
Email: jesusdavid.ieft@gmail.com