bess Z. Zhang
Internet-Draft Juniper Networks
Intended status: Standards Track K. Patel
Expires: 11 January 2023 Arrcus
10 July 2022

Transporting IP/UDP Payload-only in VPNs
draft-zzhang-bess-ipvpn-payload-only-00

Abstract

This document specifies an option for IP-VPN to transport IP/UDP payload only, without transporting IP/UDP headers, which are removed by an ingress PE and re-added by an egress PE.

Status of This Memo

Copyright Notice

Table of Contents

1. Introduction

Consider the following 5G network [_3GPP-23.501]:

```
     gNB1 ---\
     gNB2 ---- PE1 ----- PE2 --- UPF
     gNB3 ---/     \
                    ---- PE3 --- UPF2
```

Where gNB and UPF are 5G Network Function (NF) elements [3GPP-23.501]. They are IPv4 or IPv6 hosts connected via an IPVPN over an IPv6 transport infrastructure (it is believed that only IPv6 can scale to the requirements of 5G transport network but that's outside the scope of this document).

Per 3GPP specifications, the gNBs and UPF communicate over GPRS Tunnelling Protocol (GTP) tunnels, whose encapsulation includes (IP, UDP, GTP) headers. Some operators prefer IPv6/SRv6 [RFC8986] data plane instead of MPLS, so when PE1 receives the GTP traffic from gNBx, by default it would impose another IPv6 header and send to PE2.

There have been proposals to replace GTP with SRv6 tunnels for the following benefits:

* Traffic Engineering (TE) and Service Function Chaining (SFC) capability provided by SRv6

* Bandwidth savings because UDP and GTP headers are no longer needed

While 3GPP has not adopted the proposal, and GTP can be transported over SRv6 (instead of being replaced by SRv6), some operators still prefer to replace GTP with SRv6 "under the hood". That is, while RAN/UPF still use 3GPP signaling, the actual tunnel are no longer GTP but SRv6 based on GTP parameters signaled per 3GPP specifications.

The SRv6 tunnel could be between two NFs, or a GW (e.g. PE1/PE2) could be attached to an NF that still use traditional GTP and the GW will convert GTP to/from SRv6. This is specified in [I-D.ietf-dmm-srv6-mobile-uplane].

With that approach, the GTP information is encoded in SRv6 destination address and no additional IPv6 header is added by the PEs either. It is important to point out that when the GW is used, the original IP payload is reconstructed (UDP/GTP header removed or added and IPv6 header modified).

In this particular scenario, the reconstruction of IP payload is acceptable to some operators because they control all the network/host elements. With that premise, if an operator prefers MPLS data plane, only the GTP header and its payload (i.e., the UDP payload after UDP header) need to be transported - the IP/UDP header is removed or added by the PEs. Compared with SRv6 based approach, it is even more bandwidth efficient (no need for a minimum 40-byte IPv6 header) and SR-MPLS can also provide TE/SFC capabilities.

For example, PE1 can advertise a label for a (source, destination, IP/UDP payload type) tuple with the local semantics being that incoming traffic will be encapsulated in an IP or IP+UDP header and then routed out. When PE2 receives IP or IP+UDP traffic from the UPF, if there is a label for the corresponding (source, destination, IP/UDP payload type) tuple, it removes the IP or IP/UDP headers and simply transport the remaining payload. In this 5G scenario, it is still GTP - just that the IP/UDP headers are not present between PE1 and PE2.

The processing logic/burden and hardware requirement on PE1 and PE2 for the adding/removing of the IP/UDP header are not different from the "endpoint behaviors" required in the SRv6 approach even though they’re not called "End.xyz".

While this is inspired by the 5G GTP use case, the solution is generally applicable wherever it is acceptable to reconstruct the IP/UDP payload.

2. Specifications

2.1. Signaling Encoding

A new SAFI value TBD is to be assigned for the signaling of transporting IPVPN traffic w/o IP/UDP headers. The NLRI for this SAFI is encoded per [RFC3107], where the prefix consists of an 8-octet RD followed by an 4-octet IPv4 or 16-octet IPv6 host prefix (depending on the AFI) that is referred to as the DST Address, and optionally some or all of (SRC Prefix, DST UDP, SRC UDP).

```
+------------------------------------+
|   Length (1 octet)                 |
+------------------------------------+
|   Label (3 octets)                 |
+------------------------------------+
|   RD (8 octets)                    |
+------------------------------------+
|   DST Address (4/16 octets)        |
+------------------------------------+
|   SRC Prefix (optional/variable)   |
+------------------------------------+
|   DST UDP (optional, 2 octets)     |
+------------------------------------+
|   SRC UDP (optional, 2 octets)     |
+------------------------------------+
```

The DST UDP is present only if the SRC Prefix is a host prefix, which can be determined from the fact that the total NLRI length is at least 176 bits (22 octets, for IPv4) or 368 bits (46 octets, for IPv6).

The SRC UDP is present only if DST UDP is also present, which means that the total NLRI length is 192 bits (24 octets, for IPv4) or 384 bits (48 octets, for IPv6).

BGP updates with the TBD SAFI carry Route Targets as with the SAFI 128 (MPLS-labeled VPN address) updates so that corresponding forwarding state can be installed into appropriate VRFs. In fact, the SAFI 128 NLRI can be viewed as a special case of the NLRI for the new TBD SAFI (the DST UDP and SRC UDP are not present), though customer traffic is forwarded without any alteration with signaling via SAFI 128.

2.2. Control Word

To prevent intermediate LSRs that do ECMP hashing based on 5-tuple from mistaking the payload after the label as an IP packet, a Control Word SHOULD be used, with its first nibble set to 0. Other parts of the Control Word may be specified for various purposes.

Use of Control Word can be signaled via an Extended Community. Details to be added later.

2.3. Procedures

When a PE receives a route of the TBD SAFI, it installs forwarding state into corresponding VRFs according to the Route Targets [RFC4364].

The forwarding state causes traffic matching the (DST Address, SRC prefix, DST UDP, SRC UDP) in the NLRI to be forwarded using the label in the NLRI, with with IP or IP+UDP header removed.

As an implementation example, the forwarding state can be routes with prefix being concatenated (DST Address, SRC prefix, 17, DST UDP, SRC UDP), where 17 is the IP Protocol number for UDP and others are from the NLRI (any trailing part can be omitted and that is indicated by the prefix length). Longest-match using a concatenated (DST address, SRC address, Next Header, DST UDP, SRC UDP) key constructed from the packets is done, to produce a forwarding "nexthop" that:

- Removes IP or IP+UDP header and optionally prepends a control word

- Forward resulting payload using the label in the NLRI, over a base "tunnel" that is resolved for the Protocol Next Hop

This is just like how IPVPN works, except the matching of additional fields against a longer prefix and the removal of IP or IP+UDP header.

Specifically:

- If the SRC UDP is present, all fields MUST match. Both IP and UDP headers are removed.

- Otherwise, if the DST UDP is present, all (DST address, SRC address, DST UDP) MUST match. Both IP and UDP headers are removed.

- Otherwise, only (DST address, SRC address) need to match (longest), and only IP header is removed.

For a PE that originates a route of the TBD SAFI, the following apply:

- A distinct label MUST be allocated for each route.

- If the SRC UDP is encoded in the NLRI, incoming traffic with the label MUST be encapsulated (after removing the label and Control Word) with an IP and UDP header whose addresses and UDP ports MUST be set according to the NLRI.

- If the SRC UDP is not but DST UDP is encoded in the NLRI, then the added UDP header MUST have its source port set to a locally configured value.

- Otherwise, incoming traffic with the label (after removing the label and Control World) MUST be encapsulated with an IP header with its DST Address set as encoded in the NLRI. If the SRC Prefix in the NLRI is a host address, the added IP header's SRC Address is set accordingly, otherwise it MUST be set to a locally configured address.

The routes are attached with Route Targets to control the propagation and importation. They MAY be the same RTs for the relevant VPN, or MAY be RTs that target specific VRFs. For example, if the Source Prefix is encoded and it is known that matching traffic comes from a known set of VRFs, then RTs for those VRFs can be used.

2.4. Usage in Global Table

The encodings and procedures specified above equally apply to Global Table (default/master routing instance) instead of limited to VRFs. Specifically, a RD can be provisioned for the Global Table just like for a VRF, whether it is an all-zero value or not, and whether each PE has its distinct RD or they all use the same RD for the Global Table. Similarly, RTs are used to control the propagation and importation of the routes for the new SAFI even for the Global Table.

3. Security Considerations

To be provided.

4. References

4.1. Normative References

[RFC3107] Rekhter, Y. and E. Rosen, "Carrying Label Information in BGP-4", RFC 3107, DOI 10.17487/RFC3107, May 2001, <https://www.rfc-editor.org/info/rfc3107>.

[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>.

4.2. Informative References

[I-D.ietf-dmm-srv6-mobile-uplane]
Matsushima, S., Filsfils, C., Kohno, M., Garvia, P. C., Voyer, D., and C. E. Perkins, "Segment Routing IPv6 for Mobile User Plane", Work in Progress, Internet-Draft, draft-ietf-dmm-srv6-mobile-uplane-21, 9 May 2022, <https://www.ietf.org/archive/id/draft-ietf-dmm-srv6-mobile-uplane-21.txt>.

[RFC8986] Filsfils, C., Ed., Camarillo, P., Ed., Leddy, J., Voyer, D., Matsushima, S., and Z. Li, "Segment Routing over IPv6 (SRv6) Network Programming", RFC 8986, DOI 10.17487/RFC8986, February 2021, <https://www.rfc-editor.org/info/rfc8986>.

[_3GPP-23.501]
"System architecture for the 5G System (5GS), V17.3.0", December 2021.

Authors' Addresses

Zhaohui Zhang
Juniper Networks

Email: zzhang@juniper.net

Keyur Patel
Arrcus

Email: keyur@arrcus.com