INTERNET-DRAFT
Intended Status: Proposed Standard

Patrice Brissette
Samir Thoria
Ali Sajassi
Cisco Systems

Expires: April 25, 2019

October 22, 2018

EVPN multi-homing port-active load-balancing
draft-brissette-bess-evpn-mh-pa-02

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. 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 draft, port-active load-balancing mode is also referred to as per interface active/standby.

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. 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, 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.

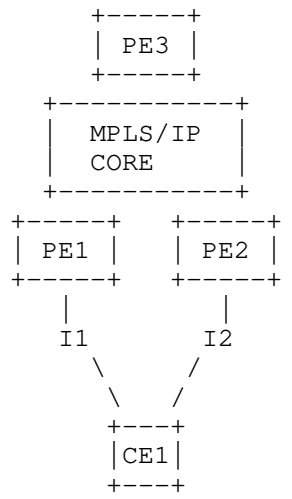

Figure 1. MC-LAG topology

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. 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. 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. ICCP-based protocol has been used for that purpose. EVLAG simplifies greatly that solution. Along with the simplification comes few assumptions:

- Links in the Ethernet Bundle MUST operate in all-active load-balancing mode

- Same LACP parameters MUST be configured on peering PEs such as system id, port priority, etc.

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 EVLAG to support port-active load-balancing mode:

1- 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.

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

3- Peering PEs MAY exchange only Ethernet-Segment route (Route Type-4)

4- PEs in the redundancy group leverages DF election defined in [draft-ietf-bess-evpn-df-election-framework] 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 [draft-ietf-bess-evpn-df-election-framework] 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 MUST keep corresponding access interface in up and forwarding active state for that Ethernet-Segment

6- 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.

4. Algorithm to elect per port-active PE

The default DF Election algorithm, or modulus-based algorithm as in [RFC7432], is used here also, at the granularity of <ES> only. For Modulo calculation, byte 10 of the ESI is used.

Highest Random Weight (HRW) algorithm defined in [draft-ietf-bess-evpn-df-election-framework] MAY also be used and signaled, 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-framework] 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.

5. Port-active over Integrated Routing-Bridging Interface

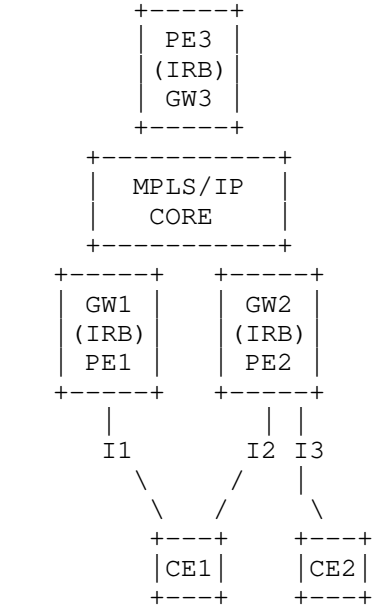

Figure 2. EVPN-IRB Port-active load-balancing

Figure 2 shows a simple network where EVPN-IRB is used for inter-subnet connectivity. IRB interfaces on PE1 and PE2 are configured in anycast gateway (same MAC, same IP). CE1 device is multi-homed to both PE1 and PE2. The Ethernet-segment load-balancing mode, of the connected CE1 to peering PEs, can be of any type e.g. all-active, single-active or port-active. CE2 device is connected to a single PE (PE2). It operates as single-homed device via an orphan port I3. Finally, port-active load-balancing is apply to IRB interface on peering PEs (PE1 and PE2). Manual Ethernet-Segment Identifier is assigned per IRB interface. ESI auto-generation is also possible based on the IRB anycast IP address.

DF election is performed between peering PE over IRB interface (per ESI/EVI). Designed forwarder (DF) IRB interface remains in up state. Non-designated forwarder (NDF) IRB interface may goes in down state. Furthermore, if all access interfaces connected to an IRB interface are down state (failure or admin) OR in blocked forward state(NDF), IRB interface is brought down. For example, interface I3 fails at the same time than interface I2 (in single-active load-balancing mode) is in blocked forwarding state.

In the example where IRB on PE2 is NDF, all L3 traffic coming from

PE3 is going via PE1. An IRB interface in down state doesn't attract traffic from core side. CE2 device reachability is done via an L2 subnet stretch between PE1 and PE2. Therefore L3 traffic coming from PE3 destinated to CE2 goes via GW1 first, then via an L2 connection to PE2 and finally via interface I3 to CE2 device.

There are many reasons of configuring port-active load-balancing mode over IRB interface:
- Ease replacement of legacy technology such VRRP / HSRP

- Better scalability than legacy protocols

- Traffic predictability

- Optimal routing and entirely independent of load-balancing mode configured on any access interfaces

6. 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 / MLD cache may be synchronized. Moreover, associated VRF tables may also be synchronized. For L2 services, MAC table synchronization may be considered. Finally, using bundle-Ethernet interface, where LACP is running, is usually a smart thing since it provides the ability to set the "standby" port in "out-of-sync" state aka "warm-standby".

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 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 overlay encapsulation.

7. Overall 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, E-LINE, 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.

8 Security Considerations

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

9 IANA Considerations

There are no new IANA considerations in this document.

10. Acknowledgements

Authors would like to thank Luc Andre Burdet for valuable reviews and inputs.

11 References

11.1 Normative References

[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>.

[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>.

11.2 Informative 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>.

[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>.

Authors' Addresses

Patrice Brissette

Cisco Systems
EMail: pbrisset@cisco.com

Samir Thoria
Cisco Systems
EMail: sthoria@cisco.com

Ali Sajassi
Cisco Systems
EMail: sajassi@cisco.com