INTERNET-DRAFT   Sami Boutros
Intended Status: Standard Track   VMware
                                                          Ali Sajassi
                                                       Cisco Systems
                                                          John Drake
                                                     Juniper Networks
                                                       Jorge Rabadan
                                                               Nokia

Expires: August 17, 2018   February 13, 2018

                      EVPN control plane for Geneve
                    draft-boutros-bess-evpn-geneve-01.txt

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 in NVO3
   solutions. EVPN control plane can be used by a Network Virtualization
   Endpoints (NVEs) to express as well what Geneve tunnel option TLV(s)
   that they can transmit and/or receive.

Table of Contents

1 Introduction

The Network Virtualization over Layer 3 (NVO3) develop solutions for
network virtualization within a data center (DC) environment that
assumes 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 have been
working on different dataplane encapsulations. The Generic Network
Virtualization Encapsulation [GENEVE] have been recently recommended
to be the proposed standard for network virtualization overlay
encapsulation.

This document describes how the EVPN control plane can signal Geneve
encapsulation type in the BGP Tunnel Encapsulation Extended Community
defined in [TUNNEL-ENCAP]. 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 [GENEVE].

[DT-ENCAP] recommends that a control plane determines how Network
Virtualization Edge devices (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 a Network Virtualization
Edge (NVE) to express what Geneve option TLV types it is capable to
receive or to send over the Geneve tunnel to 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.

1.1 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 RFC 2119 [RFC2119].

Most of the terminology used in this documents comes from [RFC7432]
and [NVO3-FRWK].

   NVO3: Network Virtualization Overlay 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.

2. GENEVE extensions

This document adds some extensions to the [GENEVE] encapsulation that
are relevant to the operation of EVPN.

2.1 Source Ethernet Segment option TLV

[RFC7432] defines the use of ESI-labels on egress NVEs that are part
of multi-homed Ethernet Segments (ES), so that they can determine if
the received packet was originated from an ES that is also locally
defined. Based on the ESI-label, the egress NVE can apply split-
horizon for the local ES'es and avoid packet duplication in all-
active multi-homing. In single-active multi-homing, packet
duplication can also happen during transient times, therefore the use
of ESI-labels and split-horizon checks are strongly recommended too.

In addition, [EVPN-ETREE] makes use of the ESI-labels to indicate BUM
leaf AC originated packets, so that the egress NVEs can filter BUM
traffic between leaf ACs.

In order to apply existing EVPN data plane split-horizon and E-Tree
filtering procedures, this document describes a new option TLV for
GENEVE:

```
 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=EVPN         | Type=ESI    |R|R|R| Length  |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

```
        |      Rsvd    |                 Source-ESI                |
        +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Where:
- Option Class is set to EVPN (new Option Class requested to IANA)
- Type is set to ESI (new type requested to IANA) and C bit must be set.
- Source-ESI 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 [EVPN-ETREE] for leaf-to-leaf BUM filtering.

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

Note that [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.
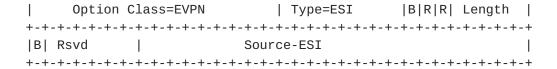
While "local-bias" MUST be supported along with GENEVE encapsulation, the use of the EVPN/ESI option-TLV is RECOMMENDED in cases where local-bias does not work, for instance: single-active multi-homing, multi-homing ES'es in Inter-AS option B scenarios and EVPN-ETREE.

2.2 BUM bit for Bridge Ethernet Protocol Type

As described in [EVPN-OVERLAY], in EVPN, 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 packet encapsulates an ingress replicated BUM frame. This is required to avoid transient packet duplication in all-active multi-homing scenarios.

This document defines the B bit as the ingress-replicated BUM traffic indication and encodes it in Source Ethernet Segment option TLV.

```
     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=EVPN        | Type=ESI     |B|R|R| Length  |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|B| Rsvd      |              Source-ESI                       |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Where:
    - Option Class is set to EVPN.
    - Type is set to ESI.
    - B bit is set to 1 for BUM traffic.
    - Source-ESI is set to 0 or to a value identifying the ingress
    NVE.

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).

3. BGP Extensions

As per [EVPN-OVERLAY] the BGP Encapsulation extended community
defined in [TUNNEL-ENCAP] 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.

3.1 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)     |
        |                                  |
        +----------------------------------+
```
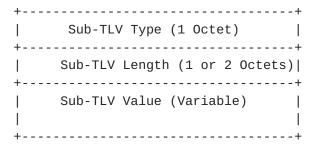
        Figure 1: Geneve tunnel option types sub-TLV

   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  |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

   Where, an NVE receiving the above sub-TLV, will send GENEVE packets
   to the originator NVE with with only the option TLVs the receiver NVE
   is capable of receiving, and following the same order. Also the high
   order bit in the type, is the critical bit, MUST be set accordingly.

4. 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  ------>|
```
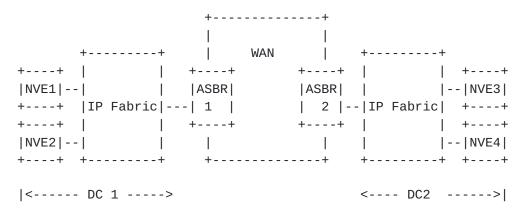
                Figure 2: Data Center Interconnect with ASBR

   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
only encapsulate overlay packets with the Geneve option TLV(s) that
NVE1 is capable of receiving.

[TBD] How do we propagate the signaling of GENEVE option-TLVs end to
end using other encaps?

5. 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
[EVPN-OVERLAY] are equally applicable.

6. 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 Source Ethernet Segment option TLV.

    Option Class   Description
    ------------   ------------------------------

      XXXX   Source Ethernet Segment option

7. Acknowledgements

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

8. References

8.1 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, <http://www.rfc-editor.org/info/rfc2119>.

[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, <http://www.rfc-editor.org/info/rfc7432>.

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

[RFC5226] Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 5226, May 2008, <http://www.rfc-editor.org/info/rfc5226>.

[GENEVE] Gross, et al. "Geneve: Generic Network Virtualization Encapsulation", draft-ietf-nvo3-geneve-05, work in progress, September, 2017.

[DT-ENCAP] Boutros, et al. "NVO3 Encapsulation Considerations", draft-ietf-nvo3-encap-01, work in progress, October, 2017.

8.2 Informative References

[NVO3-FRWK] Lasserre et al., "Framework for DC Network Virtualization", RFC 7365, October 2014.

[TUNNEL-ENCAP] Rosen et al., "The BGP Tunnel Encapsulation Attribute", draft-ietf-idr-tunnel-encaps-07, work in progress, July, 2017.

[EVPN-OVERLAY] Sajassi-Drake et al., "A Network Virtualization Overlay Solution using EVPN", draft-ietf-bess-evpn-overlay-10.txt, work in progress, December, 2017

Authors' Addresses

    Sami Boutros
    VMware, Inc.
    Email: sboutros@vmware.com

    Ali Sajassi
    Cisco
    Email: sajassi@cisco.com

    John Drake
    Juniper Networks
    Email: jdrake@juniper.net

    Jorge Rabadan
    Nokia
    Email: jorge.rabadan@nokia.com>