Workgroup: BESS Working Group
Internet-Draft: draft-ietf-bess-evpn-mh-pa-10
Published: 4 March 2024
Intended Status: Standards Track
Expires: 5 September 2024
Authors: P. Brissette, Ed. — Cisco Systems
LA. Burdet, Ed. — Cisco Systems
B. Wen — Comcast
E. Leyton — Verizon Wireless
J. Rabadan — Nokia

EVPN Port-Active Redundancy Mode

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 redundancy modes. This document expands on existing redundancy mechanisms supported by EVPN and introduces a new Port-Active redundancy mode.

Status of This Memo

Copyright Notice

Table of Contents

1. Introduction
  1.1. Requirements Language
2. Multi-Chassis Link Aggregation (MC-LAG)
3. Port-Active Redundancy Mode
  3.1. Overall Advantages
  3.2. Port-Active Redundancy Procedures
4. Designated Forwarder Algorithm to Elect per Port-Active PE
  4.1. Capability Flag
  4.2. Modulo-based Algorithm
  4.3. HRW Algorithm
  4.4. Preference-based DF Election
  4.5. AC-Influenced DF Election
5. Convergence considerations
  5.1. Primary / Backup per Ethernet-Segment
  5.2. Backward Compatibility
6. Applicability
7. IANA Considerations
8. Security Considerations
9. Acknowledgements
10. Contributors
11. References
  11.1. Normative References
  11.2. Informative References
Authors' Addresses

1. Introduction

EVPN [RFC7432] defines the All-Active and Single-Active redundancy modes. All-Active redundancy provides per-flow load-balancing for multi-homing, and Single-active redundancy provides service carving where only one of the PEs in a 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 an active/standby multi-homing at the interface level is useful and required. The main consideration for this new 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 multi-homing at the interface level is also required for certain QoS features to work. While using this mode, customers also expect fast convergence during failure and recovery.

This document defines the Port-Active redundancy mode as a new type of multi-homing in EVPN and describes how this new mode operates and is to be supported via EVPN.

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 (MC-LAG)

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 between them. Interchassis Communication Protocol (ICCP) [RFC7275] has historically been used to achieve this. EVPN in [RFC7432] describes the case where a CE is multihomed to multiple PE nodes, using a LAG as a means to greatly simplify the procedure. The simplification, however, comes with a few assumptions:

* a CE device connected to EVPN multi-homing PEs MUST have a single LAG with all its links connected to the EVPN multi-homing PEs in a redundancy group.

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

This document relies on proper LAG operation as in [RFC7432]. Discrepancies from the list above are out of the scope of this document, as are LAG misconfiguration and miswiring detection across peering PEs.

```
     +-----+
     | 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. The core, shown as IP or MPLS enabled, provides a 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. In Port-Active redundancy mode, only one of the two interfaces I1 or I2 would be in forwarding and 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.

3. Port-Active Redundancy Mode

3.1. Overall Advantages

The use of Port-Active redundancy brings the following benefits to EVPN networks:

a. Open standards-based active/standby redundancy at the interface level which 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 Designated Forwarder (DF) election algorithms e.g. modulo, HRW, etc.

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

   * Efficiently supports 1+N redundancy mode (with EVPN using BGP RR) whereas ICCP requires a full mesh of LDP sessions among PEs in the redundancy group.

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

3.2. Port-Active Redundancy Procedures

The following steps describe the proposed procedure with EVPN LAG to support Port-Active redundancy 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. The access interface MAY be a Layer-2 or Layer-3 interface. The use of ESI over a Layer-3 interface is newly described in this document.

b. Ethernet-Segment (ES) MUST be configured in Port-Active redundancy mode on peering PEs for specific access interface.

c. When ESI is configured on a Layer-3 interface, the Ethernet-Segment (ES) route (Route Type-4) may be the only route exchanged by PEs in the redundancy group.

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, the DF election is done per [ES] in Port-Active redundancy mode. 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 SHOULD implement a bidirectional blocking scheme for all traffic comparable to [RFC7432] Single-Active blocking scheme, albeit across all VLANs.

   * Non-DF routers MAY bring and keep peering access interface attached to it in an 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 an interface down state. This allows for better convergence on standby to active transition.

g. The primary/backup bits of EVPN Layer 2 Attributes Extended Community [RFC8214] SHOULD be used to achieve better convergence as decribed in section Section 5.1.

4. Designated Forwarder Algorithm to Elect per Port-Active PE

The ES routes, running in Port-Active redundancy mode, are advertised with the new Port Mode Load-Balancing capability bit in the DF Election Extended Community defined in [RFC8584]. Moreover, the ES associated with the port leverages the 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 even though the choice of algorithm in this solution doesn't affect complexity or performance as in other redundancy modes.

4.1. Capability Flag

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

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].

```
                     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

This document defines the following value and extends the Bitmap field:

Bit 5:

Port Mode Designated Forwarder Election (P bit hereafter), determines that the DF Election 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 <ES> 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 redundancy mode, the Ethernet Tag is simply omitted from the CRC computation and all references to (V, Es) are replaced by (Es), as repeated and summarised below.

DF(Es) denotes the DF and BDF(Es) denote the BDF for the Ethernet Segment 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 preference-based DF Election algirithm in [I-D.ietf-bess-evpn-pref-df] is modified to consider the port only and not any associated Ethernet Tags. Furthermore, the Port Mode capability MUST be compatible with the 'Don't Preempt' bit. When an interface recovers, a peering PE signaling D bit will enable non-revertive behavior at the port level.

4.5. AC-Influenced DF Election

The AC-DF bit defined in [RFC8584] MUST be set to 0 when advertising Port Mode Designated Forwarder Election 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 algorithms.

Upon receiving the 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 the Port-Active redundancy mode is used, some advanced synchronization between peering PEs may be required. Port-Active is challenging in the sense that the "standby" port may be in a down state. It takes some time to bring a "standby" port to an 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 Extended Community ("L2-Attr") defined in [RFC8214] SHOULD be advertised in the Ethernet A-D per ES route for fast convergence.

Only the P and B bits of the Control Flags field in the L2-Attr Extended Community are relevant to this document, and only in the context of Ethernet A-D per ES routes:

* When advertised, the L2-Attr Extended Community SHALL have only P or B bits in the Control Flags field set, and all other bits and fields MUST be zero.

* A remote PE receiving the optional L2-Attr Extended Community in Ethernet A-D per ES routes SHALL consider only P and B bits and ignore other values.

For L2-Attr 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 SHOULD be considered overridden by "parent" bits when advertised in the Ethernet A-D per ES.

* 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 specification) will not advertise the EVPN Layer 2 Attributes 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 L2-Attr Extended Community in Ethernet A-D per ES routes will ignore it and continue to use the default path resolution algorithm:

* The remote ESI Label Extended Community ([RFC7432]) signals Single-Active (Section 4)

* the remote MAC and/or Ethernet A-D per EVI routes are unchanged, and since the L2-Attr Extended Community in Ethernet A-D per ES route is ignored, the P and B bits in the L2-Attr Extended Community in Ethernet A-D per EVI routes are used.

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 a VPN context [RFC4364] or in a 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 active/standby redundancy at the interface level is desired.

A possible alternate solution to the one described in this document 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 document with EVPN does not mandate the need to use LDP or ICCP and is independent of the underlay encapsulation.

7. IANA Considerations

This document solicits the allocation of the following values from the "BGP Extended Communities" registry group :

*Bit 5 in the [RFC8584] DF Election Capabilities registry, "P bit - Port Mode Designated Forwarder Election".

8. 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.

9. Acknowledgements

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

10. Contributors

In addition to the authors listed on the front page, the following coauthors have also contributed to this document:

Ali Sajassi
Cisco Systems
United States of America

Email: sajassi@cisco.com

Samir Thoria
Cisco Systems
United States of America

Email: sthoria@cisco.com

11. References

11.1. Normative References

[I-D.ietf-bess-evpn-pref-df] Rabadan, J., Sathappan, S., Lin, W., Drake, J., and A. Sajassi, "Preference-based EVPN DF Election", Work in Progress, Internet-Draft, draft-ietf-bess-evpn-pref-df-13, 9 October 2023, <https://datatracker.ietf.org/doc/html/draft-ietf-bess-evpn-pref-df-13>.

[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>.

[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-08, 24 October 2022, <https://datatracker.ietf.org/doc/html/draft-ietf-bess-evpn-vpws-fxc-08>.

[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

Luc Andre Burdet (editor)
Cisco Systems
Canada

Email: lburdet@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