Workgroup: BESS Working Group
Internet-Draft: draft-ietf-bess-evpn-mh-pa-06
Published: 10 June 2022
Intended Status: Standards Track
Expires: 12 December 2022

Authors: P. Brissette, Ed.   A. Sajassi      LA. Burdet, Ed.
         Cisco Systems       Cisco Systems   Cisco Systems
         S. Thoria       B. Wen    E. Leyton
         Cisco Systems   Comcast   Verizon Wireless
         J. Rabadan
         Nokia

# EVPN multi-homing port-active load-balancing

## Abstract

The Multi-Chassis Link Aggregation Group (MC-LAG) technology enables establishing a logical link-aggregation connection with a redundant group of independent nodes. The purpose of multi-chassis LAG is to provide a solution to achieve higher network availability, while providing different modes of sharing/balancing of traffic. RFC7432 defines EVPN based MC-LAG with single-active and all-active multi-homing load-balancing mode. The current draft expands on existing redundancy mechanisms supported by EVPN and introduces support for port-active load-balancing mode.

## Status of This Memo

## Copyright Notice

## Table of Contents

## 1. Introduction

EVPN, as per [RFC7432], provides all-active per flow load-balancing for multi-homing. It also defines single-active with service carving mode, where one of the PEs, in redundancy relationship, is active per service.

While these two multi-homing scenarios are most widely utilized in data center and service provider access networks, there are scenarios where active-standby per interface multi-homing load-balancing is useful and required. The main consideration for this mode of load-balancing is the determinism of traffic forwarding through a specific interface rather than statistical per flow load-balancing across multiple PEs providing multi-homing. The determinism provided by active-standby per interface is also required for certain QOS features to work. While using this mode, customers also expect minimized convergence during failures.

A new type of load-balancing mode, port-active load-balancing, is defined. This draft describes how the new load-balancing mode can be supported via EVPN. The new mode may also be referred to as per interface active/standby.

```
              +-----+
              | PE3 |
              +-----+
           +-----------+
           |  MPLS/IP  |
           |  CORE     |
           +-----------+
         +-----+   +-----+
         | PE1 |   | PE2 |
         +-----+   +-----+
            |         |
            I1       I2
              \     /
               \   /
               +---+
               |CE1|
               +---+
```

Figure 1: MC-LAG Topology

Figure 1 shows a MC-LAG multi-homing topology where PE1 and PE2 are part of the same redundancy group providing multi-homing to CE1 via interfaces I1 and I2. Interfaces I1 and I2 are members of a LAG running LACP protocol. The core, shown as IP or MPLS enabled, provides wide range of L2 and L3 services. MC-LAG multi-homing functionality is decoupled from those services in the core and it focuses on providing multi-homing to the CE. With per-port active/standby load-balancing, only one of the two interface I1 or I2 would be in forwarding, the other interface will be in standby. This also implies that all services on the active interface are in active mode and all services on the standby interface operate in standby mode.

### 1.1. 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.

## 2. Multi-Chassis Link Aggregation

When a CE is multi-homed to a set of PE nodes using the [IEEE.802.1AX_2014] Link Aggregation Control Protocol (LACP), the PEs must act as if they were a single LACP speaker for the Ethernet links to form and operate as a Link Aggregation Group (LAG). To achieve this, the PEs connected to the same multi-homed CE must synchronize LACP configuration and operational data among them. Interchassis Communication Protocol (ICCP) [RFC7275] has been used for that purpose. EVPN LAG simplifies greatly that solution. Along with the simplification come a few assumptions:

*a CE device connected to multi-homing PEs may have a single LAG with all its active links i.e. links in the LAG operate in all-active load-balancing mode.

*Same LACP parameters MUST be configured on peering PEs such as system id, port priority and port key.

Any discrepancies from this list are out of the scope of this document, as are mis-configuration and mis-wiring detection across peering PEs.

## 3. Port-active Load-balancing Procedure

Following steps describe the proposed procedure with EVPN LAG to support port-active load-balancing mode:

a. The Ethernet-Segment Identifier (ESI) MUST be assigned per access interface as described in [RFC7432], which may be auto derived or manually assigned. Access interface MAY be a Layer-2 or Layer-3 interface. The usage of ESI over Layer-3 interface is newly described in this document.

b. Ethernet-Segment (ES) MUST be configured in port-active load-balancing mode on peering PEs for specific access interface.

c. Peering PEs MAY exchange only Ethernet-Segment (ES) route (Route Type-4) when ESI is configured on a Layer-3 interface.

d. PEs in the redundancy group leverage the DF election defined in [RFC8584] to determine which PE keeps the port in active mode and which one(s) keep it in standby mode. While the DF election defined in [RFC8584] is per [ES, Ethernet Tag] granularity, for port-active mode of multi-homing, the DF election is done per <ES>. The details of this algorithm are described in Section 4.

e. DF router MUST keep corresponding access interface in up and forwarding active state for that Ethernet-Segment

f. Non-DF routers will by default implement a bidirectional blocking scheme for all traffic in line with [RFC7432] Single-Active blocking scheme, albeit across all VLANS.

* Non-DF routers MAY bring and keep peering access interface attached to it in operational down state.

* If the interface is running LACP protocol, then the non-DF PE MAY also set the LACP state to OOS (Out of Sync) as opposed to interface state down. This allows for better convergence on standby to active transition.

g. For EVPN-VPWS service, the usage of primary/backup bits of EVPN Layer-2 attributes extended community [RFC8214] is highly recommended to achieve better convergence.

## 4. Designated Forwarder Algorithm to Elect per Port-active PE

The ES routes, running in port-active load-balancing mode, are advertised with the new Port Mode Load-Balancing capability in the DF Election Extended Community defined in [RFC8584]. Moreover, the ES associated to the port leverages existing procedure of Single-Active, and signals Single-Active Multihomed site redundancy mode along with Ethernet-AD per-ES route (Section 7.5 of [RFC7432]). Finally the ESI-label based split-horizon procedures in Section 8.3 of [RFC7432] should be used to avoid transient echo'ed packets when Layer-2 circuits are involved.

The various algorithms for DF Election are discussed in Sections 4.2 to 4.5 for completeness, although the choice of algorithm in this solution doesn't affect complexity or performance as in other load-balancing modes.

### 4.1. Capability Flag

[RFC8584] defines a DF Election extended community, and a Bitmap field to encode "capabilities" to use with the DF election algorithm in the DF algorithm field. Bitmap (2 octets) is extended by the following value:

```
                     1 1 1 1 1 1
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|D|A|     |P|                   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 2: Amended Bitmap field in the DF Election Extended Community

**Bit 0:**

D bit or 'Don't Preempt' bit, as explained in [I-D.ietf-bess-evpn-pref-df].

**Bit 1:** AC-DF Capability (AC-Influenced DF election), as explained in [RFC8584].

**Bit 5:** (corresponds to Bit 29 of the DF Election Extended Community and it is defined by this document): 'Port Mode Load-Balancing' Capability (P bit hereafter), determines that the DF-Algorithm should be modified to consider the port ES only and not the Ethernet Tags.

## 4.2. Modulo-based Algorithm

The default DF Election algorithm, or modulus-based algorithm as in [RFC7432] and updated by [RFC8584], is used here, at the granularity of ES only. Given that ES-Import Route Target extended community may be auto-derived and directly inherits its auto-derived value from ESI bytes 1-6, many operators differentiate ESI primarily within these bytes. As a result, bytes 3-6 are used to determine the designated forwarder using Modulo-based DF assignment, achieving good entropy during Modulo calculation across ESIs:
Assuming a redundancy group of N PE nodes, the PE with ordinal i is the DF for an <EE> when (Es mod N) = i, where Es represents bytes 3-6 of that ESI.

## 4.3. HRW Algorithm

Highest Random Weight (HRW) algorithm defined in [RFC8584] MAY also be used and signaled, and modified to operate at the granularity of <ES> rather than per <ES, VLAN>.

Section 3.2 of [RFC8584] describes computing a 32 bit CRC over the concatenation of Ethernet Tag and ESI. For port-active load-balancing mode, the Ethernet Tag is simply removed from the CRC computation.

DF(Es) denotes the DF and BDF(Es) denote the BDF for the ESI es; Si is the IP address of PE i; and Weight is a function of Si, and Es.

1. DF(Es) = Si| Weight(Es, Si) >= Weight(Es, Sj), for all j. In the case of a tie, choose the PE whose IP address is numerically the least. Note that 0 <= i,j < number of PEs in the redundancy group.

2. BDF(Es) = Sk| Weight(Es, Si) >= Weight(Es, Sk), and Weight(Es, Sk) >= Weight(Es, Sj). In the case of a tie, choose the PE whose IP address is numerically the least.

Where:

- *DF(Es) is defined to be the address Si (index i) for which Weight(Es, Si) is the highest; 0 <= i < N-1.

- *BDF(Es) is defined as that PE with address Sk for which the computed Weight is the next highest after the Weight of the DF. j is the running index from 0 to N-1; i and k are selected values.

### 4.4. Preference-based DF Election

When the new capability 'Port-Mode' is signaled, the algorithm is modified to consider the port only and not any associated Ethernet Tags. Furthermore, the "port-based" capability MUST be compatible with the "Don't Preempt" bit. When an interface recovers, a peering PE signaling D-bit will enable non-revertive behaviour at the port level.

### 4.5. AC-Influenced DF Election

The AC-DF bit MUST be set to 0 when advertising Port Mode Load-Balancing capability (P=1). When an AC (sub-interface) goes down, it does not influence the DF election. The peer's Ethernet A-D per EVI is ignored in all Port Mode DF Election algorthms.

Upon receiving AC-DF bit set (A=1) from a remote PE, it MUST be ignored when performing Port-Mode DF Election.

## 5. Convergence considerations

To improve the convergence, upon failure and recovery, when port-active load-balancing mode is used, some advanced synchronization between peering PEs may be required. Port-active is challenging in a sense that the "standby" port is in down state. It takes some time to bring a "standby" port in up-state and settle the network. For IRB and L3 services, ARP / ND cache may be synchronized. Moreover, associated VRF tables may also be synchronized. For L2 services, MAC table synchronization may be considered.

Finally, for members of a LAG running LACP the ability to set the "standby" port in "out-of-sync" state a.k.a "warm-standby" can be leveraged.

### 5.1. Primary / Backup per Ethernet-Segment

The EVPN Layer 2 Attributes Control Flags extended community SHOULD be advertised in Ethernet A-D per ES route for fast convergence.

Only the P and B bits are relevant to this document, and only in the context of Ethernet A-D per ES routes:

- When advertised, the EVPN Layer 2 Attributes Control Flags extended community SHALL have only P or B bits set and all other bits and fields MUST be zero.
- A remote PE receiving the optional EVPN Layer 2 Attributes Control Flags extended community in Ethernet A-D per ES routes SHALL consider only P and B bits.

For EVPN Layer 2 Attributes Control Flags extended community sent and received in Ethernet A-D per EVI routes used in [RFC8214], [RFC7432] and [I-D.ietf-bess-evpn-vpws-fxc]:

- P and B bits received are overridden by "parent" bits on Ethernet A-D per ES above.
- Other fields and bits of the extended community are used according to the procedures of those documents.

### 5.2. Backward Compatibility

Implementations that comply with [RFC7432] or [RFC8214] only (i.e., implementations that predate this document) will not advertise the EVPN Layer 2 Attributes Control Flags extended community in Ethernet A-D per ES routes. That means that all remote PEs in the ES will not receive P and B bit per ES and will continue to receive and honour the P and B bits received in Ethernet A-D per EVI route(s). Similarly, an implementation that complies with [RFC7432] or [RFC8214] only and that receives an EVPN Layer 2 Attributes Control Flags extended community will ignore it and will continue to use the default path resolution algorithm.

## 6. Applicability

A common deployment is to provide L2 or L3 service on the PEs providing multi-homing. The services could be any L2 EVPN such as EVPN VPWS, EVPN [RFC7432], etc. L3 service could be in VPN context [RFC4364] or in global routing context. When a PE provides first hop routing, EVPN IRB could also be deployed on the PEs. The mechanism defined in this document is used between the PEs providing L2 and/or L3 services, when per interface single-active load-balancing is desired.

A possible alternate solution is the one described in this draft is MC-LAG with ICCP [RFC7275] active-standby redundancy. However, ICCP requires LDP to be enabled as a transport of ICCP messages. There are many scenarios where LDP is not required e.g. deployments with VXLAN or SRv6. The solution defined in this draft with EVPN does not mandate the need to use LDP or ICCP and is independent of the underlay encapsulation.

## 7. Overall Advantages

The use of port-active multi-homing brings the following benefits to EVPN networks:

a. Open standards based per interface single-active load-balancing mechanism that eliminates the need to run ICCP and LDP (e.g. they may be running VXLAN or SRv6 in the network).

b. Agnostic of underlay technology (MPLS, VXLAN, SRv6) and associated services (L2, L3, Bridging, E-LINE, etc).

c. Provides a way to enable deterministic QOS over MC-LAG attachment circuits.

d. Fully compliant with [RFC7432], does not require any new protocol enhancement to existing EVPN RFCs.

e. Can leverage various DF election algorithms e.g. modulo, HRW, etc.

f. Replaces legacy MC-LAG ICCP-based solution, and offers following additional benefits:

   - Efficiently supports 1+N redundancy mode (with EVPN using BGP RR) where as ICCP requires full mesh of LDP sessions among PEs in redundancy group.

   - Fast convergence with mass-withdraw is possible with EVPN, no equivalent in ICCP.

## 8. IANA Considerations

This document solicits the allocation of the following values:

- Bit 5 in the [RFC8584] DF Election Capabilities registry, with name "P" for Port Mode Load-Balancing.

## 9. Security Considerations

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

By introducing a new capability, a new requirement for unanimity (or lack thereof) between PEs is added. Without consensus on the new DF election procedures and Port Mode, the DF election algorithm falls back to the default DF election as provided in [RFC8584] and

[RFC7432]. This behavior could be exploited by an attacker that manages to modify the configuration of one PE in the ES so that the DF election algorithm and capabilities in all the PEs in the ES fall back to the default DF election. If that is the case, the PEs will be exposed to the same unfair load balancing, service disruption, and possibly black-holing or duplicate traffic mentioned in those documents and their security sections.

## 10. Acknowledgements

The authors thank Anoop Ghanwani for his comments and suggestions and Stephane Litkowski for his careful review.

## 11. References

### 11.1. Normative References

**[I-D.ietf-bess-evpn-pref-df]**
Rabadan, J., Sathappan, S., Przygienda, T., Lin, W., Drake, J., Sajassi, A., and satyamoh@cisco.com, "Preference-based EVPN DF Election", Work in Progress, Internet-Draft, draft-ietf-bess-evpn-pref-df-08, 23 September 2021, <https://www.ietf.org/archive/id/draft-ietf-bess-evpn-pref-df-08.txt>.

**[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>.

**[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>.

**[RFC8214]** Boutros, S., Sajassi, A., Salam, S., Drake, J., and J. Rabadan, "Virtual Private Wire Service Support in Ethernet VPN", RFC 8214, DOI 10.17487/RFC8214, August 2017, <https://www.rfc-editor.org/info/rfc8214>.

**[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>.

### 11.2. Informative References

**[I-D.ietf-bess-evpn-vpws-fxc]**
Sajassi, A., Brissette, P., Uttaro, J., Drake, J., Boutros, S., and J. Rabadan, "EVPN VPWS Flexible Cross-Connect Service", Work in Progress, Internet-Draft, draft-ietf-bess-evpn-vpws-fxc-05, 8 February 2022, <https://www.ietf.org/archive/id/draft-ietf-bess-evpn-vpws-fxc-05.txt>.

**[IEEE.802.1AX_2014]** IEEE, "IEEE Standard for Local and metropolitan area networks -- Link Aggregation", IEEE 802.1AX-2014, DOI 10.1109/IEEESTD.2014.7055197, 24 December 2014, <http://ieeexplore.ieee.org/servlet/opac?punumber=6997981>.

**[RFC4364]** Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, DOI 10.17487/RFC4364, February 2006, <https://www.rfc-editor.org/info/rfc4364>.

**[RFC7275]** Martini, L., Salam, S., Sajassi, A., Bocci, M., Matsushima, S., and T. Nadeau, "Inter-Chassis Communication Protocol for Layer 2 Virtual Private Network (L2VPN) Provider Edge (PE) Redundancy", RFC 7275, DOI 10.17487/RFC7275, June 2014, <https://www.rfc-editor.org/info/rfc7275>.

### Authors' Addresses

Patrice Brissette (editor)
Cisco Systems
Ottawa ON
Canada

Email: pbrisset@cisco.com

Ali Sajassi
Cisco Systems
United States of America

Email: sajassi@cisco.com

Luc Andre Burdet (editor)
Cisco Systems
Canada

Email: lburdet@cisco.com

Samir Thoria
Cisco Systems

United States of America

Email: sthoria@cisco.com

Bin Wen
Comcast
United States of America

Email: Bin_Wen@comcast.com

Edward Leyton
Verizon Wireless
United States of America

Email: edward.leyton@verizonwireless.com

Jorge Rabadan
Nokia
United States of America

Email: jorge.rabadan@nokia.com