BESS Workgroup
INTERNET-DRAFT
Intended Status: Standards Track

Ali Sajassi
Samer Salam
Dennis Cai
Cisco

Sami Boutros
VmWare

John Drake
Juniper

Luay Jalil
Verizon

Expires: October 21, 2016 March 21, 2016

Multi-homed L3VPN Service with Single IP peer to CE
draft-sajassi-bess-evpn-l3vpn-multihoming-01

Abstract

This document describes how EVPN can be used to offer a multi-homed L3VPN service leveraging EVPN Layer 2 access redundancy. The solution offers single IP peering to the Customer Edge (CE) nodes, rapid failure detection, minimal fail-over time and make-before-break paradigm for maintenance.

Status of this Memo

Copyright and License Notice

Table of Contents

1 Introduction

[RFC7432] defines EVPN, a solution for multipoint Layer 2 Virtual Private Network (L2VPN) services, with advanced multi-homing capabilities, using BGP for distributing customer/client MAC address reachability information over the core MPLS/IP network. [EVPN-IRB] and [EVPN-PREFIX] discuss how EVPN can be used to support inter-subnet forwarding among hosts across different IP subnets, while maintaining the redundancy capabilities of the original solution.

In this document, we discuss how EVPN can be used to offer a multi-homed L3VPN service leveraging its Layer 2 access redundancy. The solution offers single IP peering to the Customer Edge (CE) nodes, rapid failure detection, minimal fail-over time and make-before-break paradigm for maintenance.

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 Requirements

The network topology in question comprises of three domains: the customer network, the MPLS access network and the MPLS core network, as shown in the figure below.

```
     Customer  MPLS Access           MPLS        MPLS Access Customer
      Network    Network          Core Network      Network    Network
                                 +--------------+
                +-----------+    |              | +-------------+
                |+----+     |   +----+          | |             |
          +--+  ||    |---------|    |          | |             |
          |CE|---|APE1|------\  |SPE1|          | |             |
          +--+  |+----+     | \/+----+          | |             |
                |+----+     | / +----+        +----+       +---+|
          +--+  ||    |------/ \|    |        |    |       |APE|| +--+
          |CE|---|APE2|---------|SPE2|        |SPEr|-------|   |--|CE|
          +--+  |+----+     |   +----+        +----+       +---+| +--+
                |           |     |             | |             |
                +-----------+     |             | +-------------+
                                  +-------------+
```

Figure 1: Network Topology

The customer network connects via Customer Edge (CE) nodes to the

MPLS Access Network. The MPLS Access Network includes Access PEs (A-PEs) and MPLS P nodes (not shown for simplicity). The A-PEs provide a Virtual Private Wire Service (VPWS) to the connected CEs using Ethernet over MPLS (EoMPLS) pseudowires per [RFC5462]. The access pseudowires terminate on the service PEs (S-PE1, S-PE2,..., S-PEr). The Service PEs (S-PEs) provide inter-subnet forwarding between the CEs, i.e. L3VPN service between them. To provide redundancy, pseudowires from a given A-PE can terminate on two or more S-PEs forming a Redundancy Group. This provide multi-homed interconnect of A-PEs to S-PEs.

The solution MUST support the following requirements:

- The S-PEs in a redundancy group must provide single-active redundancy to the CEs, i.e. only one S-PE is actively forwarding traffic at any given point of time.

- The SPEs in a redundancy group must appear as a single IP peer to the CE, and a single eBGP session will be established between a given CE and its associated S-PEs.

- In the case of S-PE failure, pseudowire failure or S-PE isolation from access network, the fail-over time should be minimized by optimizing both the backup pseudowire establishment as well as the BGP convergence time. This reduces the amount of traffic loss as the active path reroutes to one of the backup S-PEs.

- The active S-PE must be able to quickly detect pseudowire failures or its isolation from the access MPLS network by means of a proactive monitoring mechanism.

- For system maintenance, it should be possible to support a make-before-break paradigm, where the backup path is in warm standby state before a given active S-PE is taken offline for service.

3 Challenges with L3VPN Multi-homing

The requirements depicted in section 2 above, especially the requirement to maintain a single eBGP session between the CE and the S-PEs, introduce challenges for standard L3VPN multi-homing solutions. In particular, the BGP prefix independent convergence (PIC) solution [BGP-PIC] cannot be used here because the backup S-PEs have no means of learning the IP prefixes from the CE: recall that the CE will only have an active eBGP session with the active S-PE. As a result, when the primary S-PE fails, the backup S-PE will have no alternate paths to the prefixes advertised by the CE. Therefore, with BGP PIC it is not possible to address the fast fail-over requirement.

4 Solution

4.1 Using Pseudowires in Access Network

The solution involves running EVPN on the S-PEs in single-active redundancy mode albeit for inter-subnet forwarding (i.e. Layer 3 forwarding). All pseudowires associated with a given CE are considered collectively as a Virtual Ethernet Segment (vES) [Virtual-ES] from the EVPN PEs perspective.

In the MPLS access network, pseudowire redundancy mechanisms are used [RFC6718][RFC6870] in either the Independent mode or the Master/Slave mode, with the S-PEs acting as the Master. The EVPN Designated Forwarder (DF) election mechanism is used to identify the active and standby S-PEs, and the pseudowire Preferential Forwarding Status Bit [RFC6870], for the access pseudowires, is derived from the outcome of the DF election, as follows:

- The S-PE that is elected as DF for a given vES MUST advertise Active in the Preferential Forwarding Status bit over the pseudowire corresponding to the vES.

- The SPE that is elected as non-DF for a given vES MUST advertise Standby in the Preferential Forwarding Status bit over the pseudowire corresponding to the vES.

On the S-PEs, the pseudowires from the Access PEs are terminated onto VRFs, such that all pseudowires within a given redundancy set terminate on a single IP endpoint on the S-PEs. To achieve this, the S-PEs in a given Redundancy Group are configured with the same Anycast IP and MAC addresses on the virtual (sub)interface corresponding to the VRF termination point.

Since the S-PEs are running in EVPN single-active redundancy mode, the S-PEs would advertise an Ethernet AD route per vES with the single-active flag set per [RFC7432]. Furthermore, the DF PE sets the Primary bit in the L2 extended community and the backup PE set the Backup bit in that extended community. Since only the DF S-PE has its access pseudowire in Active state, only that device would establish an eBGP session with the CE and receive control and data traffic. The DF S-PE advertises host prefixes that it receives, from the CE over the eBGP session, to other PEs in the EVI using EVPN route type-5, with the proper ESI set. Remote PEs learn the host prefixes and associate them with the ESI, using the advertising PE as the next-hop for forwarding.

Other S-PEs in the same Redundancy Group as the advertising PE will receive the same EVPN route type-5 advertisement, and will recognize

the associated ESI as a locally attached vES. This information will be used in the case of failure to provide a backup path to the CE. In other words, the S-PEs in the same Redundancy Group, use EVPN Aliasing procedure to synchronzie their IP-VRFs among themselves. It is worth noting here that the S-PEs in the Redundancy Group will have their ARP caches synchronized through the EVPN route type-2 advertisements from the DF PE.

4.2 Using EVPN-VPWS in Access Network

[EVPN-VPWS] can be used instead of pseudo wires in the MPLS access network, in that case all EVPN-VPWS service instances associated with a given CE are considered collectively as a Virtual Ethernet Segment (vES) [Virtual-ES].

The elected DF S-PE MUST set the Primary bit in the L2 attributes extended community associated with the EVPN-VPWS service instance Ethernet A-D route, corresponding to the vES. The non-DF S-PEs MUST set the Backup bit in the L2 attributes extended community associated with the EVPN-VPWS service instance Ethernet A-D route, corresponding to the vES.

Just as with pseudowires described in previous section, only the DF S-PE has its access EVPN-VPWS service instance in Active state, and thus establishes an eBGP session with the CE and receive control and data traffic. Just as before, the DF S-PE advertises host prefixes that it receives, from the CE over the eBGP session, to other PEs in the EVI using EVPN route type-5, with the proper ESI set. Remote PEs learn the host prefixes and associate them with the ESI, using the advertising PE as the next-hop for forwarding.

5 Failure Scenarios

5.1 Pseudowire Failure

The active (DF) S-PE can proactively monitor the health of the primary pseudowire by using a pseudowire OAM mechanism such as VCCV-BFD. As such, the S-PE can detect the failure of the primary pseudowire, and react by withdrawing both the Ethernet Segment route as well as the Ethernet A-D route associated with the vES. Note that the S-PE advertises the Ethernet A-D route per vES granularity as well as the Ethernet A-D per EVI. The withdrawal of the Ethernet Segment route serves as an indication to the backup S-PE to go active (i.e. act as a backup DF), and activate its pseudowires to the Access PE. The withdrawal of the Ethernet A-D route triggers a "mass withdraw" on the remote PEs: these PEs adjust their next-hop associated with the prefixes that were originally advertised by the

failed PE to point to the "backup path" per [RFC7432]. This provides relatively fast convergence because only a single message per Ethernet Segment is required for the remote PEs to switch over to the backup path irrespective of how many prefixes were learnt from the CE over the pseudowire. Also, note that no synchronization of VRF or ARP tables is required between the primary S-PE and its backup S-PE during the fail-over, because these tables were populated ahead of time during the original EVPN route advertisements.

As a result of the pseudowire failure, the eBGP session between the CE and the original DF PE will time out. This will cause said S-PE to start a timer in order to defer withdrawing the EVPN type-5 and type-2 routes that it had advertised for the prefixes learnt over the session from the CE. As the backup pseudowire to the backup DF PE goes active, the eBGP session will be re-established by the CE with the backup PE. Since both PEs share the same Anycast IP and MAC addresses, the CE does not recognize that it is in communication with a different PE.

To minimize disruption in data forwarding on the CE and the backup PE, the non-stop forwarding feature such as BGP Graceful Restart is used. Since the end-point IP address has not changed, this eBGP session handover between the primary S-PE and the backup S-PE, looks like a eBGP session flap with respect to the CE. Thus, the CE continues its packet forwarding operation in data-plane while synchronizing its control-plane with the backup S-PE.

5.2 EVPN VPWS Service Instance Failure

The failure scenario for an EVPN VPWS in similar to PW failure scenario described in the previous section. The failure detection of an EVPN service instance can be performed via OAM mechanisms such as VCCV-BFD and upon such failure detection, the switch over procedure to the backup S-PE is the same as the one described above.

5.3 PE Node Failure

In the case of PE node failure, the operation is similar to the steps described above, albeit that EVPN route withdrawals are performed by the Route Reflector instead of the PE.

6 Security Considerations

TBD.

7 IANA Considerations

TBD

8 References

8.1 Normative References

[RFC7432] Sajassi et al., "Ethernet VPN", RFC 7432, February 2015.

[EVPN-IRB] Sajassi et al., "Integrated Routing and Bridging in EVPN", draft-ietf-bess-evpn-inter-subnet-forwarding-00, work in progress, November 2014.

[EVPN-PREFIX] Rabadan et al., "IP Prefix Advertisement in EVPN", draft-ietf-bess-evpn-prefix-advertisement-02, work in progress, September 2015.

[RFC6718] Muley P., et al., "Pseudowire Redundancy", RFC 6718, August 2012.

[RFC6870] Muley P., et al., "Pseudowire Preferential Forwarding Status Bit", RFC 6870, February 2013.

8.2 Informative References

[BGP-PIC] Bashandy A. et al., "BGP Prefix Independent Convergence", draft-rtgwg-bgp-pic-02.txt, work in progress, October 2013.

Authors' Addresses

Ali Sajassi
Cisco
EMail: sajassi@cisco.com

Samer Salam
Cisco
EMail: ssalam@cisco.com

Dennis Cai
Cisco
EMail: dcai@cisco.com

John Drake
Juniper
EMail: jdrake@juniper.net

Luay Jalil
Verizon
EMail: luayjalil@gmail.com

Sami Boutros
VmWare
EMail: boutros.sami@gmail.com