BESS Workgroup
Internet-Draft
Intended status: Standards Track
Expires: August 25, 2021

S. Boutros, Ed.
S. Sivabalan, Ed.
H. Shah
Ciena Corporation
J. Uttaro
ATT
D. Voyer
Bell Canada
B. Wen
Comcast
L. Jalil
Verizon

February 21, 2021

A Simplified Scalable ELAN Service Model with Segment Routing Underlay
draft-boutros-bess-elan-services-over-sr-02

Abstract

This document proposes a new approach for deploying Ethernet LAN (ELAN) services with an objective of achieving high scalability, faster network convergence, and reduced operational complexity. Furthermore, it naturally brings the benefits of All-Active multihoming as well as MAC learning in data-plane.

Status of This Memo

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

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at https://datatracker.ietf.org/drafts/current/.

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."

This Internet-Draft will expire on August 25, 2021.

Table of Contents

1. Introduction

Virtual Private LAN Service(VPLS) is based on Pseudo-Wire (PW) construct which identifies both the service type and the service termination node in both control and data planes. RFCs 4761 and 4762 specify mechanisms to signal PW for VPLS services using BGP and LDP respectively. An ingress Provider Edge (PE) node needs to maintain a PW per VPLS instance for each egress PE node. So, if we assume 10K ELAN instances over a network of 100 PE nodes, each PE node needs to

setup and maintain approximately 1M PWs which can easily become a scalability bottleneck in large scale deployment.

As described in RFC7432, Ethernet Virtual Private Network (EVPN) technology builds ELAN services similar to BGP-based IP-VPN services with additional features such as MAC address learning in control lane, All-Active multihoming, etc. It eliminates the need for PWs, and hence the scale problem associated with PWs. However, an egress PE node cannot unambiguously identify ingress PE node in data-plane. As such, EVPN requires control plane mechanisms for MAC advertisement and learning which increases control plane complexity and overhead.

The goal of the proposed approach is to greatly simplify control plane functions and minimize the amount of control plane messages PE nodes have to process. In this version of the document, we assume Segment Routing (SR) underlay network. A future version of this document will generalize the underlay network to both classical MPLS and SR technologies.

The proposed approach does not require PW, and hence the control plane complexity and message overhead associated with signaling and maintaining PWs are eliminated.

An ELAN instance is uniquely identified by Segment ID (SID) regardless of the number of service termination points. Such a SID will be referred to as "Service SID" in the rest of the document. The number of states maintained at a PE node is equal to the number of ELAN instances in the corresponding broadcast domain. Referring to the above example, each PE node now needs to maintain states for 10K ELAN service instances as opposed to 1 M PWs in the case of classical VPLS model in data and control planes. A node can advertise service SID(s) of the ELAN instance(s) that it hosts via BGP for auto-discovery purpose. A Service SID can be:

- MPLS label for SR-MPLS.

- uSID (micro SID) for SRv6 representing network function associated with an ELAN service instance.

MAC address is learned in data-plane. Source node of a MAC address is identified by its node SID (assigned for regular SR operation) during MAC learning phase. In the data packets, the node SID of the source is inserted directly below the service SID so that a destination node can uniquely identify the source of the packets in an SR domain.

ELAN service instances are advertised such that a service message packs as many ELAN instances hosted by the advertising PE node as

possible at the time of advertisement. A possible approach is to use a bit-map in which each bit position represents an ELAN instance, as well as the starting value of Service SID. Using these parameters, an ingress PE receiving advertisements node can learn ELAN instance(s) hosted by an egress PE node.

All-Active multihoming redundancy is supported at the underlay level by making use of SR anycast SID. No overlay mechanism is required for this purpose.

Each node is also associated with another SID unique within the broadcast domain that is used to identify incoming Broadcast Unknown-unicast, and Multicast (BUM) traffic. We call such SID BUM SID. If node A wants to send BUM traffic to node B, it needs to use BUM SID assigned to node B as a destination SID. BUM SIDs can also be advertised via BGP for auto-discovery purpose. In order to send BUM traffic within a broadcast domain, P2MP SR policies can be used. Such policies may or may not be shared by ELAN instances.

The proposed solution can also be applicable to the EVPN control plane without compromising its benefits such as All-Active multihoming on access, multipathing in the core, auto-provisioning and auto-discovery, etc. With this approach, the need for advertising EVPN route types 1 through 4 as well Split-Horizon (HP) label is eliminated.

In the following sections, we will describe the functionalities of the proposed approach in detail.

2. 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 [RFC2119].

3. Abbreviations

BUM: Broadcast, unicast and multicast.

CE: Customer Edge node e.g., host or router or switch.

ELAN: Ethernet LAN.

EVPN: Ethernet VPN.

MAC: Media Access Control.

MAC-VRF: A Virtual Routing and Forwarding table for Media Access Control (MAC) addresses on a PE.

MH: Multi-Home.

OAM: Operations, Administration and Maintenance.

PE: Provide Edge Node.

SID: Segment Identifier.

SR: Segment Routing.

VPLS: Virtual Private LAN Service.

4. Control Plane Behavior

4.1. Service discovery

A node can discover ELAN service instances as well as the associated service SIDs hosted on other nodes via configuration or auto-discovery. With the latter, the service SIDs can be advertised using BGP. As mentioned earlier such update message will pack information about as many ELAN instances hosted by the advertising PE node to reduce the amount of update messages exchanged by PE nodes.

Similar to the service SID, an ingress PE node can discover BUM SID associated with an egress PE node via configuration or auto-discovery.

The necessary BGP extensions will be specified in a future version of this document.

4.2. All-Active Service Redundancy

An anycast SID per Ethernet Segment (ES) can be associated with the PE nodes attached to a Multi-Home (MH) CE. The anycast SIDs will be advertised in BGP by the PE nodes. Based on ES anycast SIDs, ingress PEs receiving updates can discover the redundancy membership and perform DF election. Aliasing/Multipathing can be achieved using the same mechanisms excercised by SR underlay for forwarding traffic to destinations belonging to anycast group.

4.3. Mass service withdrawal

Node failure can be detected due via IGP convergence. For faster detection of node failure, mechanism like BFD can be deployed. The

proposed approach does not require additional MAC withdrawal mechanism.

On PE-CE link failure, the corresponding PE node withdraws the route to the corresponding ES in BGP in order to stop receiving traffic to that ES. With MH case with anycast SID, upon detecting a failure on PE-CE link, a PE node may forward incoming traffic to the impacted ES(s) to other PE node(s) that is/are part of the anycast group until it withdraws routes to the impacted ES(s) for faster convergence. For example, in Figure 1, assuming PE5 and PE6 are part of an anycast group, upon link failure between PE5 and CE5, PE5 can forward the received packets from the core to PE6 until it withdraws the anycast SID associated with the ES(s).

4.4. E-Tree Support

To be covered in the next revision of this document.

5. Data Plane Behavior

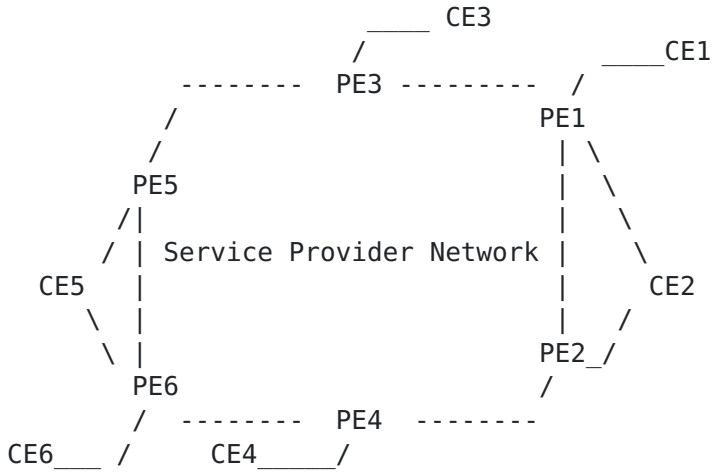

Figure 1: Reference network diagram used for examples below

5.1. Unicast Traffic

The proposed method requires unicast data packet be formed as shown in Figure 2.

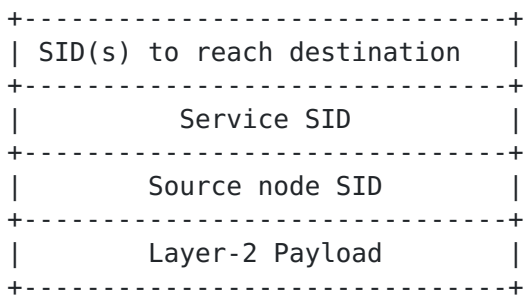

Figure 2: Data packet format for unicast traffic

- SID(s) to reach destination: depends on the intent of the underlay transport:
  - IGP shortest path: node SID of the destination. The destination can belong to an anycast group.
  - IGP path with intent: Flex-Algo SID if the destination can be reached using the Flex-Algo SID for a specific intent (e.g., low latency). The destination can belong to an anycast group.
  - SR policy (to support fine intent): a SID-list for the SR policy that can be used to reach the destination.
- Service SID: The SID that uniquely identifies an ELAN instance in a broadcast domain.
- Source node SID: The SID that uniquely identifies the source node. This can be a node SID which may be part of an anycast group. Note that such a SID is allocated as part of SR underlay operation, and the proposed approach does not impose any additional requirement.

5.2. BUM Traffic

In order to identify incoming BUM traffic a unique SID (which will be referred to as "BUM SID" in the rest of the document) per PE node is allocated. A BUM packet is formatted as shown in Figure 3:

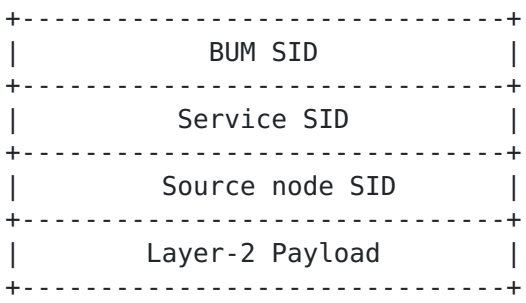

Figure 3: Data packet format for BUM traffic

In order to send BUM traffic, a P2MP SR policy may be established from a given node to rest of the nodes associated with an ELAN instance. If a dedicated P2MP SR policy is used per ELAN instance, a single SID may be used as both replication SID for the P2MP SR policy as well as to identify ELAN instance. With this approach, the number of SIDs imposed on data packet will be only two. It is possible to use a given P2MP SR policy for multiple ELAN instances in which case service SID needs to be inserted in the packet for egress PE to identify the ELAN instance for the BUM traffic.

5.3. Data Plane MAC learning

With the proposed approach, MAC address can be learned in data- plane using the packets formatted as shown in Figure 4.

Source MAC address on the received Layer 2 packet is learned against the source node SID placed directly under the service SID in the data-plane.

5.3.1. Single Home CE

In Figure 1, node 3 learns a MAC address from CE3 and floods it to all nodes configured with the same service SID. Nodes 1, 2, 4, 5 and 6 learn the MAC address as reachable via the source node SID of Node 3.

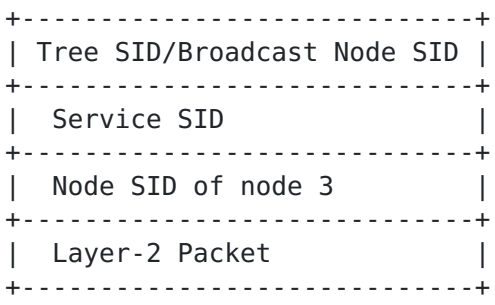

Figure 4: Packet format used for flooding

5.3.2. Multi-Home CE

Referring to Figure 1, let's assume that node 5 learns a MAC address from MH CE5, and floods it to all nodes in data-plane as per SID stack shown in Figure 5, including node 6. The receiving nodes learn the MAC address as reachable via the anycast SID belonging to node 5 and node 6. Node 6 applies SH and hence does not send the packet back to CE5, but treats the MAC address as reachable via CE5, as well floods the address to CE6.

The following diagram shows SID label stack for a Broadcast and Multicast MAC frame sent by Multi-Home PE. Note the presence of source SID after the service SID. This combination/order is necessary for the receiver to learn source MAC address (from L2 packet) associated with ingress PE (i.e. source node SID).

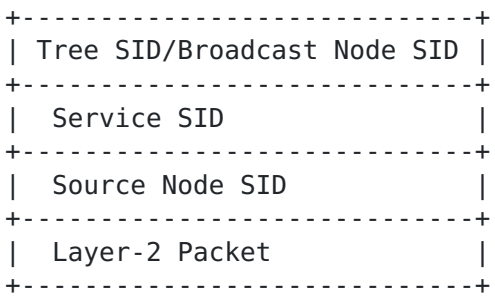

Figure 5: Data packet format for traffic sent by a MH PE

5.4. ARP suppression

Gleaning ARP packet requests and replies will be used to learn IP/MAC binding for ARP suppression. ARP replies are unicast, however flooding ARP replies can allow all nodes to learn the MAC/IP bindings for the destinations too.

5.5. Distributed Anycast Gateway

Distributed Anycast Gateway (GW) (aka inter-subnet IRB function) can be realized as follows:

- All PEs connected to the tenant subnets share the same GW IP/MAC per subnet.
- A PE MUST never learn its own GW IP/MAC via the tunnels connecting itself to other PE(s).
- ARP requests/replies from the tenant subnet are flooded via the ingress PE(s) attached to the subnet to all egress PE(s) attached to the subnet so that egress PE(s) can learn the source MAC/IP address via the ingress PE(s).
- ARP replies from tenants will be delivered to the local PE hosts the GW virtual MAC address. The local PE MUST flood the ARP replies over the tunnel to other PEs. Other PEs, including the PE which originated the ARP request, will learn the IP/MAC association of the tenant from the received ARP reply.

5.6. Multi-pathing

Packets destined to a MH CE is distributed to the PE nodes attached to the CE for load-balancing purpose. This is achieved implicitly due to the use of anycast SIDs for both ES as well as PE attached to the ES. In our example, traffic destined to CE5 is distributed via PE5 and PE6.

5.7. E-Tree Support

To be covered in the next revision of this document.

6. Benefits of ELAN over SR

The proposed approach eliminates the need for establishing and maintaining PWs as with legacy VPLS technology. This yields significant reduction in control plane overhead. Also, due to MAC learning in data-plane (conversational MAC learning), the proposed approach provides the benefits as such fast convergence, fast MAC movement, etc. Finally, using anycast SID, the proposed approach provides All-Active multihoming as well as multipathing and ARP suppression.

7. Security Considerations

The mechanisms in this document use Segment Routing control plane as defined in Security considerations described in Segment Routing control plane are equally applicable.

8. IANA Considerations

TBD.

9. Acknowledgements

10. References

10.1. Normative 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>.

[RFC8402] Filsfils, C., Ed., Previdi, S., Ed., Ginsberg, L., Decraene, B., Litkowski, S., and R. Shakir, "Segment Routing Architecture", RFC 8402, DOI 10.17487/RFC8402, July 2018, <https://www.rfc-editor.org/info/rfc8402>.

[RFC8660] Bashandy, A., Ed., Filsfils, C., Ed., Previdi, S., Decraene, B., Litkowski, S., and R. Shakir, "Segment Routing with the MPLS Data Plane", RFC 8660, DOI 10.17487/RFC8660, December 2019, <https://www.rfc-editor.org/info/rfc8660>.

[RFC8754] Filsfils, C., Ed., Dukes, D., Ed., Previdi, S., Leddy, J., Matsushima, S., and D. Voyer, "IPv6 Segment Routing Header (SRH)", RFC 8754, DOI 10.17487/RFC8754, March 2020, <https://www.rfc-editor.org/info/rfc8754>.

10.2. Informative References

[I-D.ietf-spring-segment-routing-policy] Filsfils, C., Talaulikar, K., Voyer, D., Bogdanov, A., and P. Mattes, "Segment Routing Policy Architecture", draft-ietf-spring-segment-routing-policy-09 (work in progress), November 2020.

[I-D.voyer-pim-sr-p2mp-policy]
Voyer, D., Filsfils, C., Parekh, R., Bidgoli, H., and Z. Zhang, "Segment Routing Point-to-Multipoint Policy", draft-voyer-pim-sr-p2mp-policy-02 (work in progress), July 2020.

[RFC4761] Kompella, K., Ed. and Y. Rekhter, Ed., "Virtual Private LAN Service (VPLS) Using BGP for Auto-Discovery and Signaling", RFC 4761, DOI 10.17487/RFC4761, January 2007, <https://www.rfc-editor.org/info/rfc4761>.

[RFC4762] Lasserre, M., Ed. and V. Kompella, Ed., "Virtual Private LAN Service (VPLS) Using Label Distribution Protocol (LDP) Signaling", RFC 4762, DOI 10.17487/RFC4762, January 2007, <https://www.rfc-editor.org/info/rfc4762>.

Authors' Addresses

Sami Boutros (editor)
Ciena Corporation
USA

Email: sboutros@ciena.com

Siva Sivabalan (editor)
Ciena Corporation
Canada

Email: ssivabal@ciena.com

Himanshu Shah
Ciena Corporation
USA

Email: hshah@ciena.com

James Uttaro
ATT
USA

Email: ju1738@att.com

Daniel Voyer
Bell Canada
Canada

Email: daniel.voyer@bell.ca

Bin Wen
Comcast
USA

Email: bin_wen@cable.comcast.com

Luay Jalil
Verizon
USA

Email: luay.jalil@verizon.com