Workgroup: BESS WG
Published: 20 August 2020
Intended Status: Standards Track
Expires: 21 February 2021

Authors: Y. Wang   B. Song
         ZTE Corporation   ZTE Corporation

Context Label for MPLS EVPN

Abstract

EVPN is designed to provide a better VPLS service than [RFC4761] and [RFC4762], and EVPN indeed introduced many new features which couldn't be achieved in those old VPLS implementions. But EVPN didn't inherit all features of old VPLS, and a few issues arises for EVPN only.

Some of these issues can be imputed to the MP2P nature of EVPN labels. The PW label in old VPLS is a label for P2P VC, so it contains more context than a identifier in dataplane for it's VSI instance.But the EVPN label just identifies it's VSI instnace and it can't stand for the ingress PE in dataplane. So the following issues arises with MPLS EVPN service:

*MPLS EVPN statistics can't be done per ingress PE.

*MPLS EVPN can't support hub/spoke use case which the spoke PE can only connect to each other by the hub PE.

*MPLS EVPN can't support AR REPLICATOR.

*MPLS EVPN can't support anycast SR-MPLS tunnel on the SPE nodes.

This document introduces a compound label stack to take advantage of both P2P VC and MP2P evpn labels.

Status of This Memo

This Internet-Draft will expire on 21 February 2021.

Copyright Notice

Table of Contents

1. Terminology and Acronyms

This document uses the following acronyms and terms:

BUM - Broadcast, Unknown unicast, and Multicast.

CE - Customer Edge equipment.

PE - Provider Edge equipment.

OPE - Originating PE - the original Router of an EVPN route.

ORIP - Originating Router's IP address.

Root ORIP - The ORIP in the Route Key field of a Leaf A-D route is called as that Leaf A-D route's Root ORIP. The Route Key field is the "Route Type specific" field of an PMSI route for which that Leaf A-D route is generated. When that PMSI route is an IMET route, that Leaf A-D route's root ORIP is that IMET route's "Originating Router's IP Address".

Leaf ORIP - The "Originator's Addr" field of the "Route Type specific" field of a Leaf A-D route is called as that Leaf A-D route's Leaf ORIP.

Self ORIP - The Leaf ORIP of a Leaf A-D route is also called as that Leaf A-D route's "self ORIP".

PTA - PMSI Tunnel Attribute.

PTA label - The MPLS label field of PMSI Tunnel Attribute.

PTA flags - The flags field of PMSI Tunnel Attribute.

IR - Ingress Replication.

AR - Assisted Replication.

IR PTA - PMSI Tunnel Attribute with tunnel-type = IR.

AR PTA - PMSI Tunnel Attribute with tunnel-type = AR.

IRL - Ingress Replication List, the list for Ingress-Replication BUM packets forwarding.

LS - Label Space.

CL - Context-identifying Label, or Context Label, A "context label" is one label that identifies a label table in which the label immediately below the context label should be looked up. It is defined in [RFC5331] section 3.

CLS - Context Label-Space, a Label Space that is identified by a Context Label. In [RFC5331], it is called as "Context-specific Label Space".

CSL - Context-Specific Label, a MPLS label that is allocated in a Context Label Space (CLS) which is identified by a CL. Note that a CSL is totally different from a CL. A CL itself is typically allocated in the per-platform label space.

CSL-Entry - Context-Specific Label Entry, a specific MPLS label allocated in a specific Context Label Space (CLS).

EVI - EVPN Instance.

EVL - EVI label, a MPLS label identifies an EVPN Instance (EVI) or a MAC/IP/Prefix in that EVPN instance.

EVPN label - The MPLS label in an EVPN route. It includes ESI label and EVL.

Admin EVI - A EVPN Instance (EVI) that is responsible for specific signalling work for other EVIs.

EC - Extended Community

SR-TL - Segment Routing (SR) Tunnel Label (TL).

ILM - Incoming Label Map, which is defined in [RFC3031] section 3.11.

CLS ILM - CLS-Specific ILM, an ILM that is installed in a "Context-specific Label Space".

Transit LSP - A LSP whose label operation for its ILM label is Label Swapping.

IMET-H Route - An IMET route that is advertised from hub-PE to spoke-PEs.

IMET-S Route - An IMET route that is advertised from spoke-PE to hub-PE.

VCL - VC Label.

Context VC - Context-identifying VC, a VC that identifies the ingress PE (see Section 5.1) or egress PE (see Section 5.4) of a data packet.

2. Problem Statement

EVPN is designed to provide a better VPLS service than RFC4761/RFC4762, and EVPN indeed introduced many new features which couldn't be achieved in those old VPLS implemention.But EVPN didn't inherit all features of old VPLS, and a few issues arises for EVPN only.

Some of these issues can be imputed to the MP2P nature of EVPN labels. The PW label in old VPLS is a label for P2P VC, so it contains more context than an identifier in dataplane for it's VSI instance. But the EVPN label just identifies it's VSI instnace and it can't stand for the ingress PE in dataplane. So the following issues arises with MPLS EVPN service:

*MPLS EVPN statistics can't be done per ingress PE. All flows from remote PEs share the same statistics on egress PE, because they share the same EVPN label and the egress PE can't pick them out in the dataplane.

*MPLS EVPN can't support hub/spoke usecase, where the spoke PEs can only connect to each other through the hub PE. Especially when at least two of the spoke PEs are connected to a common route reflector.

*MPLS EVPN can't work as an AR-REPLICATOR. Because the AR-REPLICATOR will apply replication for the ingress AR-LEAF too. But a packet shoud not be sent back to the AR-LEAF where it is received from.

*MPLS EVPN SPE cannot make use of SR-MPLS anycast tunnel because the two SPEs of the anycast tunnel will assign different EVPN labels for the same EVPN route.

So this document introduces an compound label stack to take advantage of both P2P VC and MP2P EVPN labels.

3. Using VC Label to Add Context Data to EVPN Flows

In order to add as much context as old VPLS to EVPN data packet, We can construct an infrastructure by a full-mesh of context-VCs among the EVPN PEs.

Take the context-VCs between PE-i and PE-j as an example, VC-ij is the context-VC from PE-i to PE-j, and VC-ji is the context-VC from PE-j to PE-i. The VC-ij identifies the PE-i node on PE-j. The VC-ji identifies PE-j node on PE-i. The VC-label for VC-ij is called as L-ij, and the VC-label for VC-ji is called as L-ji.

So the PE-i can push the L-ij onto the EVPN data packet for PE-j to distinguish the packet of PE-i from other data packets. Because that the L-ij identifies the ingress PE of the data packet.

There are two styles of context-VC in this draft. One style is named as shared context-VC, the other style is named as per-EVI context VC.

3.1. The Shared Context VCs

The shared context-VCs are dedicated to identify the context for a data packet while the EVPN label still identifies the EVPN instance.

Note that typically a shared context VC can be shared by all EVPN instances between it's ingress PE and egress PE. In other words, we don't have to establish a dedicated mesh of context VCs for each specified EVPN service. So we called the shared context VCs as a common infrastructure for those EVPN services.

3.2. The per-EVI Context VCs

The per-EVI context VCs are used to identify both the context (typically the ingress-PE) and the EVPN instance for a data packet at the same time. In other words, we have to establish a dedicated set of per-EVI context VCs for each specified EVPN service.

4. Context VC Signalling Procedures

The IMET route per [RFC7432] have a corresponding route-type in MVPN. It is, in effect, the Intra-AS I-PMSI route per [RFC6514]. The difference between them is that an IMET route won't handle a responding Leaf A-D route, but an Intra-AS I-PMSI route will.

The Leaf A-D route per [I-D.ietf-bess-evpn-bum-procedure-updates] is required for per-EVI context VCs. In this draft, we use the Leaf A-D route with IR-PTA to establish per-EVI context-VCs. The Leaf A-D route is generated for an IMET route.

It is an update for [RFC7432]. The backward compatibility will be described in Section 4.4.

4.1. Construct Leaf A-D Route for IR

PE1 will construct a Leaf A-D route with IR-PTA for EVI1 in response to an IMET route R1 with IR-PTA. The IMET route R1 is received from PE2 previously. The key fields of the IMET route is included in the "Route Key" field of the Leaf A-D route (say R2) along with the ORIP of PE1 itself. We call the ORIP of PE1 itself as the Leaf A-D route's "self-ORIP" in order to distinguish it from the Leaf A-D route's Root-ORIP. So the "Route Type sepcific" field of the Leaf A-D route is per <EVI1, PE2> basis.

4.1.1. Advertising Per-platform VC Label

The MPLS label field in the IR-PTA of the Leaf A-D route is allocated per <EVI1, PE2> basis in per-platform label space on PE1. So the per-EVI context VC can identify the EVI1 too.

Note that PE1 may already advertise an IMET route R3 to PE2 before the advertisement of above Leaf A-D route.

4.1.2. Advertising Context-Specific VC label

Note that the per <EVI,Ingress PE> basis label allocation (see Section 4.1.1) may consume too many labels in per-platform label space. Sometimes we want to use the same EVPN label in all Leaf A-D routes and IMET routes of the same EVI. So we allocate a context-specific label (CSL) for a context VC in this section.

The EVPN label is still allocated from per-platform label space, and it identifies the EVPN instance as per [RFC7432]. But it also identifies a context label space CLS1. The VC label of the context VC is allocated in CLS1. So we say that the VC label is a context-specific VC label.

The encapsulation of the Context-Specific VC Label is illustrated as the following figure:

```
                +----------------------------------+
                |  underlay ethernet header        |
                +----------------------------------+
                |  PSN tunnel label                |
                +----------------------------------+
                |  EVPN label                      |
                +----------------------------------+
                |  Context VC Label                |
                +----------------------------------+
                |  overlay ethernet or IP header   |
                +----------------------------------+
```

Figure 1: Encapsulation of EVI-Specific VC Label for EVPN Payload

Note that the Context-VC Label here is not the Context Label Space (CLS) ID of the EVPN Label. But the EVPN label is the CLS-ID of the Context-VC Label. That's why the Context Label Space ID EC (see section 3.1 of [I-D.ietf-bess-mvpn-evpn-aggregation-label]) is not appropriate for such encapsulation.

We introduce a new BGP Extended Community called Context-specific Label (CSL) Entry Extended Community, the CSL-Entry EC was used to carry the downstream-assigned context-specific VC labels.

4.1.3. CSL-Entry Extended Community

CSL-Entry Extended Community is a new Transitive Opaque EC with the following structure (Sub-Type value to be assigned by IANA):

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| 0x03 or 0x43  |   Sub-Type    |      CSL-Type                 |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                        CSL-Value                              |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 2: CSL-Entry Extended Community

- CSL-Type: A 2-octet field that specifies the type of Context-Specific Label (CSL). In this document, the CSL-Type is 0, indicating that the CSL-Value field is a downstream-assigned label whose label space is identified by the EVPN label of the same route.

- CSL-Value: A 4-octet field that specifies the value of CSL-Entry. When it is a label (with CSL-Type 0), the most significant 20-bit is set to the label value.

The "sub-type" field of the CSL-Entry EC has a different codepoint from the CLS-ID EC. The CSL-Value of the CSL-Entry EC is a MPLS label in a context-specific label space identified by the PTA label. And the MPLS label in the CSL-Entry EC will be pushed onto the label stack before the PTA label by the ingress PE. Typically, the MPLS label of the CSL-Entry EC is a downstream assigned label, which means that it will be used as an outgoing label by the PE receiving the CSL-Entry EC, not as incomming label.

When constructing the Leaf A-D route, the IR-PTA label is the EVPN Label, as per [RFC7432]. But the CSL-value in the CSL-Entry EC is a label (say L1) that is allocated per TPE basis in CLS1. In fact, L1 is the context-specific VC label of the context VC of that ingress TPE. That's why the context VC is called as CSL-based context VC. So the CLS1-specific VC label need to be pushed onto the label stack before EVPN Label (which identifies CLS1) on ingress PEs.

Note that when the PTA label is changed to a new value (caused by the BGP nexthop rewriting) by the SPE nodes, the CSL-Entry in the same EVPN route won't be rewrite. This is similar to the behavior of ESI Label EC of EAD per ES route.

Note that when an ingress PE sends traffic, it imposes the CSL before it imposes the PTA label of the same EVPN route. That's the obvious difference from the Context Label Space ID EC (see section 3.1 of [I-D.ietf-bess-mvpn-evpn-aggregation-label]).

4.2. The Sharing of Context VCs

4.2.1. Using Admin-EVI

Note that the CSL-Entry ECs (for different EVIs) received from the same TPE may be the same label, because that all EVLs on the same PE may identify the same Context-specific Label Space (CLS). So we can select a single EVI to use the Leaf A-D route with CSL-Entry EC in such case. This EVI is called as administrating EVI (admin-EVI). The context VC label carried in the Leaf A-D routes of the admin-EVI will be used to take the place of the PTA label of the IMET route with the same ORIP in all other ordinary EVIs in such case.

Note that all other ordinary EVIs don't have to use the Leaf A-D routes with IR-PTA in their signalling procedures, they can use ordinary IMET routes instead. The admin-EVI need to be configured on all EVPN-PEs in such case.

4.2.2. Using per-platform VC labels

The admin EVIs may allocate the VC labels of its context VCs in the per-platform label space.

Even if these VC labels are allocated in the per-platform label space, The CSL-Entry EC is still necessary.

Because that the Context VC is not expected to be used to determine the forwarding path of the data packet. They just identify the context of the data packet (typically the ingress PE of the data packet). So they should not be rewritten even when the nexthop of the EVPN route is rewriten. When the Leaf routes use PTA label (instead of CSL-entry EC) to carry the Context VC Label, The PTA label will be rewritten on SPE nodes. It is just not our expectation.

Note that the SPEs don't have to recognize the CSL-Entry EC, because that it's a transitive opaque EC. The EVPN label of the admin-EVIs in the PTA is unuseful. But they should also be included so as to pass though the SPE nodes.

4.3. Establish Ingress Replication List by Leaf A-D Route

```
                         T1: ===R3(IMET)======>
                    PE1----------------------------PE2
                         T2: <==R1(IMET)=======
                         T3: ===R2(Leaf A-D)==>
```

Figure 3: The Leaf A-D Route's Ordinal Number

PE2 receives (at time T3) the responding Leaf A-D route (say R2) of the IMET route R1 which is previously (at time T2) advertised by itself, and PE2 preiously (at time T1) received an IMET route R3 whose ORIP is the same as the self-ORIP of R2 . Given that R1,R2 and R3 both have a IR-PTA, PE2 SHOULD use R2 to install the Ingress Replication List (IRL) item for PE1 instead, and R3 will not used to install the IRL-item for PE1 from then on.

Note that when R2 included a CSL-Entry EC, the CSL-value of the CSL-Entry EC will be used as the outgoing label of the IRL-item. The MPLS label of the IR-PTA will be used as the context label (CL) of the CSL-Entry in NHLFE. No ILM entry will be installed for the CSL of R2 on PE2.

Note that when PE2 sent R1 (at time T2) to other PEs(including PE1), it will set the LIR flag of R1 to one. At the same time, PE2 will add an import RT (say RT1) to the EVPN instance. When PE1 send R2 (at time T3) to PE2 as R1's reply, PE1 will add the same RT1 to R2, because PE1 assumes that RT1 should have been added to the EVPN instance by PE2 as an import RT. Otherwise PE2 wouldn't set the LIR flag to one during the advertisement of R1.

4.4. Backward Compatibility

In [RFC7432], the LIR flag of IMET route is required to be zero when it is advertised and to be ignored on receipt.

It means that the LIR flag is reserved by [RFC7432], but it technically can be used in the future. What should the LIR flag be restrained in the future use is no more severer than any other reserved PTA flags in the IMET routes.

So when PE2 set the LIR flag to one in the IMET route and send it to PE1, PE2 won't expect that the IMET route must be responded by a Leaf A-D. When the corresponding Leaf A-D route can't be received from PE1, the IMET route from PE1 still be used as per [RFC7432]. But when PE1 is a new PE following this draft, PE1 will indeed respond a Leaf A-D route for the IMET route.

The EVPN S-PMSI routes for (C-*,C-*) may be a succedaneum of IMET routes. Such EVPN S-PMSI route is called wildcard S-PMSI route in these document. The LIR flag of wildcard S-PMSI route is not reserved. So there won't be any controversial issues on the compatibilities of Leaf A-D route responsing. But we prefer the IMET routes, because the IMET routes with non-zero LIR flag won't conflict with [RFC7432] either.

5. Solutions

5.1. Solution for Source-Squelching in Hub-Spoke Scenarios

```
       PEs3--------RR2--------+
                              PEh------CE3
CE1----PEs1--------RR1--------+
                   /
CE2----PEs2-------/
```

Figure 4: Hub PE and Spoke PEs

Now take above use case for example, there are three spoke PEs and one hub PE. The spoke PEs are PEs1, PEs2 and PEs3. The hub PE is PEh. Two of the spoke PEs (PEs1 and PEs2) are connected to the same RR group and the third one connects to another RR group.

Although we can advertise different EVPN labels for different RR groups, we can't advertise different EVPN labels for PEs1 and PEs2 in the form of current IMET routes.

In order to advertise different EVPN labels for PEs1 and PEs2, PEh may use BGP ADD-PATH to advertise two NLRIs (of the same EVI) to RR1, each along with a different EVPN label.

But that approach will cause some issues. Because that the BGP ADD-PATH is typically used for ECMP purpose. But we don't expect RR1 to do ECMP for these two NLRIs. That's the first issue. The second issue is that we can't expect that PEs1/PEs2 will import (and install) one of (and exactly only one of) these two NLRIs, even if PEh adds some intended RTs along with these NLRIs. Because that such RT may be not so intended from the view point of PEs1/PEs2. The third issue is that when RR1 receives these two NLRIs, RR1 may just reflect one of them to its clients because that their nexthop is the same IP address.

Note that although one (that is for PEs1) of those two IMET-NLRIs can be advertised (by PEh) along with an IP-based Route Target Extended Community which is constructed by placing the IP address carried in the Next Hop (that is PEs1's node address) of the received (from PEs1) IMET route in the Global Administrator field of that EC (with the Local Administrator field of that EC set to 0).

When PEs1 receives that IMET-H route (say R5), PEs1 either won't consider R5 as an IMET route for itself (If PEs1 didn't configure its nexthop address as an import RT previously), or will import R5 to all (because that R5 doesn't have any service-delimiting information in it) of its EVPN Instances, or will import all (not R5 only) of those two IMET-H NLRIs into the corresponding EVI-instance (if R5's EVI RT are still constructed just following [RFC7432]).

In order to solve these problems, the follow three methods may be used:

M1) Make the IMET-H routes and the IMET-S routes of the same EVPN Instance use the same Route-distinguisher (RD). When an RT of an IMET-H route (which is enhanced by the "path identifer" of [RFC7911]) matches the import RT of an EVI but its RD doesn't match the EVI's RD, that IMET-H route won't be imported.

M2) Make the IMET-H route's Route Target carry the IMET-S route's RD. at the same time, the spoke PEs (auto-)configure the RD of each EVI as an import RT of the same EVI. Different spoke PEs should use different RDs in this method, and the "path identifer" of [RFC7911] should also be used in this method.

M3) Make the IMET-H route's NLRI carry the IMET-S route's NLRI. Such IMET-H route is actually another format of the Leaf A-D route for that IMET-S route.

But these three methods all have obvious defects, that's why we said that we can't advertise different EVPN labels for PEs1 and PEs2 in the form of current IMET routes.

But PEh can request PEs1 or PEs2 to push the label of the context VC from them to PEh. Benefit from the context VC label, PEh can distinguish where the packet from, in other words, PEh can decide where the packet can't be sent to.

The signaling for the hub PE to request the spoke PE to push the context VC label will be the CSL-Entry EC, regardless of the label-space type of the context VC label.

Note that although PEs1 and PEs2 can receive EVPN routes from each other, they won't import these routes because of the hub/spoke behaviors.

5.2. Solution for per ingress statistics

We use CSL-based per-EVI context-VCs(see Section 4.1.2) to do per-ingress statistics.

Note that The per-platform label space can be used as CLS1 at the same time. In such case, the inner context-VC label is similar to the downstream-assigned ESI-label in ILM-lookup behavior. Such context-VC is very similar to the shared context VC too.

Note that when PE1 sends a Leaf A-D route with a CSL-Entry EC to PE2, but PE2 don't recognize the CSL-Entry EC, then PE2 will encapsulate the EVPN label without the inner context-VC label. If CLS1 is actually identical to the per-platform label space, this will work as well as [RFC7432], although the per-ingress statistics can't be executed.

Note that legacy PEs will not send a Leaf A-D route in response to an IMET route even if the LIR flag in the IMET route is set to one. So when legacy PEs and new PEs following this section coexist in the same EVI, they can interwork well, but only the new PEs can do per-ingress statistics.

5.3. Solution for AR REPLICATOR in MPLS EVPN

```
                  LEAF1--------REPLICATOR1--------RNVE1
                                 /
                  LEAF2-----------/
```

Figure 5: AR REPLICATOR in MPLS EVPN

When REPLICATOR1 node recieves an IMET Route with AR-role = AR-LEAF from LEAF1 node, REPLICATOR1 SHOLD respond to it with an Leaf A-D route with AR-PTA. The MPLS label field of the AR-PTA (say AR-PTA Label) will be allocated following the same rules as the IR-PTA Label in Section 4.1. When AR-LEAF1 receives above Leaf A-D route, the Leaf A-D route is treated as a Replicator-AR route for the same ORIP, then the control-plane procedures follows [I-D.ietf-bess-evpn-optimized-ir]. When REPLICATOR1 receives data packets from the AR-PTA Label, REPLICATOR1 will do source-squelching for LEAF1, which means that these data packets will not be forwarded back to LEAF1.

Note that the old Replicator-AR route (from AR-REPLICATOR) which is in terms of IMET route will not be installed in dataplane by MPLS EVPN AR-LEAF. Because that the Leaf A-D route (from AR-RPELICATOR) will take it's place on AR-LEAF node. But the old Regular-IR route (from AR-REPLICATOR) should still be installed by MPLS EVPN AR-LEAFs.

As what is discussed in Section 4.4, the wildcard S-PMSI route can be used to replace the IMET route between AR-REPLICATOR and its AR-LEAFs.

5.4. Solution for anycast tunnel usage on SPE

```
                    /--------SPE1-------\
          CE1-----TPE1                 TPE2-----CE2
                    \--------SPE2-------/
```

Figure 6: SPE with Anycast Tunnel

Now take above use case for example, the two SPEs are the egress nodes of an anycast SR-MPLS tunnel. The anycast SR-MPLS tunnel is used to transport flows from TPE1 to either SPE1 or SPE2 according to load balancing procedures. So SPE1 and SPE2 have to advertise the same EVPN label independently for a given EVPN route.

5.4.1. Control-plane

When TPE2 send a MAC/IP advertisement route (say R8) to SPE1 and SPE2, a "Downstream Context-specific Label Space (CLS) ID Extended Community" can be included in R8 along with an EVPN label (say EVL4).

5.4.1.1. Downstream-CLS ID Extended Community

The downstream-CLS ID Extended Community is a new Transitive Opaque EC with the following structure (Sub-Type value to be assigned by IANA):

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| 0x03 or 0x43  |   Sub-Type    |O|     ID-Type                 |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                            ID-Value                           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 7: Downstream-CLS ID Extended Community

*ID-Type: A 2-octet field that specifies the type of Label Space ID. In this section, the ID-Type is 0. The ID-Type 0 indicating that the ID-Value field is a MPLS label in DCB, and it has global uniqueness across the EVPN domain.

- ID-Value: A 4-octet field that specifies the value of Label Space ID. When it is a label (with ID-Type 0), the most significant 20-bit is set to the label value.

- O bit: The OPE-Flag. When the Extended Community is advertised by the Originating PE of the corresponding EVPN route, The O bit should be set to 1, otherwise the O bit should be set (or rewritten) to 0.

Note that although the downstream-CLS ID EC is highly similar to the Context Label Space ID Extended Community (see section 3.1 of [I-D.ietf-bess-mvpn-evpn-aggregation-label]) in their encodings, they have absolutely different behaviors in data-plane. The CLS-ID EC should be treated as an incomming label in data-plane, but the downstream-CLS ID EC should be treated as an outgoing label in data-plane. So they couldn't share the same code-point in the signalling procedures.

5.4.1.2. Context-specific Label Swapping

When SPE1 and SPE2 receive R8 from TPE2, they should advertise R8 to TPE1 independently, and the next-hop of R8 should be changed to the common anycast node address (say IP_12) of SPE1 and SPE2 before the advertisement. But SPE1 and SPE2 can simply keep R8's EVPN label (the EVL4 from TPE2) unchanged.

The contex-VC label (say VCL4) in the "downstream-CLS ID EC" is also kept unchanged. But the O bit of the "downstream-CLS ID EC" is rewritten to 0.

Note that although the EVL4 and VCL4 is unchanged, a CLS-specific ILM whose label operation is "label swapping" should also be installed, because that the outgoing PSN tunnel information should be resolved.

Note that the two outgoing-labels of the label-swapping have the same value (EVL4 and VCL4) as the two incomming-labels. The VCL4 is an optional outgoing-label because that the O bit of its "Downstream-CLS ID EC" is 1.

5.4.2. Data-plane

The label stack on the anycast SR-MPLS tunnel is constructed by TPE1 as the following:

```
+---------------------------------+
|  underlay ethernet header       |
+---------------------------------+
|  Anycast SR-TL = SR_LSP_to_SPEs |
+---------------------------------+
|  Context-VC Label = VCL4        |
+---------------------------------+
|  EVPN label = EVL4              |
+---------------------------------+
|  overlay ethernet or IP header  |
+---------------------------------+
```

Figure 8: Anycast SPE dataplane

Note that the SR Tunnel Label (TL) in the label stack is the anycast SR-LSP label from TPE1 to the SPE1 or SPE2. And the VCL4 in the label stack is mandatory (from the viewpoint of TPE1) because that the O bit of its "Downstream-CLS ID EC" is 0.

Note that the context-VC is constructed (on SPE1 and SPE2) in per-platform label space, and VC labels from TPE2 to SPE1 and SPE2 will be the same value (VCL4). so the label stacks (from the viewpoint of TPE1) are the same for SPE1 and SPE2. That's why the anycast tunnel from TPE1 to SPE1 and SPE2 can be used for R8 by TPE1.

When SPE1/SPE2 receives that data packet, then SPE1/SPE2 will perform CLS-specific ILM lookup for the EVPN label in the "TPE2-specific label space" which is identified by the context-VC label VCL4. The label operation will be "swapping", and the new outgoing EVPN label will be the same value (as EVL4). Note that the optional (from the viewpoint of SPE1/SPE2) VCL4 is suggested to be absent in the label stack.

5.4.3. The Generating of Downstream-CLS ID EC on SPE

When TPE2 don't advertise the Downstream-CLS ID EC to SPE1 and SPE2, They have to generate that EC by themselves.

In such case, TPE2 should advertise the OPE TLV for R8. And a context-VC infrastructure should be established previously. The context-VC infrastructure should assure that the context-VCs from TPE2 to any other TPEs/SPEs have the same VCL value.

Then the SPE1 can set the ID-Value of the Downstream-CLS ID EC to the VCL of the contex VC from TPE2 to itself. The ID-Type of the Downstream-CLS ID EC is set to 0. The O bit of the Downstream-CLS ID EC is set to 0. So the same Downstream-CLS ID EC can be generated by the SPEs independently.

It is feasible for such context-VC infrastructure to be implemented on the basis of Kompella VPLS signalling or BGP SR signaling. But it will be better for the admin-EVI (as the context-VC infrastructure) and EVPN VPLS to use the same signalling framework.

So we can just transplant the SRGB or VE-Block configuration model into the admin-EVI but the admin-EVI still use the signalling framework of Section 5.4.1.

6. Security Considerations

This section will be added in future versions.

7. IANA Considerations

This document introduces two new Transitive Opaque Extended Communities "Downstream CLS ID Extended Community" and "Context-Specific Label Entry Extended Community". An IANA request will be submitted later for two code-points in the BGP Transitive Opaque Extended Community Sub-Types registry.

8. Acknowledgements

The authors would like to thank the following for their comments and review of this document:

Benchong Xu.

9. References

9.1. Normative References

[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>.

[RFC6514] Aggarwal, R., Rosen, E., Morin, T., and Y. Rekhter, "BGP Encodings and Procedures for Multicast in MPLS/BGP IP VPNs", RFC 6514, DOI 10.17487/RFC6514, February 2012, <https://www.rfc-editor.org/info/rfc6514>.

[I-D.ietf-bess-evpn-optimized-ir] Rabadan, J., Sathappan, S., Lin, W., Katiyar, M., and A. Sajassi, "Optimized Ingress Replication solution for EVPN", Work in Progress, Internet-Draft, draft-ietf-bess-evpn-optimized-ir-07, 13 July 2020, <https://tools.ietf.org/html/draft-ietf-bess-evpn-optimized-ir-07>.

[I-D.ietf-bess-evpn-bum-procedure-updates]
Zhang, Z., Lin, W., Rabadan, J., Patel, K., and A. Sajassi, "Updates on EVPN BUM Procedures", Work in Progress, Internet-Draft, draft-ietf-bess-evpn-bum-procedure-updates-08, 18 November 2019, <https://tools.ietf.org/html/draft-ietf-bess-evpn-bum-procedure-updates-08>.

9.2. Informative References

[RFC3031] Rosen, E., Viswanathan, A., and R. Callon, "Multiprotocol Label Switching Architecture", RFC 3031, DOI 10.17487/RFC3031, January 2001, <https://www.rfc-editor.org/info/rfc3031>.

[RFC5331] Aggarwal, R., Rekhter, Y., and E. Rosen, "MPLS Upstream Label Assignment and Context-Specific Label Space", RFC 5331, DOI 10.17487/RFC5331, August 2008, <https://www.rfc-editor.org/info/rfc5331>.

[I-D.ietf-bess-mvpn-evpn-aggregation-label]
Zhang, Z., Rosen, E., Lin, W., Li, Z., and I. Wijnands, "MVPN/EVPN Tunnel Aggregation with Common Labels", Work in Progress, Internet-Draft, draft-ietf-bess-mvpn-evpn-aggregation-label-03, 24 October 2019, <https://tools.ietf.org/html/draft-ietf-bess-mvpn-evpn-aggregation-label-03>.

Authors' Addresses

Yubao Wang
ZTE Corporation
No. 50 Software Ave, Yuhuatai Distinct
Nanjing
China

Email: wang.yubao2@zte.com.cn

Bing Song
ZTE Corporation
No. 50 Software Ave, Yuhuatai Distinct
Nanjing
China

Email: song.bing@zte.com.cn