Workgroup: BESS Workgroup
Internet-Draft: draft-ietf-bess-evpn-geneve-06
Published: 26 May 2023
Intended Status: Standards Track
Expires: 27 November 2023
Authors: S. Boutros, Ed. A. Sajassi J. Drake
Ciena Cisco Systems Juniper Networks
J. Rabadan S. Aldrin
Nokia Google

EVPN control plane for Geneve

Abstract

This document describes how Ethernet VPN (EVPN) control plane can be used with Network Virtualization Overlay over Layer 3 (NVO3) Generic Network Virtualization Encapsulation (Geneve) encapsulation for NVO3 solutions.

EVPN control plane can also be used by Network Virtualization Edges (NVEs) to express Geneve tunnel option TLV(s) supported in the transmission and/or reception of Geneve encapsulated data packets.

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 27 November 2023.

Copyright Notice

respect to this document. Code Components extracted from this document must include Revised BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Revised BSD License.

Table of Contents

1. Introduction

The Network Virtualization over Layer 3 (NVO3) solutions for network virtualization in data center (DC) environment are based on an IP-based underlay. An NVO3 solution provides layer 2 and/or layer 3 overlay services for virtual networks enabling multi-tenancy and workload mobility.

This document describes how the EVPN control plane defined in [RFC7432] can signal Geneve encapsulation type in the BGP Tunnel Encapsulation Extended Community defined in [RFC9012]. In addition, this document defines how to communicate the Geneve tunnel option types using BGP Tunnel Encapsulation Attribute sub-TLV. The Geneve tunnel options are encapsulated as TLVs after the Geneve base header in the Geneve packet as described in [RFC8926].

[I-D.ietf-nvo3-encap] recommends that a control plane determine how Network Virtualization Edges (NVEs) use the Geneve option TLVs when sending/receiving packets. In particular, the control plane negotiates the subset of option TLVs supported, their order and the total number of option TLVs allowed in the packets. This negotiation capability allows, for example, interoperability with hardware-based NVEs that can process fewer options than software-based NVEs.

This EVPN control plane extension will allow an NVE to express what Geneve option TLV types it is capable of receiving, or sending over the Geneve tunnel with its peers.

In the datapath, a transmitting NVE MUST NOT encapsulate a packet destined to another NVE with any option TLV(s) the receiving NVE is not capable of processing.

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 and Terminology

NVO3: Network Virtualization Overlays over Layer 3

Geneve: Generic Network Virtualization Encapsulation.

NVE: Network Virtualization Edge.

VNI: Virtual Network Identifier.

MAC: Media Access Control.

OAM: Operations, Administration and Maintenance.

PE: Provide Edge Node.

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

EVPN: Ethernet VPN.

ES: Ethernet segment.

ESI: Ethernet Segment Identifier.

EVI: An EVPN instance spanning the Provider Edge (PE) devices participating in that EVPN.

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

4. Geneve extension

This document adds an extension to the [RFC8926] encapsulation that is relevant to the operation of EVPN.

4.1. Ethernet option TLV

[RFC8365] describes when an ingress NVE uses ingress replication to flood unknown unicast traffic to the egress NVEs, the ingress NVE needs to indicate to the egress NVE that the Encapsulated packet is a BUM packet. This is required to avoid transient packet duplication in all-active multi-homing scenarios. For Geneve, we need a bit for this purpose.

[RFC8317] uses an MPLS label for leaf indication of BUM traffic originated from a leaf attachment circuit (AC) in an ingress NVE so that the egress NVEs can filter BUM traffic toward their leaf ACs. For Geneve, we need a bit for this purpose.

Although the default mechanism for split-horizon filtering of BUM traffic on an Ethernet segment for IP-based encapsulations such as VxLAN, GPE, NVGRE, and Geneve, is local-bias as defined in section 8.3.1 of [RFC8365], there can be an incentive to leverage the same split-horizon filtering mechanism of [RFC7432] that uses a 20-bit MPLS label so that a) the a single filtering mechanism is used for all encapsulation types and b) the same PE can participate in a mix of MPLS and IP encapsulations. For this purpose a 20-bit label field MAY be defined for Geneve encapsulation. The support for this label is OPTIONAL.

If an NVE wants to use local-bias procedure, then it sends the new option TLV without ESI-label (e.g., length=4):

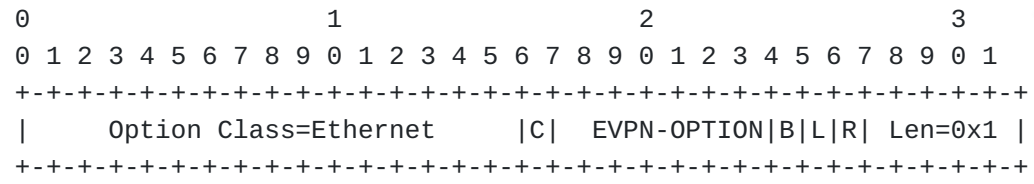

Figure 1: Ethernet Option TLV without ESI label

If an NVE wants to use ESI-label, then it sends the new option TLV with ESI-label (e.g., length=8)

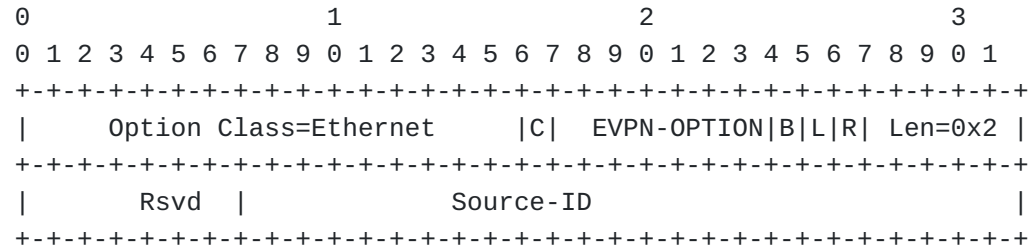

Figure 2: Ethernet Option TLV with ESI label

Where:

- Option Class is set to Ethernet (new Option Class requested to IANA)

- Type is set to EVPN-OPTION with value = 0, and C bit must be set.

- B bit is set to 1 for BUM traffic.

- L bit is set to 1 for Leaf-Indication.

- R bit is set to 1 for Root-Indication.

- Source-ID is a 24-bit value that encodes the ESI-label value signaled on the EVPN Autodiscovery per-ES routes, as described in [RFC7432] for multi-homing and [RFC8317] for leaf-to-leaf BUM filtering. The ESI-label value is encoded in the high-order 20 bits of the Source-ID field.

The egress NVEs that make use of ESIs in the data path because they have a local multi-homed ES or support [RFC8317] SHOULD advertise their Ethernet A-D per-ES routes along with the Geneve tunnel sub-TLV in addition to the ESI-label Extended Community. The ingress NVE can then use the Ethernet option-TLV when sending Geneve packets based on the [RFC7432] and [RFC8317] procedures. The egress NVE will use the Source-ID field in the received packets to make filtering decisions.

Note that [RFC8365] modifies the [RFC7432] split-horizon procedures for NVO3 tunnels using the "local-bias" procedure. "Local-bias" relies on tunnel IP source address checks (instead of ESI-labels) to determine whether a packet can be forwarded to a local ES.

While "local-bias" MUST be supported along with Geneve encapsulation, the use of the Ethernet option-TLV is RECOMMENDED to follow the same procedures used by EVPN MPLS.

An ingress NVE using ingress replication to flood BUM traffic MUST send B=1 in all the Geneve packets that encapsulate BUM frames. An egress NVE SHOULD determine whether a received packet encapsulates a BUM frame based on the B bit. The use of the B bit is only relevant to Geneve packets with Protocol Type 0x6558 (Bridged Ethernet).

5. BGP Extensions

As per [RFC8365] the BGP Encapsulation extended community defined in [RFC9012] is included with all EVPN routes advertised by an egress NVE.

This document uses the Geneve Encapsulation BGP Tunnel Encapsulation Typei from the IANA BGP Tunnel Encapsulation Types registry, Value = 19.

5.1. Geneve Tunnel Option Types sub-TLV

The Geneve tunnel option types is a new BGP Tunnel Encapsulation Attribute Sub-TLV.

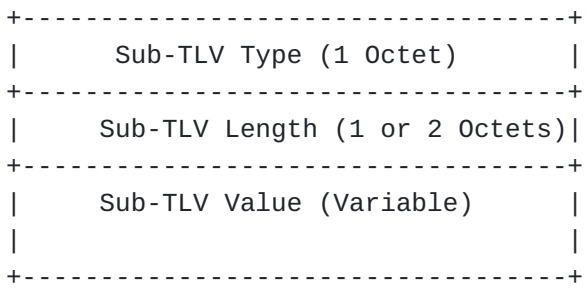

Figure 3: Geneve tunnel option types sub-TLV

The Sub-TLV Type field contains a value in the range from 192-252. To be allocated by IANA.

Sub-TLV value MUST match exactly the first 4-octets of the option TLV format. For instance, if we need to signal support for two option TLVs:

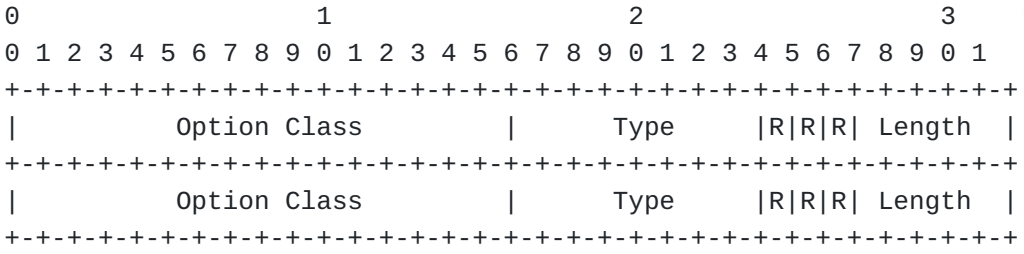

Figure 4: Geneve Option TLVs

An NVE receiving the above sub-TLV, MUST send Geneve packets to the originator NVE with only the option TLVs the receiver NVE is capable of receiving, and following the same order.

The above sub-TLV(s) MAY be included with only Ethernet A-D per-ES routes.

6. Operation

The following figure shows an example of an NVO3 deployment with EVPN.

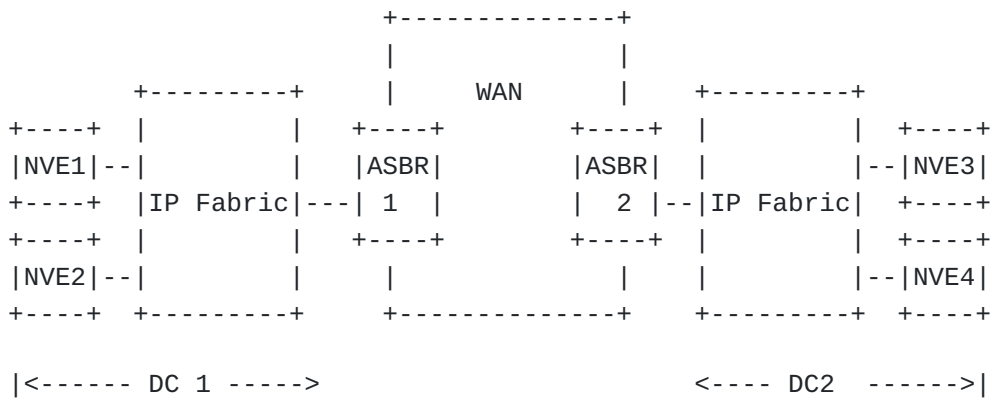

Figure 5: Data Center Interconnect with ASBR

iBGP sessions are established between NVE1, NVE2, ASBR1, possibly via a BGP route-reflector. Similarly, iBGP sessions are established between NVE3, NVE4, ASBR2.

eBGP sessions are established among ASBR1 and ASBR2.

All NVEs and ASBRs are enabled for the EVPN SAFI and exchange EVPN routes. For inter-AS option B, the ASBRs re-advertise these routes with NEXT_HOP attribute set to their IP addresses as per [RFC4271].

NVE1 sets the BGP Encapsulation extended community defined in all EVPN routes advertised. NVE1 sets the BGP Tunnel Encapsulation Attribute Tunnel Type to Geneve tunnel encapsulation, and sets the Tunnel Encapsulation Attribute Tunnel sub-TLV for the Geneve tunnel option types with all the Geneve option types it can transmit and receive.

All other NVE(s) learn what Geneve option types are supported by NVE1 through the EVPN control plane. In the datapath, NVE2, NVE3 and NVE4 MUST only encapsulate overlay packets with the Geneve option TLV(s) that NVE1 is capable of receiving, and in case more than one option TLV is being used, they MUST be in the order specified by NVE1.

A PE advertises the BGP Encapsulation extended community defined in [RFC5512] if it supports any of the encapsulations defined in [RFC8365]. A PE advertises the BGP Tunnel Encapsulation Attribute defined in [RFC9012] if it supports Geneve encapsulation, setting the type to Geneve Encapsulation.

7. Security Considerations

The mechanisms in this document uses EVPN control plane as defined in [RFC7432]. Security considerations described in [RFC7432] are equally applicable.

This document uses IP-based tunnel technologies to support data plane transport. Security considerations described in [RFC7432] and in [RFC8365] are equally applicable.

8. IANA Considerations

IANA is requested to assign a new option class from the "Geneve Option Class" registry for the Ethernet option TLV.

Option Class	Description	Reference
XXXX	Ethernet option	This document

IANA is requested to assign a new BGP Tunnel Encapsulation Attribute Sub-TLV from the BGP Tunnel Encapsulation Attribute Sub-TLVs registry.

BGP Tunnel Attribute Sub-TLV	Description	Reference
XXXX	Geneve tunnel option type	This document

9. Acknowledgements

The authors wish to thank T. Sridhar, for his input, feedback, and helpful suggestions.

10. References

10.1. Normative References

[I-D.ietf-nvo3-encap] Boutros, S. and D. E. Eastlake, "Network Virtualization Overlays (NVO3) Encapsulation Considerations", Work in Progress, Internet-Draft, draft-ietf-nvo3-encap-09, 7 October 2022, <https://datatracker.ietf.org/doc/html/draft-ietf-nvo3-encap-09>.

[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>.

[RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, DOI 10.17487/RFC4271, January 2006, <https://www.rfc-editor.org/info/rfc4271>.

[RFC5512] Mohapatra, P. and E. Rosen, "The BGP Encapsulation Subsequent Address Family Identifier (SAFI) and the BGP

Tunnel Encapsulation Attribute", RFC 5512, DOI 10.17487/RFC5512, April 2009, <https://www.rfc-editor.org/info/rfc5512>.

[RFC7432] Sajassi, A., Ed., Aggarwal, R., Bitar, N., Isaac, A., Uttaro, J., Drake, J., and W. Henderickx, "BGP MPLS-Based Ethernet VPN", RFC 7432, DOI 10.17487/RFC7432, February 2015, <https://www.rfc-editor.org/info/rfc7432>.

[RFC8317] Sajassi, A., Ed., Salam, S., Drake, J., Uttaro, J., Boutros, S., and J. Rabadan, "Ethernet-Tree (E-Tree) Support in Ethernet VPN (EVPN) and Provider Backbone Bridging EVPN (PBB-EVPN)", RFC 8317, DOI 10.17487/RFC8317, January 2018, <https://www.rfc-editor.org/info/rfc8317>.

[RFC8365] Sajassi, A., Ed., Drake, J., Ed., Bitar, N., Shekhar, R., Uttaro, J., and W. Henderickx, "A Network Virtualization Overlay Solution Using Ethernet VPN (EVPN)", RFC 8365, DOI 10.17487/RFC8365, March 2018, <https://www.rfc-editor.org/info/rfc8365>.

[RFC8926] Gross, J., Ed., Ganga, I., Ed., and T. Sridhar, Ed., "Geneve: Generic Network Virtualization Encapsulation", RFC 8926, DOI 10.17487/RFC8926, November 2020, <https://www.rfc-editor.org/info/rfc8926>.

[RFC9012] Patel, K., Van de Velde, G., Sangli, S., and J. Scudder, "The BGP Tunnel Encapsulation Attribute", RFC 9012, DOI 10.17487/RFC9012, April 2021, <https://www.rfc-editor.org/info/rfc9012>.

10.2. Informative References

[RFC7365] Lasserre, M., Balus, F., Morin, T., Bitar, N., and Y. Rekhter, "Framework for Data Center (DC) Network Virtualization", RFC 7365, DOI 10.17487/RFC7365, October 2014, <https://www.rfc-editor.org/info/rfc7365>.

Authors' Addresses

Sami Boutros (editor)
Ciena
United States of America

Email: sboutros@ciena.com

Ali Sajassi
Cisco Systems
United States of America

Email: sajassi@cisco.com

John Drake
Juniper Networks
United States of America

Email: jdrake@juniper.net

Jorge Rabadan
Nokia
United States of America

Email: jorge.rabadan@nokia.com

Sam Aldrin
Google
United States of America

Email: aldrin.ietf@gmail.com