INTERNET-DRAFT   Patrice Brissette
Intended Status: Proposed Standard   Samir Thoria
                                                           Cisco Systems

Expires: May 3, 2018   October 30, 2017

EVPN multi-homing port-active load-balancing
draft-brissette-bess-evpn-mh-pa-00

Abstract

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

Table of Contents

1 Introduction

EVPN, as per [RFC 7432], currently 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. Main consideration for this mode of redundancy is the determinism of traffic forwarding through 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 customer 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.

```
                 +-----+
                 | 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. The core shown as IP or MPLS enabled, can provide wide range of L2 and L3 services. MC-LAG multi-homing functionality is decoupled from the services in the core and is focused in 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. When EVPN is used to provide MC-LAG functionality, we refer to it as EVLAG in this draft.

1.1 Terminology

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 RFC 2119 [RFC2119].

2. Port-active load-balancing procedure

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

1- ESI is assigned per access interface as described in [RFC 7432],

which may be auto derived or manually assigned.
2- Ethernet-Segment is configured in per-port load-balancing mode on peering PEs for specific interface
3- Peering PEs exchange only Ethernet-Segment route (Route Type-4). No other EVPN routes are used for redundancy.
4- PEs in the redundancy group leverages DF election defined in [draft-ietf-bess-evpn-df-election] to determine which PE will keep the port in active mode and which one(s) will keep it in standby mode. While the DF election defined in draft-ietf-bess-evpn-df-lection is per <ES, VLAN> 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.
5- DF router keeps corresponding access interface in up and forwarding active state for that Ethernet-Segment
6- Non-DF router brings and keeps the 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.

3. Algorithm to elect per port-active PE

The default mode of Designated Forwarder Election algorithm remains as per [RFC7432] at the granularity of <ES>.

However, Highest Random Weight (HRW) algorithm defined in [draft-ietf-bess-evpn-df-election] is leveraged, and modified to operate at the granularity of <ES> rather than per <ES, VLAN>.

Let Active(ESI) denote the PE that will be the active PE for port with Ethernet segment identifier - ESI. The other PEs in the redundancy group will be standby PE(s) for the same port (ES). Ai is the address of the PEi and weight() is a pseudorandom function of ESi and Ai, Wrand() function defined in [draft-ietf-bess-evpn-df-election] is used as the Weight() function.

Active(ESI) = PEi: if Weight(ESI, Ai) >= Weight(ESI, Aj), for all j, 0 <= I,j <= Number of PEs in the redundancy group. In case of a tie, choose the PE whose IP address is numerically the least.

4. Applicability

A common deployment is to provide L2 or L3 service on the PEs providing multi-homing. The L2 services could include EVPN VPWS or EVPN [RFC 7432]. L3 service could be in VPN context [RFC 4364] or in global routing context. When the PE is providing 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 or L3 service,

when the requirement is to use per port active.

A possible alternate solution for the one described in this draft is MC-LAG with ICCP [RFC 7275] 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 - for example 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 overlay encapsulation.

5. Advantages

There are many advantages in EVLAG to support port-active load-balancing mode. Here is a non-exhaustive list:

- Open standards based per interface single-active redundancy mechanism that eliminates the need to run ICCP and LDP.

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

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

- Fully compliant with RFC-7432, does not require any new protocol enhancement to existing EVPN RFCs.

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

- 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

- 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.

6 Security Considerations

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

7 IANA Considerations

There are no new IANA considerations in this document.

8 References

8.1 Normative References

[RFC4684] Marques, P., Bonica, R., Fang, L., Martini, L., Raszuk, R., Patel, K., and J. Guichard, "Constrained Route Distribution for Border Gateway Protocol/MultiProtocol Label Switching (BGP/MPLS) Internet Protocol (IP) Virtual Private Networks (VPNs)", RFC 4684, DOI 10.17487/RFC4684, November 2006, <http://www.rfc-editor.org/info/rfc4684>.

[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, <http://www.rfc-editor.org/info/rfc7432>.

8.2 Informative References

[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, <http://www.rfc-editor.org/info/rfc7275>.

Authors' Addresses

Patrice Brissette
Cisco Systems
EMail: pbrisset@cisco.com

Samir Thoria
Cisco Systems
EMail: sthoria@cisco.com