Internet Working Group
INTERNET-DRAFT
Category: Informational

A. Sajassi
Cisco

R. Aggarwal
Arktan

J. Uttaro
AT&T

N. Bitar
Verizon

W. Henderickx
Alcatel-Lucent

Aldrin Isaac
Bloomberg

Expires: January 15, 2014 July 15, 2013

Requirements for Ethernet VPN (EVPN)
draft-ietf-l2vpn-evpn-req-04.txt

Status of this Memo

This Internet-Draft is submitted to IETF in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

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."

The list of current Internet-Drafts can be accessed at http://www.ietf.org/1id-abstracts.html

The list of Internet-Draft Shadow Directories can be accessed at http://www.ietf.org/shadow.html

Copyright and License Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents

(http://trustee.ietf.org/license-info) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Abstract

The widespread adoption of Ethernet L2VPN services and the advent of new applications for the technology (e.g., data center interconnect) have culminated in a new set of requirements that are not readily addressable by the current Virtual Private LAN Service (VPLS) solution. In particular, multi-homing with all-active forwarding is not supported and there's no existing solution to leverage Multipoint-to-Multipoint (MP2MP) LSPs for optimizing the delivery of multi-destination frames. Furthermore, the provisioning of VPLS, even in the context of BGP-based auto-discovery, requires network operators to specify various network parameters on top of the access configuration. This document specifies the requirements for an Ethernet VPN (EVPN) solution which addresses the above issues.

Table of Contents

1. Specification of requirements

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. Terminology

AS: Autonomous System
CE: Customer Edge
E-Tree: Ethernet tree
MAC address: Media Access Control address - simply referred to as MAC
LSP: Label Switched Path
PE: Provider Edge
MP2MP: Multipoint to Multipoint
VPLS: Virtual Private LAN Service

3. Introduction

VPLS, as defined in [RFC4664][RFC4761][RFC4762], is a proven and widely deployed technology. However, the existing solution has a number of limitations when it comes to redundancy, multicast optimization and provisioning simplicity. Furthermore, new applications are driving several new requirements for other L2VPN services such as E-TREE, and VPWS.

In the area of multi-homing current VPLS can only support multi-homing with active/standby resiliency model, for example as described in [VPLS-BGP-MH]. Flexible multi-homing with all-active Attachment Circuits (ACs) cannot be supported by current VPLS solution.

In the area of multicast optimization, [VPLS-MCAST] describes how multicast LSPs can be used in conjunction with VPLS. However, this solution is limited to Point-to-Multipoint (P2MP) LSPs, as there's no defined solution for leveraging Multipoint-to-Multipoint (MP2MP) LSPs with VPLS.

In the area of provisioning simplicity, current VPLS does offer a mechanism for single-sided provisioning by relying on BGP-based service auto-discovery [RFC4761][RFC6074]. This, however, still requires the operator to configure a number of network-side parameters on top of the access-side Ethernet configuration.

Furthermore, data center interconnect applications are driving the need for new service interface types which are a hybrid combination of VLAN Bundling and VLAN-based service interfaces. These are

referred to as "VLAN-aware Bundling" service interfaces.

Also virtualization applications are fueling an increase in the volume of MAC addresses that are to be handled by the network, which gives rise to the requirement for having the network re-convergence upon failure be independent of the number of MAC addresses learned by the PE.

In addition, there are requirements for minimizing the amount of flooding of multi-destination frames and localizing the flooding to the confines of a given site.

Moreover, there are requirements for supporting flexible VPN topologies and policies beyond those currently covered by (H-)VPLS.

The focus of this document is on defining the requirements for a new solution, namely Ethernet VPN (EVPN), which addresses the above issues.

Section 4 discusses the redundancy requirements. Section 5 describes the multicast optimization requirements. Section 6 articulates the ease of provisioning requirements. Section 7 focuses on the new service interface requirements. Section 8 highlights the fast convergence requirements. Section 9 describes the flood suppression requirement, and finally section 10 discusses the requirements for supporting flexible VPN topologies and policies.

4. Redundancy Requirements

4.1. Flow-based Load Balancing

A common mechanism for multi-homing a CE node to a set of PE nodes involves leveraging multi-chassis Ethernet link aggregation groups based on [802.1AX]. [PWE3-ICCP] describes one such scheme. In Ethernet link aggregation, the load-balancing algorithms by which a CE distributes traffic over the Attachment Circuits connecting to the PEs are quite flexible. The only requirement is for the algorithm to ensure in-order frame delivery for a given traffic flow. In typical implementations, these algorithms involve selecting an outbound link within the bundle based on a hash function that identifies a flow based on one or more of the following fields:

i. Layer 2: Source MAC Address, Destination MAC Address, VLAN
ii. Layer 3: Source IP Address, Destination IP Address
iii. Layer 4: UDP or TCP Source Port, Destination Port

A key point to note here is that [802.1AX] does not define a standard load-balancing algorithm for Ethernet bundles, and as such different implementations behave differently. As a matter of fact, a bundle operates correctly even in the presence of asymmetric load-balancing over the links. This being the case, the first requirement for active/active multi-homing is the ability to accommodate flexible flow-based load-balancing from the CE node based on L2, L3 and/or L4 header fields.

A solution MUST be capable of supporting flexible flow-based load balancing from the CE as described above. Further the MPLS network MUST be able to support flow-based load-balancing of traffic destined to the CE, even when the CE is connected to more than one PE. Thus the solution MUST be able to exercise multiple links connected to the CE, irrespective of the number of PEs that the CE is connected to. It should be noted that when a CE is multi-homed to several PEs, there could be multiple ECMP paths from each remote PE to each multi-homed PE. Furthermore, for active/active multi-homed site, a remote PE can choose any of the multi-homed PEs for sending traffic destined to the multi-homed sites. Therefore, when a solution supports active/active multi-homing, it MUST exercise as many of these paths as possible for traffic destined to a multi-homed site.

A solution MAY support flow-based load balancing among PEs that are members of a redundancy group spanning multiple Autonomous Systems.

4.2. Flow-based Multi-pathing

Any solution that meets the active-active flow based load balancing requirement described in section 4.1 MUST also be able to exercise multiple paths between a given pair of PEs. For instance, if there are multiple RSVP-TE LSPs between a pair of PEs then the solution MUST be capable of load balancing traffic among those LSPs on a per flow basis. Similarly, if LDP is being used as the signaling protocol for transport LSPs, then the solution MUST be able to leverage LDP signaled equal cost LSPs. The solution MUST also be able to leverage work in the MPLS WG that is in progress to improve the load balancing capabilities of the network based on entropy labels [RFC6790].

It is worth pointing out that flow-based multi-pathing complements flow-based load balancing described in the previous section.

4.3. Geo-redundant PE Nodes

The PE nodes offering multi-homed connectivity to a CE or access network may be situated in the same physical location (co-located),

or may be spread geographically (e.g., in different COs or POPs). The latter is desirable when offering a geo-redundant solution that ensures business continuity for critical applications in the case of power outages, natural disasters, etc. An active/active multi-homing mechanism SHOULD support both co-located as well as geo-redundant PE placement. The latter scenario often means that requiring a dedicated link between the PEs, for the operation of the multi-homing mechanism, is not appealing from a cost standpoint. Furthermore, the IGP cost from remote PEs to the pair of PEs in the dual-homed setup cannot be assumed to be the same when those latter PEs are geo-redundant.

A solution MUST support active/active multi-homing without the need for a dedicated control/data link among the PEs in the multi-homed group.

A solution MUST NOT assume that the IGP cost from a remote PE to each of the PEs in the multi-homed group is the same.

A solution MUST support multi-homing across different IGP domains within the same Autonomous System.

A solution SHOULD support multi-homing across multiple Autonomous Systems.

4.4. Optimal Traffic Forwarding

In a typical network, and considering a designated pair of PEs, it is common to find both single-homed as well as multi-homed CEs being connected to those PEs. An active/active multi-homing solution SHOULD support optimal forwarding of unicast traffic for all the following scenarios. By "optimal forwarding", we mean that traffic will not be forwarded between PE devices that are members of a multi-home group unless the destination CE is attached to one of the multi-homed PEs.

i. single-homed CE to single-homed CE
ii. single-homed CE to multi-homed CE
iii. multi-homed CE to single-homed CE
iv. multi-homed CE to multi-homed CE

This is especially important in the case of geo-redundant PEs, where having traffic forwarded from one PE to another within the same multi-homed group introduces additional latency, on top of the inefficient use of the PE node's and core nodes' switching capacity. A multi-homed group (also known as a multi-chassis LAG) is a group of PEs supporting a multi-homed CE.

4.5. Flexible Redundancy Grouping Support

In order to simplify service provisioning and activation, the multi-homing mechanism SHOULD allow arbitrary grouping of PE nodes into redundancy groups where each redundancy group represents all multi-homed groups that share the same group of PEs. This is best explained with an example: consider three PE nodes - PE1, PE2 and PE3. The multi-homing mechanism MUST allow a given PE, say PE1, to be part of multiple redundancy groups concurrently. For example, there can be a group (PE1, PE2), a group (PE1, PE3), and another group (PE2, PE3) where CEs could be multi-homed to any one of these three redundancy groups.

4.6. Multi-homed Network

There are applications, which require an Ethernet network, rather than a single device, to be multi-homed to a group of PEs. The Ethernet network would typically run a resiliency mechanism such as Multiple Spanning Tree Protocol [802.1Q] or Ethernet Ring Protection Switching [G.8032]. The PEs may or may not participate in the control protocol of the Ethernet network. For a multi-homed network running [802.1Q] or [G.8032], these protocols require that each VLAN to be active only on one of the multi-homed links.

A solution MUST support multi-homed network connectivity with active/standby redundancy.

A solution MUST also support multi-homed network with active/active VLAN-based load balancing (i.e. disjoint VLAN sets active on disparate PEs).

A solution MAY support VLAN-based load balancing among PEs that are member of a redundancy group spanning multiple ASes.

A solution MAY support multi-homed network with active/active MAC-based load balancing (i.e. different MAC addresses on a VLAN are reachable via different PEs).

5. Multicast Optimization Requirements

There are environments where the usage of MP2MP LSPs may be desirable for optimizing multicast, broadcast and unknown unicast traffic in order to reduce the amount of multicast states in the core routers. [VPLS-MCAST] precludes the usage of MP2MP LSPs since current VPLS solutions require an egress PE to perform learning when it receives unknown unicast packets over a LSP. This is challenging when MP2MP

LSPs are used as MP2MP LSPs do not have inherent mechanisms to identify the sender. The usage of MP2MP LSPs for multicast optimization becomes tractable if the need to identify the sender for performing learning is lifted. A solution MUST be able to provide a mechanism that does not require learning when packets are received over a MP2MP LSP. Further a solution MUST be able to provide procedures to use MP2MP LSPs for optimizing delivery of multicast, broadcast and unknown unicast traffic.

6. Ease of Provisioning Requirements

As L2VPN technologies expand into enterprise deployments, ease of provisioning becomes paramount. Even though current VPLS has an auto-discovery mechanism, which enables single-sided provisioning, further simplifications are required, as outlined below:

- Single-sided provisioning behavior MUST be maintained.

- For deployments where VLAN identifiers are global across the MPLS network (i.e. the network is limited to a maximum of 4K services), the PE devices SHOULD derive the MPLS specific attributes (e.g., VPN ID, BGP Route Target, etc.) from the VLAN identifier. This way, it is sufficient for the network operator to configure the VLAN identifier(s) for the access circuit, and all the MPLS and BGP parameters required for setting up the service over the core network would be automatically derived without any need for explicit configuration.

- Implementations SHOULD revert to using default values for parameters as and where applicable.

7. New Service Interface Requirements

[MEF] and [IEEE 802.1Q] have the following services specified:

- Port mode: in this mode, all traffic on the port is mapped to a single bridge domain and a single corresponding L2VPN service instance. Customer VLAN transparency is guaranteed end-to-end.

- VLAN mode: in this mode, each VLAN on the port is mapped to a unique bridge domain and corresponding L2VPN service instance. This mode allows for service multiplexing over the port and supports optional VLAN translation.

- VLAN bundling: in this mode, a group of VLANs on the port are collectively mapped to a unique bridge domain and corresponding L2VPN service instance. Customer MAC addresses must be unique across all

VLANs mapped to the same service instance.

For each of the above services a single bridge domain is assigned per service instance on the PE supporting the associated service. For example, in case of the port mode, a single bridge domain is assigned for all the ports belonging to that service instance regardless of number of VLANs coming through these ports.

It is worth noting that the term 'bridge domain' as used above refers to a MAC forwarding table as defined in the IEEE bridge model, and does not denote or imply any specific implementation.

[RFC4762] defines two types of VPLS services based on "unqualified and qualified learning" which in turn maps to port mode and VLAN mode respectively.

A solution is required to support the above three service types plus two additional service types which are primarily intended for hosted data center applications and are described below.

For hosted data center interconnect applications, network operators require the ability to extend Ethernet VLANs over a WAN using a single L2VPN instance while maintaining data-plane separation between the various VLANs associated with that instance. This gives rise to two new service interface types: VLAN-aware Bundling without Translation, and VLAN-aware Bundling with Translation.

The VLAN-aware Bundling without Translation service interface has the following characteristics:

- The service interface MUST provide bundling of customer VLANs into a single L2VPN service instance.

- The service interface MUST guarantee customer VLAN transparency end-to-end.

- The service interface MUST maintain data-plane separation between the customer VLANs (i.e. create a dedicated bridge-domain per VLAN).

- In the special case of all-to-one bundling, the service interface MUST NOT assume any a priori knowledge of the customer VLANs. In other words, the customer VLANs shall not be configured on the PE, rather the interface is configured just like a port-based service.

The VLAN-aware Bundling with Translation service interface has the following characteristics:

- The service interface MUST provide bundling of customer VLANs into

a single L2VPN service instance.

- The service interface MUST maintain data-plane separation between the customer VLANs (i.e. create a dedicated bridge-domain per VLAN).

- The service interface MUST support customer VLAN translation to handle the scenario where different VLAN Identifiers (VIDs) are used on different interfaces to designate the same customer VLAN.

The main difference, in terms of service provider resource allocation, between these new service types and the previously defined three types is that the new services require several bridge domains to be allocated (one per customer VLAN) per L2VPN service instance as opposed to a single bridge domain per L2VPN service instance.

8. Fast Convergence

A solution MUST provide the ability to recover from PE-CE attachment circuit failures as well as PE node failure for the case of both multi-homed device and multi-homed network. The recovery mechanism(s) MUST provide convergence time that is independent of the number of MAC addresses learned by the PE. This is particularly important in the context of virtualization applications which are fueling an increase in the number of MAC addresses to be handled by the Layer 2 network. Furthermore, the recovery mechanism(s) SHOULD provide convergence time that is independent of the number of service instances associated with the attachment circuit or the PE.

9. Flood Suppression

The solution SHOULD allow the network operator to choose whether unknown unicast frames are to be dropped or to be flooded. This attribute needs to be configurable on a per service instance basis.

In addition, for the case where the solution is used for data-center interconnect, it is required to minimize the flooding of broadcast frames outside the confines of a given site. Of particular interest is periodic ARP traffic.

Furthermore, it is required to eliminate any unnecessary flooding of unicast traffic upon topology changes, especially in the case of multi-homed site where the PEs have a priori knowledge of the backup paths for a given MAC address.

10. Supporting Flexible VPN Topologies and Policies

A solution MUST be capable of supporting flexible VPN topologies that are not constrained by the underlying mechanisms of the solution. One example of this is E-TREE topology where one or more sites in the VPN are roots and the others are leaves. The roots are allowed to send traffic to other roots and to leaves, while leaves can communicate only with the roots. The solution MUST provide the ability to support E-TREE topology. Further the solution MUST provide the ability to apply policies at the MAC address granularity to control which PEs in the VPN learn which MAC address and how a specific MAC address is forwarded. It MUST be possible to apply policies to allow only some of the member PEs in the VPN to send or receive traffic for a particular MAC address.

A solution MUST be capable of supporting both inter-AS option-C and inter-AS option-B scenarios as described in [RFC4364].

11. Contributors

Samer Salam, Cisco, ssalam@cisco.com
John Drake, Juniper, jdrake@juniper.net
Clarence Filsfils, Cisco, cfilsfil@cisco.com

12. Security Considerations

For scenarios where MAC learning is performed in the data-plane, there are no additional security aspects beyond those considered in [RFC4761] and [RFC4762]. And for scenarios where MAC learning is performed in the control plane (via BGP), there are no additional security aspects beyond those considered in [RFC4364].

13. IANA Considerations

None.

14. Normative References

[RFC2119] "Key words for use in RFCs to Indicate Requirement Levels", August 1996.

[RFC4761] Kompella, K. and Y. Rekhter, "Virtual Private LAN Service (VPLS) Using BGP for Auto-Discovery and Signaling", RFC 4761, January 2007.

[RFC4762] Lasserre, M. and V. Kompella, "Virtual Private LAN Service (VPLS) Using Label Distribution Protocol (LDP) Signaling", RFC 4762, January 2007.

[RFC4364] "BGP/MPLS IP Virtual Private Networks (VPNs)", February 2006.

[802.1AX] IEEE Std. 802.1AX-2008, "IEEE Standard for Local and metropolitan area networks - Link Aggregation", IEEE Computer Society, November 2008.

[802.1Q] IEEE Std. 802.1Q-2011, "IEEE Standard for Local and metropolitan area networks - Virtual Bridged Local Area Networks", 2011.

[RFC6074] E. Rosen and B. Davie, "Provisioning, Auto-Discovery, and Signaling in Layer 2 Virtual Private Networks (L2VPNs)", January 2011.

14. Informative References

[RFC4664] "Framework for Layer 2 Virtual Private Networks (L2VPNs)", September 2006.

[VPLS-BGP-MH] Kothari et al., "BGP based Multi-homing in Virtual Private LAN Service", draft-ietf-l2vpn-vpls-multihoming-05, work in progress, February, 2013.

[PWE3-ICCP] Martini et al., "Inter-Chassis Communication Protocol for L2VPN PE Redundancy", draft-ietf-pwe3-iccp-11.txt, work in progress, February, 2013.

[VPLS-MCAST] R. Aggarwal, et al., "Multicast in VPLS", draft-ietf-l2vpn-vpls-mcast-14.txt, work in progress, July 2013.

[MEF] MEF 6.1 Technical Specification, "Ethernet Service Definitions", April 2008.

[RFC6790] K. Kompella et al., "The Use of Entropy Labels in MPLS Forwarding", RFC 6790, November 2012.

15. Author's Address

Ali Sajassi
Cisco
Email: sajassi@cisco.com

Rahul Aggarwal
Arktan
Email: raggarwa_1@yahoo.com

Wim Henderickx
Alcatel-Lucent
Email: wim.henderickx@alcatel-lucent.com

Aldrin Isaac
Bloomberg
Email: aisaac71@bloomberg.net

James Uttaro
AT&T
Email: uttaro@att.com

Nabil Bitar
Verizon Communications
Email : nabil.n.bitar@verizon.com