Internet Engineering Task Force (IETF)
Request for Comments: 7209
Category: Informational
ISSN: 2070-1721

A. Sajassi
Cisco
R. Aggarwal
Arktan
J. Uttaro
AT&T
N. Bitar
Verizon
W. Henderickx
Alcatel-Lucent
A. Isaac
Bloomberg
May 2014

Requirements for Ethernet VPN (EVPN)

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, multihoming with all-active forwarding is not supported, and there's no existing solution to leverage Multipoint-to-Multipoint (MP2MP) Label Switched Paths (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.

Status of This Memo

This document is not an Internet Standards Track specification; it is published for informational purposes.

This document is a product of the Internet Engineering Task Force (IETF). It represents the consensus of the IETF community. It has received public review and has been approved for publication by the Internet Engineering Steering Group (IESG). Not all documents approved by the IESG are a candidate for any level of Internet Standard; see Section 2 of RFC 5741.

Information about the current status of this document, any errata, and how to provide feedback on it may be obtained at http://www.rfc-editor.org/info/rfc7209.

Table of Contents

1. Introduction

Virtual Private LAN Service (VPLS), as defined in [RFC4664], [RFC4761], and [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 Ethernet Tree (E-Tree) and Virtual Private Wire Service (VPWS).

In the area of multihoming, current VPLS can only support multihoming with the single-active redundancy mode (defined in Section 3), for example, as described in [VPLS-BGP-MH]. Flexible multihoming with all-active redundancy mode (defined in Section 3) cannot be supported by the current VPLS solution.

In the area of multicast optimization, [RFC7117] 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.

In the area of data-center interconnect, applications are driving the need for new service interface types that are a hybrid combination of VLAN bundling and VLAN-based service interfaces. These are referred to as "VLAN-aware bundling" service interfaces.

Virtualization applications are also fueling an increase in the volume of MAC (Media Access Control) addresses that are to be handled by the network; this gives rise to the requirement for having the network reconvergence upon failure be independent of the number of MAC addresses learned by the Provider Edge (PE).

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

There are also requirements for supporting flexible VPN topologies and policies beyond those currently covered by VPLS and Hierarchical VPLS (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.

2. 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].

This document is not a protocol specification and the key words in this document are used for clarity and emphasis of requirements language.

3. Terminology

AS: Autonomous System

CE: Customer Edge

E-Tree: Ethernet Tree

MAC address: Media Access Control address - referred to as MAC

LSP: Label Switched Path

PE: Provider Edge

MP2MP: Multipoint to Multipoint

VPLS: Virtual Private LAN Service

Single-Active Redundancy Mode: When a device or a network is multihomed to a group of two or more PEs and when only a single PE in such a redundancy group can forward traffic to/from the multihomed device or network for a given VLAN, such multihoming is referred to as "Single-Active".

All-Active Redundancy Mode: When a device is multihomed to a group of two or more PEs and when all PEs in such redundancy group can forward traffic to/from the multihomed device or network for a given VLAN, such multihoming is referred to as "All-Active".

4. Redundancy Requirements

4.1. Flow-Based Load Balancing

A common mechanism for multihoming a CE node to a set of PE nodes involves leveraging multi-chassis Ethernet link aggregation groups (LAGs) 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 all-active multihoming is the ability to accommodate flexible flow-based load balancing from the CE node based on L2, L3, and/or L4 header fields.

(R1a) A solution MUST be capable of supporting flexible flow-based load balancing from the CE as described above.

(R1b) A solution MUST also 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 multihomed to several PEs, there could be multiple Equal-Cost Multipath (ECMP) paths from each remote PE to each multihoming PE. Furthermore, for an all-active multihomed CE, a remote PE can choose any of the multihoming PEs for sending

traffic destined to the multihomed CE. Therefore, when a solution supports all-active multihoming, it MUST exercise as many of these paths as possible for traffic destined to a multihomed CE.

(R1c) A solution SHOULD support flow-based load balancing among PEs that are members of a redundancy group spanning multiple Autonomous Systems.

4.2. Flow-Based Multipathing

Any solution that meets the all-active redundancy mode (e.g., flow-based load balancing) described in Section 4.1, also needs to exercise multiple paths between a given pair of PEs. For instance, if there are two or more LSPs between a remote PE and a pair of PEs in an all-active redundancy group, then the solution needs to be capable of load balancing traffic among those LSPs on a per-flow basis for traffic destined to the PEs in the redundancy group. Furthermore, if there are two or more ECMP paths between a remote PE and one of the PEs in the redundancy group, then the solution needs to leverage all the equal-cost LSPs. For the latter, the solution can also leverage the load-balancing capabilities based on entropy labels [RFC6790].

(R2a) A solution MUST be able to exercise all LSPs between a remote PE and all the PEs in the redundancy group with all-active multihoming.

(R2b) A solution MUST be able to exercise all ECMP paths between a remote PE and any of the PEs in the redundancy group with all-active multihoming.

For example, consider a scenario in which CE1 is multihomed to PE1 and PE2, and CE2 is multihomed to PE3 and PE4 running in all-active redundancy mode. Furthermore, consider that there exist three ECMP paths between any of the CE1's and CE2's multihomed PEs. Traffic from CE1 to CE2 can be forwarded on twelve different paths over the MPLS/IP core as follows: CE1 load balances traffic to both PE1 and PE2. Each of PE1 and PE2 have three ECMP paths to PE3 and PE4 for a total of twelve paths. Finally, when traffic arrives at PE3 and PE4, it gets forwarded to CE2 over the Ethernet channel (aka link bundle).

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

4.3. Geo-redundant PE Nodes

The PE nodes offering multihomed 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 Central Offices (COs) or Points of Presence (POPs)). The latter is needed when offering a geo-redundant solution that ensures business continuity for critical applications in the case of power outages, natural disasters, etc. An all-active multihoming mechanism needs to 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 multihoming 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.

(R3a) A solution MUST support all-active multihoming without the need for a dedicated control/data link among the PEs in the multihomed group.

(R3b) A solution MUST support different IGP costs from a remote PE to each of the PEs in a multihomed group.

(R3c) A solution MUST support multihoming across different IGP domains within the same Autonomous System.

(R3d) A solution SHOULD support multihoming across multiple Autonomous Systems.

4.4. Optimal Traffic Forwarding

In a typical network, when considering a designated pair of PEs, it is common to find both single-homed as well as multihomed CEs being connected to those PEs.

(R4) An all-active multihoming 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 multihomed group unless the destination CE is attached to one of the multihoming PEs.

i. single-homed CE to multihomed CE
ii. multihomed CE to single-homed CE
iii. multihomed CE to multihomed CE

This is especially important in the case of geo-redundant PEs, where having traffic forwarded from one PE to another within the same

multihomed group introduces additional latency, on top of the inefficient use of the PE node's and core nodes' switching capacity. A multihomed group (also known as a multi-chassis LAG) is a group of PEs supporting a multihomed CE.

4.5. Support for Flexible Redundancy Grouping

(R5) In order to support flexible redundancy grouping, the multihoming mechanism SHOULD allow arbitrary grouping of PE nodes into redundancy groups where each redundancy group represents all multihomed devices/networks that share the same group of PEs.

This is best explained with an example: consider three PE nodes -- PE1, PE2, and PE3. The multihoming 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 multihomed to any one of these three redundancy groups.

4.6. Multihomed Network

There are applications that require an Ethernet network, rather than a single device, to be multihomed 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 multihomed network running [802.1Q] or [G.8032], these protocols require that each VLAN to be active only on one of the multihomed links.

(R6a) A solution MUST support multihomed network connectivity with single-active redundancy mode where all VLANs are active on one PE.

(R6b) A solution MUST also support multihomed networks with single-active redundancy mode where disjoint VLAN sets are active on disparate PEs.

(R6c) A solution SHOULD support single-active redundancy mode among PEs that are members of a redundancy group spanning multiple ASes.

(R6d) A solution MAY support all-active redundancy mode for a multihomed network with 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 use 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. [RFC7117] precludes the use of MP2MP LSPs since current VPLS solutions require an egress PE to perform learning when it receives unknown unicast packets over an LSP. This is challenging when MP2MP LSPs are used, as they do not have inherent mechanisms to identify the sender. The use of MP2MP LSPs for multicast optimization becomes tractable if the need to identify the sender for performing learning is lifted.

(R7a) A solution MUST be able to provide a mechanism that does not require MAC learning against MPLS LSPs when packets are received over a MP2MP LSP.

(R7b) A solution SHOULD 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 automated discovery of member PEs belonging to a given VPN instance over the MPLS/IP core network, further simplifications are required, as outlined below:

(R8a) The solution MUST support auto-discovery of VPN member PEs over the MPLS/IP core network, similar to the VPLS auto-discovery mechanism described in [RFC4761] and [RFC6074].

(R8b) The solution SHOULD support auto-discovery of PEs belonging to a given redundancy or multihomed group.

(R8c) The solution SHOULD support auto-sensing of the site ID for a multihomed device or network and support auto-generation of the redundancy group ID based on the site ID.

(R8d) The solution SHOULD support automated Designated Forwarder (DF) election among PEs participating in a redundancy (multihoming) group and be able to divide service instances (e.g., VLANs) among member PEs of the redundancy group.

(R8e) 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.

(R8f) Implementations SHOULD revert to using default values for parameters for which no new values are configured.

7. New Service Interface Requirements

[MEF] and [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 the 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.

(R9a) A solution MUST support the above three service types (port mode, VLAN mode, and VLAN bundling).

For hosted applications for data-center interconnect, 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 is referred to as 'VLAN-aware bundling service'.

(R9b) A solution MAY support VLAN-aware bundling service.

This gives rise to two new service interface types: VLAN-aware bundling without translation and VLAN-aware bundling with translation.

The service interface for VLAN-aware bundling without translation has the following characteristics:

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

- The service interface guarantees customer VLAN transparency end to end.

- The service interface maintains data-plane separation between the customer VLANs (i.e., creates 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 service interface for VLAN-aware bundling with translation has the following characteristics:

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

- The service interface maintains data-plane separation between the customer VLANs (i.e., creates a dedicated bridge-domain per VLAN).

- The service interface supports customer VLAN ID 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

(R10a) A solution MUST provide the ability to recover from PE-CE attachment circuit failures as well as PE node failure for the cases of both multihomed device and multihomed network.

(R10b) 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.

(R10c) 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

(R11a) 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.

(R11b) In addition, for the case where the solution is used for data-center interconnect, the solution SHOULD minimize the flooding of broadcast frames outside the confines of a given site. Of particular interest is periodic Address Resolution Protocol (ARP) traffic.

(R11c) Furthermore, the solution SHOULD eliminate any unnecessary flooding of unicast traffic upon topology changes, especially in the case of a multihomed site where the PEs have a priori knowledge of the backup paths for a given MAC address.

10. Supporting Flexible VPN Topologies and Policies

(R12a) 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.

(R12b) The solution MAY provide the ability to apply policies at the granularity of the MAC address to control which PEs in the VPN learn which MAC address and how a specific MAC address is forwarded. It should 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.

(R12c) A solution MUST be capable of supporting both inter-AS option-C and inter-AS option-B scenarios as described in [RFC4364].

11. Security Considerations

Any protocol extensions developed for the EVPN solution shall include the appropriate security analysis. Besides the security requirements covered in [RFC4761] and [RFC4762] when MAC learning is performed in data-plane and in [RFC4364] when MAC learning is performed in control plane, the following additional requirements need to be covered.

(R13) A solution MUST be capable of detecting and properly handling a situation where the same MAC address appears behind two different Ethernet segments (whether inadvertently or maliciously).

(R14) A solution MUST be capable of associating a MAC address to a specific Ethernet segment (aka "sticky MAC") in order to help limit malicious traffic into a network for that MAC address. This capability can limit the appearance of spoofed MAC addresses on a network. When this feature is enabled, the MAC mobility for such sticky MAC addresses are disallowed, and the traffic for such MAC addresses from any other Ethernet segment MUST be discarded.

12. Normative References

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

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

[G.8032] ITU-T, "Ethernet ring protection switching", ITU-T Recommendation G.8032, February 2012.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC4364] Bersani, F. and H. Tschofenig, "The EAP-PSK Protocol: A Pre-Shared Key Extensible Authentication Protocol (EAP) Method", RFC 4764, January 2007.

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

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

[RFC6074] Rosen, E., Davie, B., Radoaca, V., and W. Luo, "Provisioning, Auto-Discovery, and Signaling in Layer 2 Virtual Private Networks (L2VPNs)", RFC 6074, January 2011.

13. Informative References

[VPLS-BGP-MH] Kothari, B., Kompella, K., Henderickx, W., Balue, F., Uttaro, J., Palislamovic, S., and W. Lin, "BGP based Multi-homing in Virtual Private LAN Service", Work in Progress, July 2013.

[PWE3-ICCP] Martini, L., Salam, S., Sajassi, A., and S. Matsushima, "Inter-Chassis Communication Protocol for L2VPN PE Redundancy", Work in Progress, March 2014.

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

[RFC4664] Andersson, L., Ed., and E. Rosen, Ed., "Framework for Layer 2 Virtual Private Networks (L2VPNs)", RFC 4664, September 2006.

[RFC6790] Kompella, K., Drake, J., Amante, S., Henderickx, W., and L. Yong, "The Use of Entropy Labels in MPLS Forwarding", RFC 6790, November 2012.

[RFC7117] Aggarwal, R., Ed., Kamite, Y., Fang, L., Rekhter, Y., and C. Kodeboniya, "Multicast in Virtual Private LAN Service (VPLS)", RFC 7117, February 2014.

14. Contributors

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

Authors' Addresses

Ali Sajassi
Cisco
EMail: sajassi@cisco.com

Rahul Aggarwal
Arktan
EMail: raggarwa_1@yahoo.com

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

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

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

Aldrin Isaac
Bloomberg
EMail: aisaac71@bloomberg.net