BESS Workgroup   S. Boutros, Ed.
Internet-Draft   Ciena
Intended status: Informational   A. Sajassi
Expires: December 7, 2021   Cisco Systems
                                                            J. Drake
                                                     Juniper Networks
                                                          J. Rabadan
                                                               Nokia
                                                           S. Aldrin
                                                              Google
                                                        June 5, 2021

                       EVPN control plane for Geneve
                       draft-ietf-bess-evpn-geneve-02

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.

Copyright Notice

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. The NVO3 working group has evaluated different
   dataplane encapsulations. The Generic Network Virtualization
   Encapsulation [I-D.ietf-nvo3-geneve] has been recently recommended to
   be the proposed standard for NVO3 encapsulation.

   This document describes how the EVPN control plane can signal Geneve
   encapsulation type in the BGP Tunnel Encapsulation Extended Community
   defined in [I-D.ietf-idr-tunnel-encaps]. In addition, this document
   defines how to communicate the Geneve tunnel option types in a new

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 [I-D.ietf-nvo3-geneve].

[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 from, or sending
to, 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.

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 [I-D.ietf-nvo3-geneve]
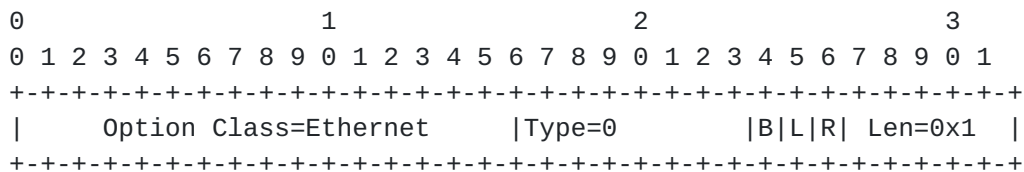encapsulation that are relevant to the operation of EVPN.

4.1. Ethernet option TLV

[I-D.ietf-bess-evpn-overlay] 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 traffic type. This is required to avoid
transient packet duplication in all-active multi-homing scenarios.
For GENEVE, we need a bit to for this purpose.

[RFC8317] uses MPLS label for leaf indication of BUM traffic
originated from a leaf 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 ecnapsulations such as
VxLAN, GPE, NVGRE, and GENVE, is local-bias as defined in section
8.3.1 of [I-D.ietf-bess-evpn-overlay], 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 GENVE 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):

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|    Option Class=Ethernet     |Type=0      |B|L|R| Len=0x1   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

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

```
        0                   1                   2                   3
        0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
       |     Option Class=Ethernet    |Typ=EVPN-OPTION|B|L|R| Len=0x2 |
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
       |     Rsvd   |              Source-ID                          |
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```
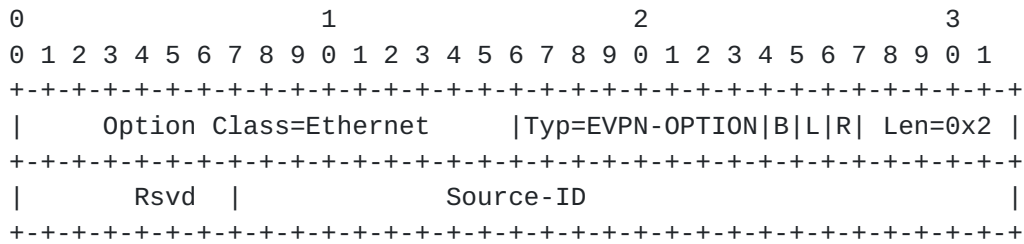
Where:

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

- Type is set to EVPN-OPTION (new type requested to IANA) 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-ESI 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
and 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 [I-D.ietf-bess-evpn-overlay] 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 [I-D.ietf-bess-evpn-overlay] the BGP Encapsulation extended
community defined in [I-D.ietf-idr-tunnel-encaps] is included with
all EVPN routes advertised by an egress NVE.

This document specifies a new BGP Tunnel Encapsulation Type for
Geneve and a new Geneve tunnel option types sub-TLV as described
below.

5.1. Geneve Tunnel Option Types sub-TLV

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

```
            +-----------------------------------+
            |      Sub-TLV Type (1 Octet)       |
            +-----------------------------------+
            |     Sub-TLV Length (1 or 2 Octets)|
            +-----------------------------------+
            |     Sub-TLV Value (Variable)      |
            |                                   |
            +-----------------------------------+
```
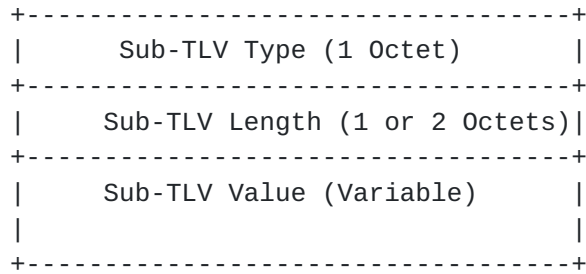
        Figure 1: Geneve tunnel option types sub-TLV

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:

```
   0                   1                   2                   3
   0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |          Option Class         |      Type     |R|R|R| Length  |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |          Option Class         |      Type     |R|R|R| Length  |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```
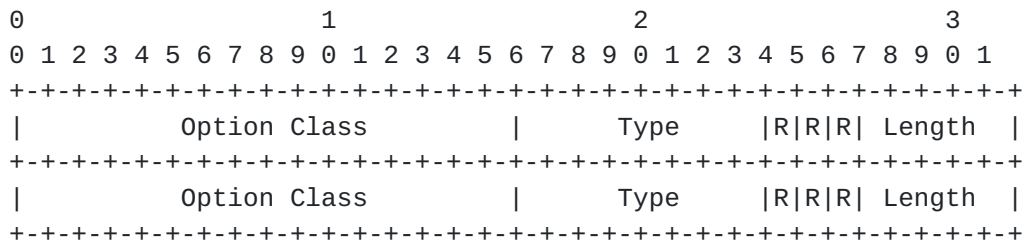
An NVE receiving the above sub-TLV, MUST send GENEVE packets to the
originator NVE with only the option TLVs the receiver NVE is capable
of receiving, and following the same order.

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.

```
                             +--------------+
                             |              |
             +---------+     |    WAN       |     +---------+
    +----+   |         |     +----+         +----+   |         |   +----+
    |NVE1|--|         |     |ASBR|         |ASBR|   |         |--|NVE3|
    +----+   |IP Fabric|---| 1  |         |  2 |--|IP Fabric|   +----+
    +----+   |         |     +----+         +----+   |         |   +----+
    |NVE2|--|         |     |              |         |         |--|NVE4|
    +----+   +---------+     +--------------+     +---------+   +----+

    |<------ DC 1 ----->                           <---- DC2   ------>|
```
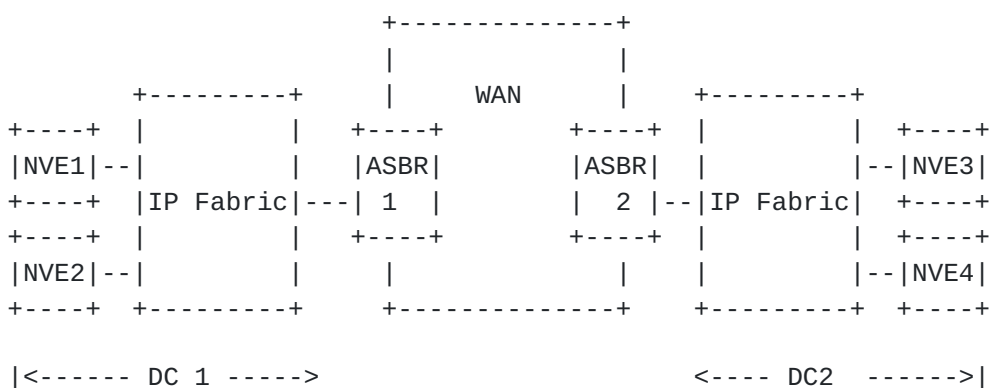
                    Figure 2: Data Center Interconnect with ASBR

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
[I-D.ietf-bess-evpn-overlay]. A PE advertises the BGP Tunnel
Encapsulation Attribute defined in [I-D.ietf-idr-tunnel-encaps] if it
supports Geneve encapsulation.

7. Security Considerations

The mechanisms in this document use 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
[I-D.ietf-bess-evpn-overlay] are equally applicable.

8. IANA Considerations

IANA is requested to allocate the following:

   BGP Tunnel Encapsulation Attribute
     Tunnel Type:

     XX   Geneve Encapsulation

BGP Tunnel Encapsulation Attribute Sub-TLVs a Code point from the
range of 192-252 for Geneve tunnel option types sub-TLV.

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

     Option Class   Description
     ------------   --------------
        XXXX   Ethernet option

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-bess-evpn-overlay]
          Sajassi, A., Drake, J., Bitar, N., Shekhar, R., Uttaro,
          J., and W. Henderickx, "A Network Virtualization Overlay
          Solution Using Ethernet VPN (EVPN)", draft-ietf-bess-evpn-
          overlay-12 (work in progress), February 2018.

[I-D.ietf-idr-tunnel-encaps]
          Patel, K., Velde, G. V. D., Sangli, S. R., and J. Scudder,
          "The BGP Tunnel Encapsulation Attribute", draft-ietf-idr-
          tunnel-encaps-22 (work in progress), January 2021.

[I-D.ietf-nvo3-encap]
          Boutros, S., "NVO3 Encapsulation Considerations", draft-
          ietf-nvo3-encap-05 (work in progress), February 2020.

[I-D.ietf-nvo3-geneve]
          Gross, J., Ganga, I., and T. Sridhar, "Geneve: Generic
          Network Virtualization Encapsulation", draft-ietf-
          nvo3-geneve-16 (work in progress), March 2020.

[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>.

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
   USA

   Email: sboutros@ciena.com

   Ali Sajassi
   Cisco Systems
   USA

   Email: sajassi@cisco.com

   John Drake
   Juniper Networks
   USA

   Email: jdrake@juniper.net

   Jorge Rabadan
   Nokia
   USA

   Email: jorge.rabadan@nokia.com

   Sam Aldrin
   Google
   USA

   Email: aldrin.ietf@gmail.com