SPRING Workgroup
Internet-Draft
Intended status: Standards Track
Expires: May 6, 2021

S. Boutros, Ed.
S. Sivabalan, Ed.
Ciena Corporation
J. Uttaro
AT&T
D. Voyer
Bell Canada
B. Wen
Comcast
L. Jalil
Verizon

November 2, 2020

A Simplified Scalable L3VPN Service Model with Segment Routing Underlay
draft-boutros-bess-l3vpn-services-over-sr-00

Abstract

This document proposes a new approach for realizing classical L3VPN (vpnv4/vpnv6/6PE/6VPE) over Segment Routing (SR) networks. It significantly improves scalability and convergence of the L3VPN control plane. Furthermore, it naturally brings the benefits of All-Active multi-homing support to the classical L3VPN.

Status of This Memo

Copyright Notice

Table of Contents

1. Introduction

Layer 3 VPN (L3VPN) enables a service provider to use an Internet Protocol (IP) backbone to provide IP VPNs for customers. This approach uses a peer model, in which the Customer Edge (CE) nodes send their routes to the Service Provider Edge (PE) nodes. BGP is used to exchange the routes of a particular VPN among the PE nodes that are attached to that VPN. This is done in a way that ensures that routes from different VPNs remain distinct and separate, even if two VPNs have an overlapping address space. The PE nodes distribute to the CE nodes in a particular VPN, the routes from other the CE nodes in that VPN. The CE nodes do not peer with each other. Each L3VPN route (v4/v6) advertisement is prepended with an 8-byte Route Distinguisher (RD) to allow the IP address space to be reused by multiple VPNs. Each L3VPN route is associated with a set of extended communities, i.e., Route Targets (RTs). Each L3VPN route can be associated with other attributes such as local preferences, MED

(Multi_EXIT_DISC attribute), color, etc. Each L3VPN route is associated with a tunnel encapsulation, i.e., MPLS label.

Current mechanisms require control plane scale to distribute a large number of VPN routes in the service provider network. In this document we propose a new approach that intends to simplify and improve the scalability of existing control plane to support L3VPN options A, B, and C using a global service SID per VPN across AS domains. Non mesh, hub/spoke and extranet topology can be realized using different SIDs or possibly RTs associated with the L3VPN services attached to different service routes. Non mesh topologies can then be realized by applying different import, export rules. The proposed control plane can be realized through protocols like BGP or using a centralized controller.

The proposed approach takes advantage of the inherent properties of SR. It maintains the existing L3VPN semantics to (1) allow overlapping IP addresses to be used across multiple VPNs and (2) associate routes with attributes. Further, it allows operators to represent an L3VPN instance by one or more globally allocated service Segment Identifiers (SID(s)). The VPN route import/export is governed by SID and allows the operator to deploy extranet, hub-and-spoke, and mesh VPN topologies. RT-based import/export can also be used to support non-mesh L3VPN sites. Also, the proposed approach provides All-Active redundancy and multi-pathing using SR anycast SIDs for Multi-Homed (MH) L3VPN sites. It significantly reduces the BGP overhead for L3VPN control planes by at least two orders of magnitude and, in mesh deployments by up to four orders of magnitude. At the same time, it does not compromise the desired benefits of L3VPN and EVPN prefix advertisements (RT-5), such as support of All-Active redundancy on access, multi-pathing in the core, auto-provisioning and auto-discovery.

The crux of the proposal is how the routes are advertised. All VPN routes originating from a PE node share the same tunnel encapsulation (ENCAP) to that PE node. Thus, we propose to advertise the tunnel encapsulation as the unique route, and the VPN prefixes as the attributes of the route. A new BGP message (to be specified in a future version of this document) will be used to advertise the route and attributes in the new format. The goal is to pack as many VPN prefixes as possible in a single BGP message. About 10k VPNv4 prefixes can be packed in a 64k message. With SRv6 and uSID, the ENCAP will be an IPv6 prefix that contains both the Node SID for the PE node as well as the Service SID representing the VPN. In common cases, this will be a /64 globally unique prefix.

A SID identifying a L3VPN instance (we call it "Service SID" in the rest of the document) can be:

- an MPLS label for SR-MPLS.

- a uSID (micro SID) for SRv6 representing network function associated with a VPLS instance. The new function will be specified in a future version of this document.

In the data packets, the service SID uniquely identify the L3VPN service in an SR domain.

Thanks to SR anycast SID capability, the proposed approach inherently provides All-Active multi-homing support.

A node can advertise service SID(s) of the L3VPN instance(s) that it is associated with via BGP for auto-discovery purpose. In the case of SR-MPLS, a service SID can be carried as a range of absolute values or an index into an Segment Routing Global Block (SRGB), and in the case of SRv6, a service SID can be carried as uSID in BGP updates. The objective is to pack information about all L3VPN service instances supported (at the time of sending update) on a transmitting node in single BGP update so as to reduce the amount of of overall BGP update messages in a network.

The proposed solution can also be applicable to EVPN control plane without compromising its benefits such as multi-active redundancy on access, multipathing in the core, auto-provisioning and auto-discovery, etc. With this approach, the need for advertisement of EVPN route types 1 and 5.

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

L3VPN: Layer 3 Virtual Private Network.

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

EVPN: Ethernet VPN.

MAC: Media Access Control.

VRF: A Virtual Routing and Forwarding table for Customer Routes on a PE.

MH: Multihome.

OAM: Operations, Administration and Maintenance.

PE: Provide Edge Node.

SID: Segment Identifier.

SR: Segment Routing.

BGP PIC: BGP Prefix independent convergence.

4. Control Plane Functionality

4.1. Service discovery

A node can discover L3VPN services instances as well as the associated service SIDs on other nodes via configuration or auto-discovery. With the latter, the service SIDs can be advertised using BGP. As mentioned earlier, the service SIDs can be MPLS label (absolute value or index into an SRGB) or SRv6 uSID.

VPNv4/v6 prefixes and operation type, i.e., to inform BGP neighbors whether prefixes are added or deleted, can be advertised in a new TLV. The prefixes will be packed efficiently; prefix length followed by prefixes sharing the same prefix length. With this format, at least 12k VPNv4 prefixes can be encoded in the message. A single route will carry a large number of VPN prefixes (e.g., ~10k VPNv4 prefixes), instead of advertising one route per each VPN prefix. In the case of VPNv4, this results in approximately four orders of magnitude reduction in BGP messages. L3VPN Service SIDs may be allocated from an SRGB range dedicated only for L3VPN services.

```
                                        ____ CE3
                                       /               ____CE1
                            --------  PE3 ---------  /
                           /                       PE1
                          /                         | \
                         PE5                        |  \
                        /|                          |   \
                       / | Service Provider Network |    CE2
                   CE5   |                          |   /
                      \  |                          |  /
                       \ |                         PE2/
                         PE6                       /
                         /  --------  PE4  --------
                 CE6___ /     CE4_____/
```

Figure 1: A Reference L3VPN Network

Each PE nodes (PE1 through PE6 in Figure 1) advertises, via IGP/BGP, (1) a regular Node SID to be used by the PE nodes when an L3VPN service is attached to local Single-Home sites, and/or (2) an anycast SID per MH site when an L3VPN service is attached to the MH site. For example, in Figure 1, the PE nodes PE3 and PE4 could advertise a Node SID for an L3VPN associated with the CE5 and CE3, respectively. For MH, the PE5 and PE6 can advertise an anycast SID for an L3VPN associated with the CE2. With the use of anycast SID per MH site, shared by PEs attached to the site, there is no need to implement any BGP PIC techniques at the L3VPN layer, as the routing convergence relies on the underlay of SR. The data plane can be MPLS or SRv6.

5. Data Plane Behavior

The proposed method requires L3 data packet be formed as shown in Figure 2.

```
                         +-------------------------------+
                         | SID(s) to reach destination   |
                         +-------------------------------+
                         |          Service SID          |
                         +-------------------------------+
                         |        Layer-3 Payload        |
                         +-------------------------------+
```

Figure 2: Data packet format for sending L3VPN traffic

o 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 path). 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 a L3VPN instance in an SR domain.

6. Service discovery

A node can discover L3VPN services instances as well as the associated service SIDs on other nodes via configuration or auto-discovery. With the latter, the service SIDs can be advertised using BGP. As mentioned earlier, the service SIDs can be MPLS label (absolute value or index into an SRGB) or SRv6 uSID.

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

7. All-Active service Redundancy

Referring to Figure 1, an anycast SID per MH Site is configured on all PE nodes PE1, PE2, PE5, and PE6 attached to the MH sites, such as CE2 and CE5. These anycast SIDs are advertised via BGP for reachability. Each PE node 1, 2 and 5, 6 attached to the MH site, advertises the same anycast SID to allow other nodes to discover the membership (auto-discovery). With SRv6, L3VPN routes associated with an MH site can be advertised as a single route containing both anycast SID of the egress PEs and service SIDs. Multi-pathing/Fast convergence achieved using the same mechanisms used for anycast SID. Single-Active redundancy is the same as the All-Active model except that the backup egress PE node advertises its route with a higher cost than the primary egress PE node.

8. 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 Figure 1, traffic destined to CE5 is distributed via PE5 and PE6.

9. Mass service withdrawal

Node failure is detected by IGP/BGP will converge. Technique like BFD shall be deployed for fast detection of failure.

On PE-CE link failure, the 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 nodes 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 MH site.

10. Benefits of L3VPN over SR

The proposed approach significantly reduces the control plane overhead, provides fast convergence, and All-Active multi-homing as well as multipathing benefits. The proposed approach eliminates the need for BGP PIC.

11. Security Considerations

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

12. IANA Considerations

TBD.

13. Acknowledgement

14. References

14.1. 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>.

14.2. 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-08 (work in progress), July 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.

[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

Sami Boutros (editor)
Ciena Corporation
Canada

Email: sboutros@ciena.com

Siva Sivabalan (editor)
Ciena Corporation
USA

Email: ssivabal@ciena.com

James Uttaro
AT&T
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