

BESS WorkGroup
Internet-Draft
Intended status: Standards Track
Expires: 8 January 2026

A. Sajassi
M. Mishra
S. Thoria
Cisco Systems
J. Rabadan
Nokia
J. Drake
Independent
7 July 2025

Per multicast flow Designated Forwarder Election for EVPN
draft-ietf-bess-evpn-per-mcast-flow-df-election-12

Abstract

This document defines an enhancement to the Designated Forwarder (DF) election process in Ethernet Virtual Private Network (EVPN) environments. While traditional DF election operates at a per Virtual Local Area Network (VLAN) or per group of VLANs (in case of VLAN bundle or VLAN-aware bundle service) level, such granularity may not be sufficient for applications requiring optimized or isolated multicast forwarding. This specification introduces a refined DF election mechanism that extends existing hash-based methods to operate at a more granular level specifically at the tuple of Ethernet Segment Identifier (ESI), VLAN, and multicast flow. This approach enables improved traffic distribution, enhanced load balancing, and greater deployment flexibility for multicast delivery in EVPN based networks. The proposed method is designed to remain compatible with existing DF election procedures while offering targeted improvements for multicast scenarios.

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 January 2026.

Copyright Notice

Copyright (c) 2025 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. Specification of Requirements	5
3. Terminology	5
4. The DF Election Extended Community	5
5. HRW base per multicast flow EVPN DF election	7
5.1. DF election for Multicast (S,G) membership request	8
5.2. DF election for Multicast (*,G) membership request	10
5.3. Default DF election procedure	10
6. Procedure to use per multicast flow DF election algorithm	11
7. Triggers for DF re-election	13
8. Security Considerations	14
9. IANA Considerations	14
10. Acknowledgement	14
11. Normative References	14
Authors' Addresses	15

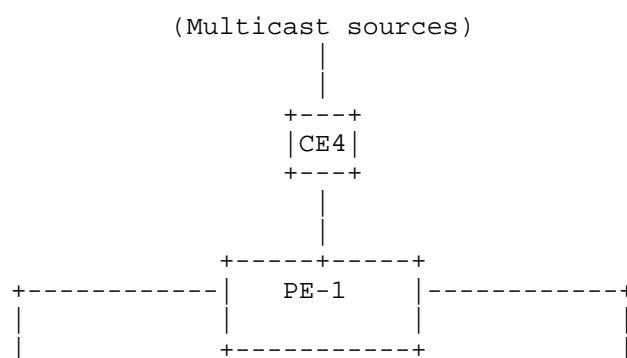
1. Introduction

[RFC7432] defines the procedures for Designated Forwarder (DF) election in Ethernet Virtual Private Network (EVPN) networks at the granularity of the Ethernet Segment Identifier (ESI) and the EVPN Instance (EVI), which typically maps to a single Virtual Local Area Network (VLAN) or a group of VLANs in the case of VLAN-bundle or VLAN-aware bundle services. This per-VLAN granularity, however, is not always sufficient to meet the operational and performance needs of certain multicast applications.

[RFC8584] enhances the default DF election by introducing the Highest Random Weight (HRW) algorithm ([HRW1999]) to provide more deterministic and stable DF selection. Separately, [RFC9251] extends EVPN to support multicast flows by defining the route types used to synchronize multicast state and specifying the procedures for multicast DF election.

This document proposes an extension to the [HRW1999] based DF election mechanism defined in [RFC8584]. The extension enables DF election at a finer granularity specifically, per (ESI, VLAN, multicast flow). This allows for improved distribution of multicast traffic across multiple Provider Edge (PE) devices in an all-active multi-homing environment.

As defined in [RFC7432], the DF in an all-active redundancy group is responsible for forwarding Broadcast, Unknown unicast, and Multicast (BUM) traffic on a given Ethernet Segment, whether attached to a Customer Edge (CE) device or an access network. By default, the DF election process regardless of whether it uses the procedures in [[RFC7432] or the HRW enhancements in [RFC8584] selects a single Provider Edge (PE) to forward all BUM traffic for a given (ESI, VLAN) tuple. While this model works well for many use cases, it introduces limitations in scenarios where multicast flows dominate the traffic mix and require more granular distribution for optimal load balancing. For example, if a deployment requires that all multicast traffic be delivered over a single Virtual Local Area Network (VLAN), the existing DF election procedures will result in a single Provider Edge (PE) node being responsible for forwarding the entire multicast traffic load to the access network. This static forwarding responsibility can lead to suboptimal bandwidth utilization and lack of resiliency in scenarios where multiple PEs are available and capable of sharing the multicast forwarding load.



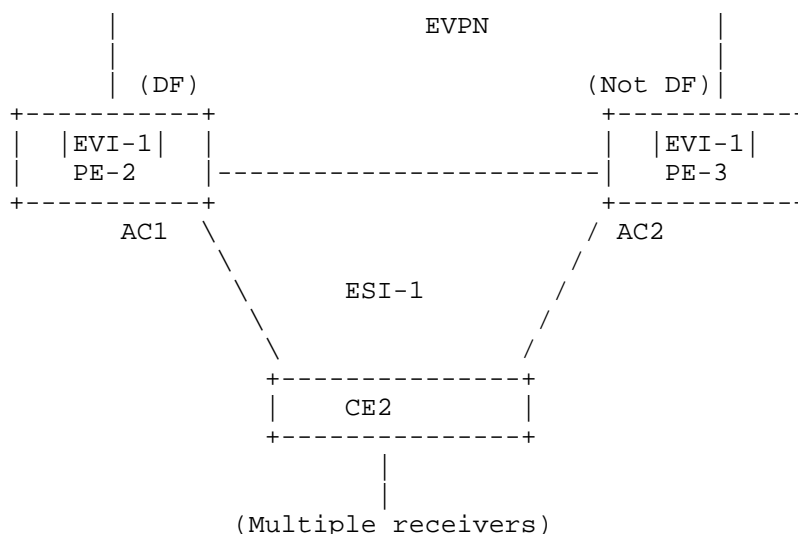


Figure 1: Multi-homing Network of EVPN
for IPTV (Internet Protocol Television) deployments

For example, consider the topology described above, which illustrates a typical residential deployment where multiple multicast receivers are connected to a Customer Edge (CE) device that is multi-homed to a set of Provider Edge (PE) routers. In such scenarios, all multicast traffic (e.g., for IPTV services) may be transported within a single Ethernet Virtual Private Network (EVPN) Instance (EVI). If PE-2 is elected as the Designated Forwarder (DF), then as per the procedures defined in [RFC7432], it becomes solely responsible for forwarding all multicast traffic to the associated Ethernet Segment. This forwarding responsibility can result in resource imbalance across PE nodes and lead to inefficient bandwidth utilization, particularly when other PEs in the redundancy group have available capacity to share the multicast forwarding load.

To address these limitations, this document defines a method to perform DF election at the granularity of (ESI, VLAN, multicast flow). By distributing multicast flows across different PE nodes within a redundancy group, the proposed mechanism improves bandwidth utilization and enables finer grained load balancing while remaining compatible with existing EVPN control plane procedures.

2. Specification of Requirements

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119] .

3. Terminology

With respect to Ethernet Virtual Private Network (EVPN), this document adheres to the terminology defined in [RFC7432]. For multicast-related terms and procedures, it follows the conventions and definitions specified in [RFC7761].

AC Attachment Circuit

BUM Broadcast, Unknown unicast, Multicast

CE Customer Edge

CRC Cyclic Redundancy Check (Appendix A of [RFC9260])

DF Designated Forwarder

nDF - non-Designated Forwarder

EVI EVPN Instance

ES Ethernet Segment

ESI Ethernet Segment Identifier

HRW Highest Random Weight ([HRW1999])

IGMP Internet Group Management Protocol

PE Provider Edge

HRW - Highest Random Weight (A paper "Using Name-Based Mappings to Increase Hit Rates")

4. The DF Election Extended Community

[RFC8584] defines an extended community, which would be used for PEs in redundancy group to reach a consensus as to which DF election procedure is desired. A PE can notify other participating PEs in redundancy group about its willingness to support Per multicast flow base DF election capability by signaling a DF election extended community along with Ethernet-Segment Route (Type-4). The current

proposal extends the existing extended community defined in [RFC8584]. This draft defines new a DF type. [RFC8584] defines a DF Election Extended Community that is used by Provider Edge (PE) devices in a redundancy group to reach consensus on the DF election procedure to be used for a given Ethernet Segment (ES). A PE indicates its supported DF election capability by attaching this extended community to its Ethernet Segment Route (Route Type 4) advertisement. This document extends the DF Election Extended Community defined in [RFC8584] by introducing a new DF Alg to signal support for per-multicast flow-based DF election. A PE that supports the procedures specified in this document MUST signal the corresponding DF Alg in its Route Type 4 advertisement. This enables all participating PEs in the redundancy group to discover and agree upon the enhanced DF election behavior in a backward-compatible and interoperable manner.

- * DF Alg (5 bits) - Encodes the DF election algorithm values (between 0 and 31) that the advertising PE desires to use for the ES. This document requests two new alg in IANA registry called "DF Alg":
 - Type 0: Default DF election algorithm, or modulus-based algorithm as defined in [RFC7432].
 - Type 1: HRW Algorithm defined in [RFC8584]
 - Type 2: Highest-Preference Algorithm defined in [RFC9785].
 - Type 3: Lowest-Preference Algorithm defined in [RFC9785].
 - Type 4: HRW base per (S,G) multicast flow DF election (explained in this document).
 - Type 5: HRW base per (*,G) multicast flow DF election (explained in this document).
 - Type 6-30: Unassigned.
 - Type 31: Reserved for Experimental Use.

- * The [RFC8584] describes encoding of capabilities associated to the DF election algorithm using Bitmap field. When these capabilities bits are set along with the DF type-4 and type-5, they need to be interpreted in context of this new DF type-4 and type-5. For example, consider a scenario where all PEs in the same redundancy group (same ES) can support both AC-DF, DF type-4 and DF type-5 and receive such indications from the other PEs in the ES. In this scenario, if a VLAN is not active in a PE, then the DF election procedure on all PEs in the ES should factor that in and exclude that PE in the DF election per multicast flow.
- * A PE SHOULD attach the DF election Extended Community to ES route and Extended Community MUST be sent if the ES is locally configured for DF type Per Multicast flow DF election. Only one DF Election Extended community can be sent along with an ES route.
- * When a PE receives the ES Routes from all the other PEs for the ES, it checks if all of other PEs have advertised their desire to proceed by Per multicast flow DF election. If all peering PEs have done so, it performs DF election based on Per multicast flow procedure. But if:
 - There is at least one PE which advertised route-4 (AD per ES Route) which does not indicate its capability to perform Per multicast flow DF election. OR
 - There is at least one PE signaling single active in the AD per ES route

it MUST be considered as an indication to support of only Default DF election [RFC7432] and DF election procedure in [RFC7432] MUST be used.

5. HRW base per multicast flow EVPN DF election

This document is an extension of [RFC8584], so this draft does not repeat the description of HRW algorithm itself. This document is an extension of [RFC8584] and leverages the Highest Random Weight (HRW) algorithm for Designated Forwarder (DF) election. Therefore, this draft does not repeat the general description of the HRW algorithm itself. Section 3.2 of [RFC8584] defines the use of the HRW algorithm for DF election in Ethernet Virtual Private Network (EVPN) environments. This document enhances that mechanism by introducing additional input parameters from the multicast flow, allowing DF election to be performed at the granularity of (, , , Es) where:

S is the multicast Source IP address

G is the multicast Group IP address

V is the VLAN Identifier

Es is the Ethernet Segment Identifier

This per-flow extension enables more granular and balanced distribution of multicast traffic across all-active PE nodes in the redundancy group.

EVPN Provider Edge (PE) devices discover redundancy groups as specified in [RFC7432]. If a redundancy group consists of N peering EVPN PEs, then upon completion of the discovery process, each PE constructs an unordered list of IP addresses corresponding to all members of the redundancy group. The procedure described in this document does not require the list of PE addresses to be ordered. Let Address[i] denote the IP address of the ith EVPN PE in the redundancy group, where $(0 < i \leq N)$.

5.1. DF election for Multicast (S,G) membership request

The DF is the PE who has maximum weight for (S, G, V, ES) where

- * S - Multicast Source
- * G - Multicast Group
- * V - VLAN ID.
- * ESI - Ethernet Segment Identifier

Address[i] is address of the ith PE. The PEs IP address length does not matter as only the lower-order 31 bits are modulo significant.

1. Weight

- * The weight of PE(i) to (S,G,VLAN ID, ESI) is calculated by function, $\text{weight}(S,G,V, ESI, \text{Address}(i))$, where $(0 < i \leq N)$, PE(i) is the PE at ordinal i.
- * $\text{Weight}(S,G,V, ESI, \text{Address}(i)) = (1103515245 \cdot ((1103515245 \cdot \text{Address}(i) + 12345) \text{ XOR } D(S,G,V,ESI)) + 12345) \pmod{2^{31}}$
- * In the event of a tie, where two or more Provider Edge (PE) devices compute the same weight for a given (S,G,V,ESI) tuple, the PE whose IP address is numerically least (i.e., lowest in value when interpreted as an unsigned integer) MUST be elected

as the Designated Forwarder (DF). This tie-breaking rule ensures deterministic and consistent DF election results across all PEs in the redundancy group.

- * Since the Weight is a pseudorandom function with the domain as the four-tuple (S,G,VLAN ID, ESI), it is an efficient and deterministic algorithm that is independent of VLAN distribution. Choosing a good hash function for the pseudorandom function is an important consideration for this algorithm to perform better than the default algorithm. As mentioned previously [RFC8584] has chosen HRW to be algorithm, this draft enhances the same.

2. Digest

- * Each PE independently computes a 31-bit digest, $D(S,G,V,ESI)$, which is the CRC-32 (Appendix A of [RFC9260]) checksum of the input tuple with the most significant bit (MSB) discarded. The CRC MUST be computed using network byte order (big-endian) serialization of the tuple fields.
- * $D(S,G,V, ESI) = \text{CRC_32}(S,G,V, ESI)$
- * Here, $D(S, G, V, ESI)$ is the 31-bit digest, computed as the CRC-32 (Appendix A of [RFC9260]) of a concatenated input stream comprising the Source IP address (S), Group IP address (G), VLAN Identifier (V), and the 10-octet Ethernet Segment Identifier (ESI). The result of the CRC-32 computation MUST discard the most significant bit (MSB), yielding a 31-bit value, as described in [HRW]. The input stream used for the CRC calculation MUST be constructed by serializing the fields in network byte order (big-endian). The CRC computation itself MUST also assume network byte order for consistency across all participating nodes. The address of the i th Provider Edge (PE), denoted as $\text{Address}[i]$, may be of any length, but only the lower-order 31 bits are considered modulo significant during weight computation..

All the benefits described in Section 3.2 of [RFC8584] with respect to HRW-based Designated Forwarder (DF) election—such as deterministic selection, consistency across nodes, and minimal churn—are fully applicable to the mechanism defined in this document. In addition, this specification provides the added benefit of more granular distribution by extending the HRW input to include multicast flow identifiers, enabling improved load balancing across Provider Edge (PE) devices.

5.2. DF election for Multicast (*,G) membership request

In the case of multicast membership requests where the source address is not specified (e.g., for (*,G) joins), the input parameters for DF election are modified from (S,G,V,ESI) to (G,V,ESI). All procedures defined in the previous section remain applicable without change, except that the source address is excluded from both the digest computation and the weight calculation. Accordingly, the digest $D(G,V,ESI)$ is computed as the CRC-32 (Appendix A of [RFC9260]) of the concatenated stream of Group address (G), VLAN ID (V), and Ethernet Segment Identifier (ESI), with the most significant bit discarded to produce a 31-bit value. The CRC computation MUST follow the same serialization and byte order rules as previously defined. The updated weight function is as follows:

1. Weight

- * The weight of PE(i) to (G,VLAN ID, ESI) is calculated by function, $\text{weight}(G,V,ESI, \text{Address}(i))$, where $(0 < i \leq N)$, PE(i) is the PE at ordinal i.
- * $\text{Weight}(G,V,ESI, \text{Address}(i)) = (1103515245.((1103515245.\text{Address}(i) + 12345) \text{ XOR } D(G,V,ESI)) + 12345) \pmod{2^{31}}$

2. Digest

- * $D(G,V,ESI) = \text{CRC_32}(G,V,ESI)$

All remaining aspects of the DF election algorithm, including tie-breaking procedures, remain unchanged.

5.3. Default DF election procedure

The per-multicast flow Designated Forwarder (DF) election procedure defined in this document is applicable only after multicast membership activity is detected, i.e., when hosts behind the Attachment Circuit (AC) of a given Ethernet Segment (ES) begin sending Internet Group Management Protocol (IGMP) or Multicast Listener Discovery (MLD) membership reports. Membership information is synchronized across the participating Provider Edge (PE) devices in the redundancy group using the procedures defined in [RFC9251]. Once synchronized, each PE MAY apply the per-flow DF election procedure to create and maintain DF state on a per multicast flow basis (i.e., per (S,G) or (,G) flow). The "Type 1" Highest Random Weight (HRW) DF election procedure specified in [RFC8584] MUST be used to perform the default DF election for the Ethernet Segment. This election SHOULD be performed at the port level, independent of

multicast membership state, and SHOULD occur prior to any IGMP or MLD membership activity.

As multicast membership requests are subsequently learned and synchronized, the default port level DF state MUST be overridden by the per-flow DF election mechanism introduced in this document. This ensures that multicast traffic forwarding is transitioned from a single designated forwarder to a more granular, per-flow basis, improving distribution and load balancing across the redundancy group.

6. Procedure to use per multicast flow DF election algorithm

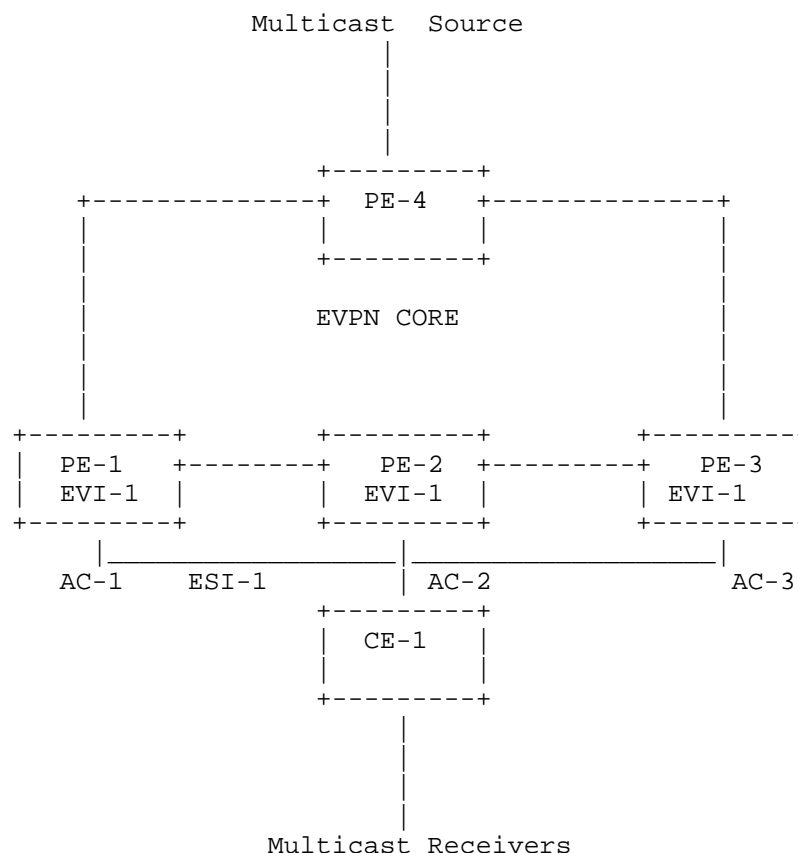


Figure-2 : Multihomed network

Figure-2 shows multihomed network. Where EVPN PE-1, PE-2, PE-3 are multihomed to CE-1. Multiple multicast receivers are behind all active multihoming segment.

1. Provider Edge (PE) devices connected to the same Ethernet Segment (ES) can automatically discover each other through the exchange of Ethernet Segment Routes. This document does not modify the discovery procedures and continues to rely on the mechanisms defined in [RFC7432].
2. Each Provider Edge (PE) device in the redundancy group advertises an Ethernet Segment (ES) Route that includes an Extended Community indicating its capability to support the per-multicast flow DF election procedure defined in this document. Since per-multicast flow DF election is applicable only after a PE learns

multicast membership state from receivers (e.g. via IGMP or MLD reports), a default DF election mechanism is required to forward Broadcast, Unknown unicast, and Multicast (BUM) traffic prior to the availability of such state. Until multicast membership state is learned, the default DF election procedure specified in Section 5.3, namely HRW per (v,Es) as defined in [RFC8584] . This ensures consistent and deterministic BUM forwarding behavior in the absence of flow-specific state.

3. When a receiver starts sending membership requests for (s1,g1), where s1 is multicast source address and g1 is multicast group address, CE-1 could hash membership request (IGMP join) to any of the PEs in redundancy group. Let's consider it is hashed to PE-2. [RFC9251] defines a procedure to sync IGMP join state among redundancy group of PEs. Now each of the PE would have information about membership request (s1,g1) and each of them run DF election procedure Section 5.1 to elect DF among participating PEs in redundancy group. Consider PE-2 gets elected as DF for multicast flow (s1,g1).

1. PE-1: non-Designated Forwarder (nDF) for flow (s1, g1), and DF for all other Broadcast, Unknown unicast, and Multicast (BUM) traffic
2. PE-2 forwarding state would be DF for flow (s1,g1) and nDF for rest other BUM traffic.
3. PE-3 forwarding state would be nDF for flow (s1,g1) and rest other BUM traffic.

4. As and when new multicast membership request comes, same procedure as above would continue.
5. If Section 4 has DF type 4, For membership request (S,G) it MUST use Section 5.1 to elect DF among participating PEs. And membership request (*,G) MUST use Section 5.2 to elect DF among participating PEs.

7. Triggers for DF re-election

There are multiple events that can trigger a Designated Forwarder (DF) re-election within the redundancy group. Some of these triggers include, but are not limited to :

1. Local ES going down due to physical failure or configuration change triggers DF re-election at peering PE.
2. Detection of new PE through ES route.

3. AC going up / down
4. ESI change
5. Remote PE removed / Down
6. Local configuration change of DF election Type and peering PE consensus on new DF Type

This document does not introduce any new mechanisms for Designated Forwarder (DF) re-election. Instead, it relies on the existing DF re-election procedures defined in [RFC7432]. Upon occurrence of any triggering event, a DF re-election is performed, resulting in the redistribution of all multicast flows among the Provider Edge (PE) devices within the redundancy group for the given Ethernet Segment (ES).

8. Security Considerations

The same Security Considerations described in [RFC7432] , [RFC8584] are valid for this document.

9. IANA Considerations

Per this document, request to allocate two new values:

Alg type 4: HRW base per (S,G) multicast flow DF election (explained in this document).

Alg type 5: HRW base per (*,G) multicast flow DF election (explained in this document).

List this document as an additional reference for the DF Election Extended Community field in the "EVPN Extended Community Sub-Types" registry on top of existing reference to RFC8584 and RFC9785.

10. Acknowledgement

Authors would like to acknowledge helpful comments and contributions of Luc Andre Burdet.

11. Normative References

- [HRW1999] Thaler, D. and C. Ravishankar, "Using Name-Based Mappings to Increase Hit Rates", IEEE/ACM Transactions in networking Volume 6 Issue 1, February 1998.

- [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>>.
- [RFC7432] Sajassi, A., Ed., Aggarwal, R., Bitar, N., Isaac, A., Uttaro, J., Drake, J., and W. Henderickx, "BGP MPLS-Based Ethernet VPN", RFC 7432, DOI 10.17487/RFC7432, February 2015, <<https://www.rfc-editor.org/info/rfc7432>>.
- [RFC7761] Fenner, B., Handley, M., Holbrook, H., Kouvelas, I., Parekh, R., Zhang, Z., and L. Zheng, "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)", STD 83, RFC 7761, DOI 10.17487/RFC7761, March 2016, <<https://www.rfc-editor.org/info/rfc7761>>.
- [RFC9251] Sajassi, A., Thoria, S., Mishra, M., Patel, K., Drake, J., and W. Lin, "Internet Group Management Protocol (IGMP) and Multicast Listener Discovery (MLD) Proxies for Ethernet VPN (EVPN)", RFC 9251, DOI 10.17487/RFC9251, June 2022, <<https://www.rfc-editor.org/info/rfc9251>>.
- [RFC8584] Rabadan, J., Ed., Mohanty, S., Ed., Sajassi, A., Drake, J., Nagaraj, K., and S. Sathappan, "Framework for Ethernet VPN Designated Forwarder Election Extensibility", RFC 8584, DOI 10.17487/RFC8584, April 2019, <<https://www.rfc-editor.org/info/rfc8584>>.
- [RFC9785] Rabadan, J., Ed., Sathappan, S., Lin, W., Drake, J., and A. Sajassi, "Preference-Based EVPN Designated Forwarder (DF) Election", RFC 9785, DOI 10.17487/RFC9785, June 2025, <<https://www.rfc-editor.org/info/rfc9785>>.
- [RFC9260] Stewart, R., Txen, M., and K. Nielsen, "Stream Control Transmission Protocol", RFC 9260, DOI 10.17487/RFC9260, June 2022, <<https://www.rfc-editor.org/info/rfc9260>>.

Authors' Addresses

Ali Sajassi
Cisco Systems
821 Alder Drive,
MILPITAS, CALIFORNIA 95035
United States
Email: sajassi@cisco.com

Mankamana Mishra
Cisco Systems
821 Alder Drive,
MILPITAS, CALIFORNIA 95035
United States
Email: mankamis@cisco.com

Samir Thoria
Cisco Systems
821 Alder Drive,
MILPITAS, CALIFORNIA 95035
United States
Email: sthoria@cisco.com

Jorge Rabadan
Nokia
777 E. Middlefield Road
Mountain View, CA 94043
United States
Email: jorge.rabadan@nokia.com

John Drake
Independent
Email: je_drake@yahoo.com