BESS Working Group
Internet-Draft
Intended status: Standards Track
Expires: August 26, 2021

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

February 22, 2021

EVPN multi-homing port-active load-balancing
draft-ietf-bess-evpn-mh-pa-01

Abstract

The Multi-Chassis Link Aggregation Group (MC-LAG) technology enables the establishment of 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. EVPN standard 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 of port-active load-balancing mode. In the current document, port-active load-balancing mode is also referred to as per interface active/standby.

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 August 26, 2021.

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 redundancy is useful

and required. The main consideration for this mode of redundancy 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 term of load-balancing mode, port-active load- balancing is then defined.

This draft describes how that new redundancy mode can be supported via EVPN

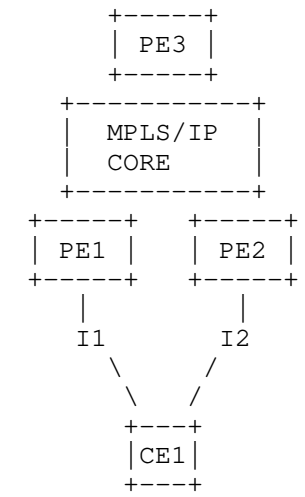

Figure 1: MC-LAG Topology

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 Bundle-Ethernet interfaces 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 CE. With per-port active/ standby redundancy, 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", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119].

2. Multi-Chassis Ethernet Bundles

When a CE is multi-homed to a set of PE nodes using the [802.1AX] Link Aggregation Control Protocol (LACP), the PEs must act as if they were a single LACP speaker for the Ethernet links to form a bundle, 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 Communicated-based Protocol (ICCP) has been used for that purpose. EVPN LAG simplifies greatly that solution. Along with the simplification comes few assumptions:

- CE device connected to Multi-homing PEs may has a single LAG with all its active links i.e. Links in the Ethernet Bundle 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 is left for future study. Furthermore, mis-configuration and mis-wiring detection across peering PEs are also left for further study.

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 Layer3 interface. The usage of ESI over L3 interfce is newly described in this document.

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

c. Peering PEs MAY exchange only Ethernet-Segment route (Route Type-4) when ESI is configured on a Layer3 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 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 Layer2 attributes extended community [RFC8214] is highly recommended to achieve better convergence.

4. Algorithm to elect per port-active PE

The ES routes, running in port-active load-balancing mode, are advertised with a new capability in the DF Election Extended Community as defined in [RFC8584]. Moreover, the ES associated to the port leverages existing procedure of single-active, and signals single-active bit along with Ethernet-AD per-ES route. Finally, as in [RFC7432], the ESI-label based split-horizon procedures should be used to avoid transient echo'ed packets when L2 circuits are involved.

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:

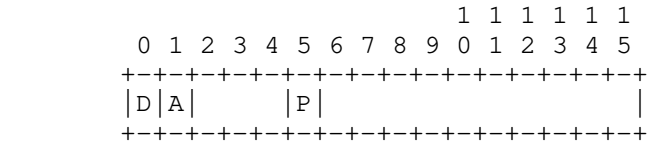

Figure 2: Amended Bitmap field in the DF Election Extended Community

Bit 0: 'Don't Preempt' bit, as explained in [PREF-DF].

Bit 1: AC-Influenced DF Election, as explained in [RFC8584].

Bit 5: (corresponds to Bit 25 of the DF Election Extended Community and it is defined by this document): P bit or 'Port Mode' bit (P hereafter), determines that the DF-Algorithm should be modified to consider the port only and not the Ethernet Tags.

4.2. Modulo-based Designated Forwarder 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 the fact, ES-Import RT community inherits from ESI only byte 1-6, many deployments differentiate ESI within these bytes only. For Modulo calculation, bytes [3-6] are used to determine the designated forwarder using Modulo-based DF assignment.

4.3. HRW Algorithm

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

[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.

4.4. Preferred-DF Algorithm

When the new capability 'Port-Mode' is signalled, 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 'DP' capability (for non-revertive). The AC-DF bit MUST be set to zero. When an AC (sub-interface) goes down, it does not influence the 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 Bundle-Ethernet interface where LACP is running the ability to set the "standby" port in "out-of-sync" state aka "warm-standby" can be leveraged.

5.1. Primary / Backup per Ethernet-Segment

The L2 Info Extended Community MAY be advertised in Ethernet A-D per ES routes for fast convergence. Only the P and B bits are relevant to this specification. When advertised, the L2 Info Extended Community SHALL have only P or B bits set and all other bits must be zero. MTU must also be zero. Remote PE receiving optional L2 Info Extended Community on Ethernet A-D per ES routes SHALL consider only P and B bits. P and B bits received on Ethernet A-D per EVI routes per [RFC8214] are overridden.

5.2. Backward Compatibility

Implementations that comply with [RFC7432] or [RFC8214] only (i.e., implementations that predate this specification) will not advertise the L2 Info 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 Ethernet A-D per EVI routes. Similarly, an implementation that complies with [RFC7432] or [RFC8214] only and that receives a L2 Info 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 draft is used between the PEs providing the L2 and/or L3 service, when the requirement is to use per port active.

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 redundancy mechanism that eliminates the need to run ICCP and LDP.

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:

g.

* 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

h. Customers want per interface single-active redundancy, but don't want to enable LDP (e.g. they may be running VXLAN or SRv6 in the network). Currently there is no alternative to this.

8. Security Considerations

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

9. IANA Considerations

This document solicits the allocation of the following values:

o Bit 5 in the [RFC8584] DF Election Capabilities registry, with name "P" (port mode load-balancing) Capability" for port-active ES.

10. References

10.1. 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>.

[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>.

[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>.

10.2. Informative References

[PREF-DF] Rabadan, J., "Preference-based EVPN DF Election", 2020.

[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>.

[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>.

Authors' Addresses

Patrice Brissette (editor)
Cisco Systems
Ottawa, ON
Canada

Email: pbrisset@cisco.com

Ali Sajassi
Cisco Systems
USA

Email: sajassi@cisco.com

Bin Wen
Comcast
USA

Email: Bin_Wen@comcast.com

Edward Leyton
Verizon Wireless
USA

Email: edward.leyton@verizonwireless.com

Jorge Rabadan
Nokia
USA

Email: jorge.rabadan@nokia.com

Luc Andre Burdet
Cisco Systems
Canada

Email: lburdet@cisco.com

Samir Thoria
Cisco Systems
USA

Email: sthoria@cisco.com