IDR
Internet Draft
Intended status: Standards Track
Expires: January 24, 2016

Luyuan Fang
Deepak Bansal
Microsoft
Chandra Ramachandran
Juniper Networks
Fabio Chiussi

Nabil Bitar
Verizon
Yakov Rekhter

July 23, 2015

BGP-LU for HSDN Label Distribution
draft-fang-idr-bgplu-for-hsdn-02

Abstract

This document describes modifications of BGP Labeled Unicast (BGP-LU) procedures for label distribution in a partitioned network. Specifically, these procedures are suitable for building the Hierarchical SDN (HSDN) control plane for the hyper-scale Data Center (DC) and cloud networks.

Table of Contents

1. Introduction

This document describes modifications to BGP Labeled Unicast (BGP-LU)-based procedures for label distribution [RFC3107] in a partitioned network where a label stack is used for forwarding. Current BGP-LU procedures do not provide mechanisms for distributing and installing operator-assigned partition-scope labels.

Specifically, the modifications described in this document are suitable for label distribution in the control plane of a MPLS-based Hierarchical SDN (HSDN) Data Center (DC) and cloud network.

Hierarchical SDN (HSDN) [I-D.fang-mpls-hsdn-for-hsdc] is an architectural solution to scale a hyper-scale cloud consisting of DCs interconnected by a Data Center Interconnect (DCI) to tens of millions of physical underlay endpoints, while efficiently handling both Equal Cost Multi Path (ECMP) load-balanced traffic and any-to-any end-to-end Traffic Engineered (TE) traffic. The HSDN reference model, operation, and requirements are described in [I-D.fang-mpls-hsdn-for-hsdc].

HSDN is designed to allow the physical decoupling of control and forwarding, and have the LFIBs configured by a controller according to a full SDN approach. Such a controller-centric approach is described in [I-D.fang-mpls-hsdn-for-hsdc].

However, the HSDN control plane can also be built in a hybrid approach, using a routing or label distribution protocol to distribute the labels, together with a controller. This hybrid approach may be particularly useful during technology migration. This document specifies the use of BGP-LU for label distribution and LFIB configuration in the HSDN control plane.

In the HSDN architecture, the DC/DCI network is partitioned into hierarchical underlay partitions (UPs) such that the number of destinations in each UP does not increase beyond the limit imposed by capabilities of network nodes. Once the DC cloud has been partitioned to the desired configuration, the traffic from a source endpoint to a destination endpoint uses a stack of labels, one label per each level in the hierarchy, whose semantics indicate to the forwarding network nodes at each level which destination in its local UP should forward the packet to. The label semantics can also identify a specific path (or group of paths) in the UP, rather than simply a destination.

In other words, the label stack indirectly represents the UPs that the packet should traverse to reach the destination end device. More precisely, the outer label specifies the destination in the partition at the highest level that the packet should traverse, while the other

labels specify the destination in each partition that the packet traverse thereafter.

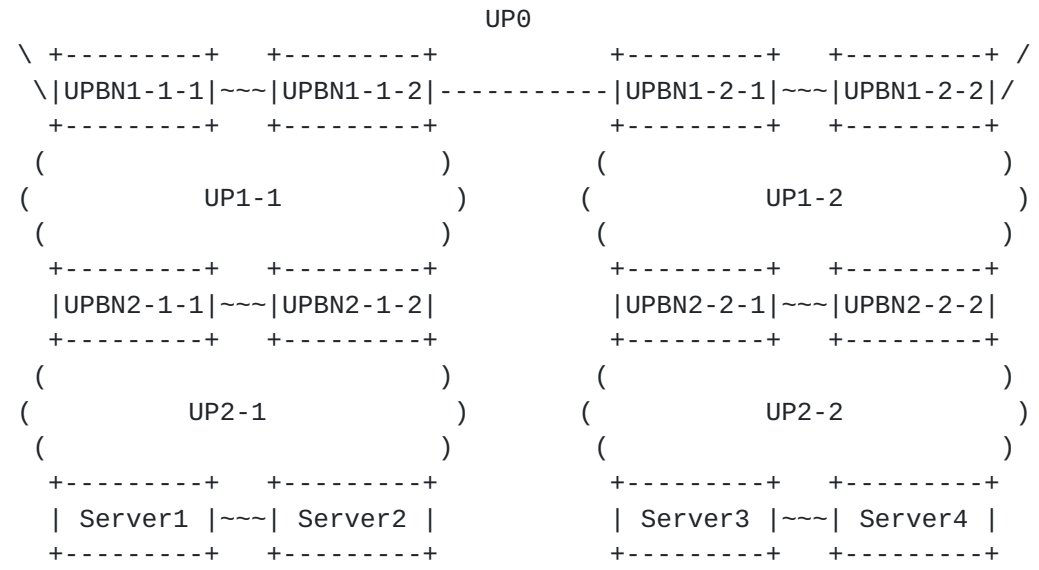


Figure 1 - Example topology with 3 levels of partitioning

In the example of Figure 1, there are 3 levels in the hierarchical partitioning. The UPs are connected by a number of Underlay Partition Border Nodes (UPBNs), grouped in Underlay Partition Border Groups (UPBGs). The UPBGs are the destinations for ECMP-forwarded traffic in each partition.

Packets from Server3 to Server1 use a label stack consisting of 3 Path Labels (PLs) for forwarding.

- Top label (PL0) forwards the packet to one of the UPBN1-1 nodes, which are grouped as UPBG1-1, connecting to UP1-1, which contains Server1 (note that, by definition of HSDN forwarding, PL0 points to UPBG1-1, i.e., the destination in UP0, rather than UPBG2-1).

- Next label (PL1) forwards the packet to one of the UPBN2-1 nodes, which are grouped as UPBG2-1, connecting to UP2-1, which contains Server1 (UPNBG2-1 is a destination in UP1-1).

- Next label (PL2) forwards the packet to Server1 (which is a destination in UP2-1)

This document proposes modified BGP-LU based procedures for:

- How each UPBN learns the destinations in its UP and the operator

assigned partition unique labels that should be installed in its LFIB to forward traffic to these destinations;

- How UPBN learns the context labels used by other UPBN destinations in the partition if the DC operator implements a policy of using separate LFIBs for installing partition unique labels on UPBNs

We also introduce an associated new extended community [RFC4360] that serves the following purposes:

1. Enables a UPBN to trigger the modified BGP-LU behavior to allow distribution of partition-unique labels to UPBNs from Label Mapping Server (LMS), and

2. Identifies which LFIB partition unique labels should be installed into (if there is ambiguity due to overlapping label name spaces), and

Such extended community allows to advertise persistent labels, which can survive across BGP session restarts.

Strictly speaking, the labels advertised with the new mechanisms described in this document are not typical downstream-advertised labels, but they are more similar to upstream-advertised labels installed in context LFIBs corresponding to upstream.

It should be noted that the BGP-LU procedures specified in this document may be implemented through operator configured policy using any existing BGP community types if some conditions are met. The minor changes to the procedures and the conditions under which policy based application of an existing BGP community can be used are described in Section 3.5.

The procedures specified in the document are applicable to ECMP traffic in mpls-based HSDN DC cloud architectures.

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 RFC 2119 [RFC2119].

This document inherits the terminology defined in [I-D.fang-mpls-hsdn-for-hsdc] and additionally introduces the following terms that apply when BGP-LU based control plane is used to realize HSDN architecture.

o Border Node (BN): A border node is a node that is present in a UP.

In HSDN architecture, UPBNi is a special BN that connects UPi with UPi-1.

- Partition Label Space: Label space that is shared by all border nodes of a UP to reach a destination in the UP. For a border node, UP destinations comprise other border nodes and end devices that are present in the UP.

- Partition Labels: Operator assigned labels that belong to partition label space corresponding to a UP. The labels need not be allocated from the platform label space on the BNs but may be directly installed in the context table corresponding the UP.

- Label Mapping Server (LMS): A BGP speaker present in each UP that allocates labels for destinations in the partition and distributes the labels to border nodes through BGP-LU.

- BGP Peer Group: Collection of BGP peers for which a set of policies are applied on a BGP speaker.

- Partition-Unique Label Info Community: A new type of BGP extended community that contains the operator assigned partition unique label for the BGP destination, origin partition and border group identifier.

- Border-group Community: This community identifies a group of border nodes that interconnect two partitions and is configured as policy on the border nodes as well as the LMS. It acts as the UPBG identifier.

- Route Resolver: A single or a collection of entities that provides the MPLS label stack to reach a destination underlay end device.

Term	Definition
BGP	Border Gateway Protocol
BGP-LU	Border Gateway Protocol Labeled Unicast
BN	Border Node
DC	Data Center
DCI	Data Center Interconnect
ECMP	Equal Cost MultiPathing
FIB	Forwarding Information Base
HSDN	Hierarchical SDN
LFIB	Label Forwarding Information Base
LMS	Label Mapping Server
MPLS	Multi-Protocol Label Switching
SDN	Software Defined Network
UP	Underlay Partition

UPBG	Underlay Partition Border Group
UPBN	Underlay Partition Border Node
TE	Traffic Engineering

3. Description of BGP-LU Procedures

This section provides an overview of how operator assigned partition label space is used to achieve end-to-end forwarding of label stacked packets. Consider the DC network that is present in the right hand side DC in Figure 1. The diagram in Figure 2 is a part of the DCI network in Figure 1 (the partitions are arranged horizontally rather than vertically as in Figure 1). UP1 in Figure 2 denotes a level 1 UP and UP2 denotes a level 2 UP. BN1 and BN2 are UPBNs of UP1, BN3 and BN4 are UPBNs of UP2. The nodes BN5 and BN6 may be some ToR switches or Servers. The nodes BN3, BN4, BN2, and BN1 are internal to the DC/DCI network (leafs and spines).

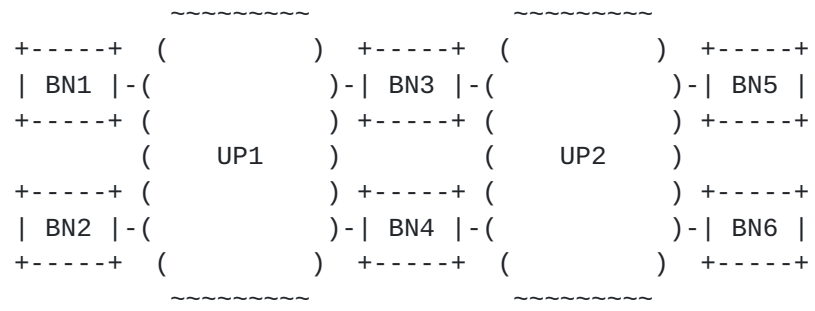


Figure 2 - Example to illustrate partition labels

If the DC network in Figure 2 ran conventional flat distributed BGP-LU control plane using router-allocated labels, when BN5 advertises itself as destination to BN3, BN3 allocates a new label (say L35) from its platform label space. If BN3 finds BN5 reachable (through say LSP35), it advertises L35 (for destination BN5) to BN1. Similarly, BN1 finds BN3 reachable (through say LSP13) and pushes two labels - bottom label is L35 and top label is the LSP13 label. In this model, BN3 stitches L35 to LSP35 that takes the packet to BN5. The same procedure runs on BN4, which allocates a label (say L45, in general different from L35) from its own platform label space for BN5 and advertises the label to BN1. This model is not suitable when end-to-end traffic from a Server behind BN1 or BN2 (not shown in the figure) to a Server behind BN5 or BN6 (not shown in the figure) needs to be forwarded using a label stack imposed by the SDN Controller with the condition that the label stack does not depend on the BN traversed to reach UP2 from UP1.

This document specifies a mechanism to implement the forwarding model

using label stacks imposed by SDN Controller but not have the limitation described in previous paragraph. The new procedures introduced in this document are explained using the above example.

1. BN5 and BN6 advertise their own loopback addresses in UP2. Assuming BN5 and BN6 do not belong to any border group, the BGP-LU advertisements from BN5 and BN6 contain NULL label. The routes will be:
   {Nlri: BN5, Label: NULL, Nh: BN5}
   {Nlri: BN6, Label: NULL, Nh: BN6}

2. BN3 and BN4 do not allocate labels for BN5 and BN6 from their own platform label space when they receive the BGP-LU advertisements. This is because BN3 and BN4 are configured to be part of a border group for UP2 destinations. Both BN3 and BN4 are configured with border group community "Border-group-2".

3. BN3 and BN4 re-advertise BN5 and BN6 as IP NLRI destinations (with BGP next-hop self) to the LMS assigned for UP2 and appends "Partition-Unique Label Info" extended community . The Partition-Unique Label Info extended community and the procedures relating to it are newly introduced in this document. Refer to Section 3.1 for the extended community format and Section 3.2 for LMS procedures. The R-bit in the extended community is set to indicate that the originator requests the receiver to assign and reflect the partition label info community with the label assigned by LMS. The routes for BN5 destination will be:
   {Nlri: BN5, Nh: BN3, Com: Border-group-2, Label-Ext-Comm: R}
   {Nlri: BN5, Nh: BN4, Com: Border-group-2, Label-Ext-Comm: R}

   If the operator has set aside a BGP community value that unambiguously indicates that the next-hop (BN3 or BN4) in the BGP route requests a label to be allocated for the destination (BN5) in UP2 partition, then the newly specified Partition label info extended community may not be added to the route. Refer to Section 3.5 for details.

4. UP2 LMS processes the IP routes for BN5 and BN6, assigns labels for them (or simply reads the labels from label mapping database configured by operator) and originates a BGP-LU route containing the label assigned for the UP2 destinations. LMS may set the P-bit to indicate that the label can be persistent and can be retained for a specified time period. For the two IP routes for BN5 originated by BN3 and BN4, the BGP-LU routes originated by LMS will be:
   {Nlri: BN5, Label: L5, Nh: BN3, Com: Border-group-2, Label-Ext-Comm: P:UP2-context}
   {Nlri: BN5, Label: L5, Nh: BN4, Com: Border-group-2, Label-Ext-

Comm: P:UP2-context}

The procedures if newly specified partition label info extended community is not used are described in Section 3.5.

5. Only when BN3 and BN4 learn the BGP-LU route for BN5 advertised by LMS of UP2, they install the label route in context table corresponding to UP2-context. Note that the operator may configure BN3 and BN4 to install the operator assigned label for BN5 in main LFIB itself (instead of UP2-context). The operator may choose this option if non-overlapping labels are assigned for different UPs.

6. BN3 and BN4 do not advertise BN5 and BN6 in UP1 but only advertise their own loopback addresses. As BN3 and BN4 are configured to be part of a border group, the border group identifier advertised as community is the same in BGP-LU advertisements from BN3 and BN4. If the partitions may have overlapping label spaces, then BN3 and BN4 advertise non-NULL labels in their BGP-LU advertisements. BN3 and BN4 install the label (that gets advertised) in default LFIB and point the label entry to the context table for UP2. In such a case, the routes from BN3 and BN4 will be:
{Nlri: BN3, Label: CL3, Nh: BN3, Com: Border-group-1}
{Nlri: BN4, Label: CL4, Nh: BN4, Com: Border-group-1}

7. BN1 and BN2 do not allocate labels for BN3 and BN4 from their platform label space when they receive BGP-LU advertisement. BN1 and BN2 only use the BGP-LU advertisement from BN3 and BN4 for determining the labels to be pushed during forwarding. Note that if there are intermediate routers between BN1/BN2 and BN3/BN4, then the labels CL3 and CL4 advertised by BN3 and BN4 will be used by those intermediate routers for determining the labels to be pushed.

8. BN1 and BN2 re-advertise BN3 and BN4 as IP destinations (with BGP next-hop self) to the LMS assigned for UP2 and appends "Partition-Unique Label Info" extended community. The R-bit is set to indicate that the originator requests the receiver to assign and reflect the partition label info community with the label assigned by LMS. The routes for BN3 destination will be:
{Nlri: BN3, Nh: BN1, Com: Border-group-1, Label-Ext-Comm: R}
{Nlri: BN3, Nh: BN2, Com: Border-group-1, Label-Ext-Comm: R}

   The procedures if newly specified partition label info extended community is not used are described in Section 3.5.

9. UP1 LMS processes the IP routes for BN3 and BN4, assigns labels for them (or simply reads the labels from label mapping database configured by operator) and originates a BGP-LU route containing

the label assigned for the UP1 destinations. As the group label advertisements will differ only in BGP next-hop, BGP add-path should be enabled on the peer group between LMS and BNs. LMS may set P-bit to indicate that the advertised label can be persistent and can be retained for specified time. For the two IP routes for BN3 originated by BN1 and BN2, the BGP-LU routes originated by LMS will be:
{Nlri: BN3, Label: L3, Nh: BN1, Com: Border-group-1, Label-Ext-Comm: P:UP1-context}
{Nlri: BN3, Label: LG2, Nh: BN1, Com: Border-group-1, Label-Ext-Comm: PG:UP1-context}
{Nlri: BN3, Label: L3, Nh: BN2, Com: Border-group-1, Label-Ext-Comm: P:UP1-context}
{Nlri: BN3, Label: LG2, Nh: BN2, Com: Border-group-1, Label-Ext-Comm: PG:UP1-context}

Note that there are two BGP-LU routes with same NLRI for advertising group label and so BGP add-path [I-D.ietf-idr-add-paths] should be enabled between LMS and BNs.

10. Only when BN1 and BN2 learn the BGP-LU route for BN3 advertised by LMS of UP1, they install the label route in context table that has been configured on BN1 and BN2 to contain all UP1 destinations. Note that the operator may configure BN1 and BN2 to install the operator assigned label for BN3 in main LFIB itself (instead of UP1-context). The operator may choose this option if non-overlapping labels are assigned for different UPs.

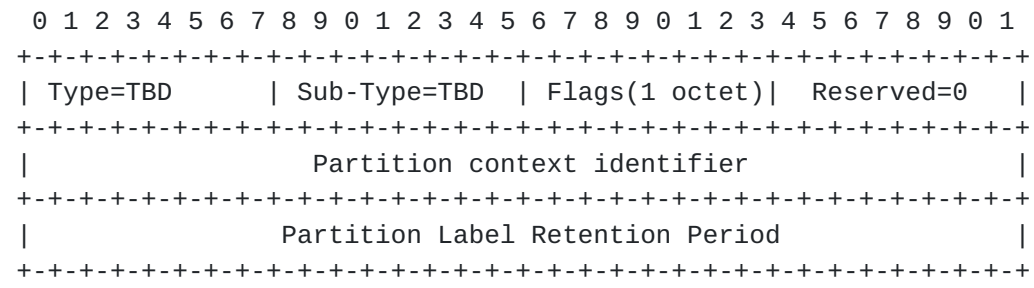
Apart from advertising partition labels to BNs, the LMSs also advertise the routes (IP routes received from BNs as well as the BGP-LU routes originated back to BNs) to Route Resolver. Resolver is logically centralized component that constructs label stacks for end-to-end traffic and it uses the routes advertised from LMSs as inputs for constructing label stacks.

The description of the procedures using the example DC network in Figure 2 provides an overview of how the LFIB states are set up for traffic entering BN1 or BN2 is forwarded to BN5 or BN6 ("downward traffic"). It should be be noted that this overview has not explained how packets from a source in a remote DC can reach BN5 or BN6. In other words, the overview has not yet explained how packets are exchanged between servers in one DC to the other DC in Figure 1. The description of how the LFIB states are setup for "upward traffic" is presented in Section 3.4.

3.1. Partitioned-Unique Label Info Extended Community

This document introduces a new extended community that enables the

originator of a BGP-LU route to convey the information specified below.



Flags

R-bit: Set to 1 if the originator requests label

G-bit: Set to 1 if the label is a group label

P-bit: Set to 1 if the receiver can retain label for specified time even if BGP peering between LMS and BN is lost

Partition context identifier: Context table identifier to which label will be installed

Partition label retention period: Timer period in seconds that the label can be retained after the BGP peering between LMS and BN is lost. This value must be zero if P-bit is not set.

3.2 Partition-Unique Label Info Extended Community Procedures

LMS is a BGP speaker that implements the following new procedures when it receives an IP route BGP advertisement containing "Partition-Unique Label Info" extended community.

- If IGP is the routing protocol with in a UP, then LMS may be implemented as a modified Route Reflector (RR) [RFC4456] assigned for the UP.

- If eBGP runs with in a UP, then the BGP peering between LMS and each border node should be configured by operator and on the BNs the eBGP peering with LMS should be configured in a peer group separate from eBGP peering with other routers in the partition. Note that even if eBGP is in use, the LMS procedures may be considered to act as a "modified reflector" because the primary goal of LMS is to return back the partition label to BN.

- LMS is configured with all the border groups that are connected to the UP where each border group is identified by a unique value of

Border-group community.

When LMS receives an IP route advertisement whose NLRI and BGP next-hop are the same, then it executes the following procedure.

1. If the operator has already assigned a label (DstLabel) for the UP destination in the NLRI, then no action is performed.

2. If the operator has not assigned a label for the UP destination, then LMS allocates a label (DstLabel) and stores the mapping between the UP destination and the label.

3. If the IP route advertisement also contains a known Border-group community and if the operator has not assigned a label for the border group, then LMS allocates a label and stores the mapping between the Border-group and the allocated label. Let the label assigned or allocated be BGLabel. LMS also stores the NLRI to the list of nodes belonging to the Border-group community contained in the route.

4. After executing the following procedures, LMS advertises the IP route to the Route Resolver.

When LMS receives an IP route advertisement whose NLRI and BGP next-hop are different, then it executes the following procedures.

1. If the IP route advertisement does not contain "Partition-Unique Label Info" extended community, then no further action is taken. Alternatively, if the LMS is configured with a policy to interpret a BGP community configured on it as equivalent to "partition label info" extended community, then the subsequent steps may be executed (refer to Section 3.5 for details).

2. If the IP route advertisement contains "Partition-Unique Label Info" extended community but the BGP next-hop does not belong to any known Border-group community configured on the LMS, then no further action is taken.

3. If none of the above conditions is true, then the LMS executes the following procedures.

   a. LMS retrieves the DstLabel label already assigned for the UP destination. LMS originates BGP-LU route with DstLabel set in the NLRI and clears the G-bit in "Partition-Unique Label Info" extended community. If the partition labels are operator assigned and is read from label mapping database, then LMS sets P-bit in the extended community flags and sets the "partition label retention period" to the value configured on LMS (default

value is 7200 seconds).

b. If the NLRI of the IP route is equal to a known Border-group community configured on the LMS, then the LMS also retrieves the BGLabel assigned for the Border-group. LMS also originates BGP-LU route with BGLabel set in the NLRI and sets the G-bit in "Partition-Unique Label Info" extended community. If the partition labels operator assigned and is read from label mapping database, then LMS sets P-bit in the extended community flags and sets the "partition label retention period" to the value configured on LMS (default value is 7200 seconds).

When the BN that originated the IP route receives the BGP-LU route "reflected" back by the LMS, it executes the following procedures.

1. BN first checks whether R-bit is cleared in "Partition-Unique Label Info" extended community. If R-bit has been reset, the label in the NLRI is installed in the context table corresponding to the "partition context identifier" present in the extended community. If "partition context identifier" is zero, then BN installs the label entry in default LFIB.

2. If P-bit is set, then BN should retain the label entry in the designated LFIB (context or default) for the time period specified in "partition label retention period" should the BGP peering with LMS is lost. After BGP peering with LMS is lost, the BN should start "label retention timer" for the labels learnt from the LMS. When the BGP peering is restored, BN should reset the "label retention timer" and re-advertise IP routes corresponding to all UP destinations it had originated before. This procedure ensures that both LMS and BNs exchange all requisite routes before reaching steady state again.

3. If P-bit is not set, then BN should delete the label entry immediately when BGP peering with LMS is lost.

4. BN should delete the label entry from the LFIB when LMS withdraws the BGP-LU route containing the "Partition-Unique Label Info" extended community.

3.3 BGP Policies on UPBNs and LMS

The BGP-LU based control plane mechanism specified in this document assumes the following set of policies be applied on various network nodes in HSDN architecture. The policy configurations required are listed below.

- Each UPBN that connects two UPs are configured with a unique Border-group to advertise membership to "border group" or UPBG. For example, in figure 1 UPBN1-1-1 and UPBN1-1-2 are configured with same Border-group community that uniquely represents the connectivity of the two BNs to UP1-1.

- Depending the routing protocol used with in a UP, each UPBN should either have iBGP or eBGP peering sessions such that all lower level UPBNs or end-devices that are connected to the UP learn each other. For example, the BNs present in UP1-1 in Figure 1 are UPBN1-1-1, UPBN1-1-2, UPBN2-1-1 and UPBN2-1-2 and each of them should learn the loopback address of the other BNs.

- Each UP should have a Label Mapping Server (LMS) that advertises to all the UPBNs the operator assigned partition labels corresponding to each UP destination. Destinations of UPi consists of all individual UPBNi+1 connected to UPi and lower level UPBGs connected to UPi. For example, destinations of UP1-1 (Figure 1) are UPBN2-1-1, UPBN2-1-2 and UPBG2-1, and LMS-1-1 will assign and advertise three labels for UP1-1.

- Each BN in a UP should also have iBGP or eBGP peering session with LMS of the UP. For example, all BNs in UP1-1 should have eBGP peering session with LMS-1-1 if UP1-1 runs eBGP routing protocol.

- UPBNj has a policy to automatically export destinations learnt from UPBNi peer group to UPj peer group (where i=j-1). But UPBNj does not export destinations learnt from UPj peer group to UPBNi peer group. This export policy on UPBNj limits the number of BGP advertisements that any network node in UPi has to process apart from limiting the number of LFIB entries in network nodes.

3.4 BGP-LU Procedures for UP0 Destinations

It should be noted that in the example topology in Figure 2, the BNs attached to UP1 and UP2 have been specified as UP destinations for illustration purposes only. Even a remote destination can be considered as a UP destination as long as the route is leaked into the UP. In HSDN architecture, even though the BNs connected to UP0 are remote for the UPBNs from level 2 down to the leaf level, as long as the normal BGP-LU route leaking policy (specified in Section 3.3) is followed, the LMS of the level 2 (or lower level) UPs will have to assign label for BNs in UP0 (or UP0 destinations). For example, UPBN2-1-1 and UPBN2-1-2 (figure 1) will learn UPBN1-1-1, UPBN1-1-2, UPBN1-2-1 and UPBN1-2-2 because UPBN1-2-1 and UPBN1-2-2 are leaked into UP1-1.

In the DC cloud network specified in Figure 1, the following

procedures are executed to enable packets with top label PL0 reach one of UPBNs connecting to UP0. To obtain end-to-end forwarding using a three label stack in a HSDN network with two levels (i.e. Servers located in UP2-x), the LMS of all UP2-x and UP1-x are set up such that they reflect the same label (i.e. PL0 label) for every UP0 destination (BNs as well as border groups).

1. UPBN1-2-1 and UPBN1-2-2 advertise their own loopback addresses in UP0. As the UPBNs are configured to be part of a border group, the border group community is the same in BGP-LU advertisements. If the partitions may have overlapping label spaces, then UPBN1-2-1 and UPBN1-2-2 advertise non-NULL labels in their BGP-LU advertisements. BN3 and BN4 install the label (that gets advertised) in default LFIB and point the label entry to the context table for UP1-2. In such a case, the routes from BN3 and BN4 will be:
   {Nlri: UPBN1-2-1, Label: CL11, Nh: UPBN1-2-1, Com: Border-group-0}
   {Nlri: UPBN1-2-2, Label: CL12, Nh: UPBN1-2-2, Com: Border-group-0}

2. For UPBN1-1-1 and UPBN1-1-2, the routes to UPBN1-2-1 and UPBN1-2-2 are in same partition (i.e. UP0). The label assigned for UPBN1-2-1, UPBN1-2-2 and UPBG1-2 are the same on LMS-0, LMS-1-1 and LMS-2-1. So all BNs in the left hand side DC in Figure 1 install the same label for UPBN1-2-1, UPBN1-2-2 and UPBG1-2.

Note that as all BNs in the DC cloud install the same label for a UP0 destination, the label range on the implementations of all BNs should have common label space (among different platform label spaces on all BNs) that can be set aside for the UP0 destinations. If this is not possible, then all BNs should be configured with a separate context table for UP0 partition. The BGP-LU procedures involving the "Partition-unique label info" community supports both forms of forwarding.

3.5 Advertising labels without partition label extended community

The procedures specified in Section 3.2 may be executed on LMS and border nodes without using the newly partition label info extended community but using an existing BGP community if all the following conditions are true.

- Each partition has a separate LMS such that border nodes connecting two partitions must have separate BGP peering with LMS of the two partitions.

- Both LMS and BNs are configured with a BGP community and both LMS and BNs interpret that community as an indication from the BGP peer that the procedures specified in Section 3 of this document should be

applied. If LMS receives IP route advertisement whose NLRI and next-hop attribute are different and contains the pre-configured BGP community, then LMS should interpret the update as label request from the BGP peer for the IP destination corresponding to the NLRI. Similarly, when BN receives BGP-LU advertisement for which the BN has originated an IP route and if the BGP-LU advertisement contains the pre-configured BGP community, then BN should interpret the update as partition label advertisement from LMS for the IP destination corresponding to the NLRI.

- BNs are configured with the LFIB to which the label advertised by the LMS should be installed. In this model, LMS cannot advertise the LFIB to which the label forwarding entry should be installed.

- Both LMS and BNs are configured with label retention policy in the event of BGP peering between LMS and BNs were to fail. For example, both LMS and BNs may be configured with label retention period of 7200 seconds so that BNs can retain the LFIB entry for 7200 seconds even if BGP peering with LMS fails.

4. Route Resolution in HSDN Architecture

As a consequence of the procedures described in Section 3, Route Resolver of the network will have the knowledge of the destinations in all UPs and the UPBNs that have advertised those UP destinations. Route Resolver uses this information to construct MPLS label stack to forward the packet to desired destination End-device.

Note that the procedure specified in this Section is only for illustration purpose and hence the implementation of Resolver is free to choose a more optimal mechanism to obtain the same result. The resolution for a given DstServer or End-device IP address works as follows.

1. Resolver should have received all BGP-LU routes of all End-devices from the LMSs of all "leaf" UPs with BGP next-hop specifying the UPBN that serves the UP. The Resolver looks up the given DstServer IP address in the resolution database. If the IP address is not present, then Resolver considers the resolution as having failed.

2. If the DstServer has been advertised by LMS of a UP, then the Resolver obtains the BGP next-hop from the BGP-LU route advertisement. The BGP next-hop is the UPBN of the leaf UP. Note that there may be multiple BGP-LU routes advertising the same DstServer. Assuming the policy is to use ECMP for the traffic, the Resolver picks the BGP-LU advertisement having G-bit set in "Partition-Unique Label Info" extended community and adds the BGLabel to the resulting stack. Assuming the DstServer is located

in second level UP and LG2 is the group label, the stack will be {LG2}.

3. Resolver then looks up the UPBN in the resolution database. If the UPBN IP address is not present, then Resolver considers the resolution as having failed. If there is one or more BGP-LU route with the UPBN as the destination, then the Resolver obtains the BGP next-hop(s). Assuming the policy is to use ECMP for the traffic, the Resolver picks the BGP-LU advertisement having G-bit set in "Partition-Unique Label Info" extended community and adds the BGLabel to the resulting stack. Assuming LG1 is the group label of level 1 UPBG, the stack will be {LG1, LG2}.

4. As the resolution has reached level 1 UPBN (that is a BN in UP0), the Resolver looks up the level 1 UPBN in resolution database. There should be multiple BGP-LU routes with level 1 UPBN as destination. Assuming the policy is to use ECMP for the traffic, the Resolver picks the BGP-LU advertisement having G-bit set in "Partition-Unique Label Info" extended community and adds the BGLabel to the resulting stack. Assuming LG0 is the group label of level 0 BG, the stack will be {LG0, LG1, LG2}. At this point the resolution is considered as successful (refer to Section 3.4) and the Resolver returns the resultant label stack to the querying system.

5. Security Considerations

The procedures defined in the document does not necessitate any security considerations.

6. IANA Considerations

This document defines a new extended community type (see Section 3.1).

7. Acknowledgments

We would like to thank Kaliraj Vairavakkalai and Balaji Rajagopalan for their valuable input and feedback.

8. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC3107] Rekhter, Y. and E. Rosen, "Carrying Label Information in BGP-4", RFC 3107, May 2001.

[RFC4456] Bates, T., Chen, E., and R. Chandra, "BGP Route Reflection: An Alternative to Full Mesh Internal BGP (IBGP)", RFC 4456, April 2006.

[RFC4360] Sangli, S., Tappan, D., and Y. Rekhter, "BGP Extended Communities Attribute", RFC 4360, February 2006.

9. Informative References

[I-D.fang-mpls-hsdn-for-hsdc] L. Fang, et. al., "MPLS-Based Hierarchical SDN for Hyper-Scale DC/Cloud", draft-fang-mpls-hsdn-for-hsdc-04 (work in progress), July 2015.

[I-D.ietf-idr-add-paths] D. Walton et al., "Advertisement of Multiple Paths in BGP", draft-ietf-idr-add-paths-10 (work in progress), Oct. 2014.

Authors' Addresses

Luyuan Fang
Microsoft
15590 NE 31st St.
Redmond, WA 98052
Email: lufang@microsoft.com

Deepak Bansal
Microsoft
15590 NE 31st St.
Redmond, WA 98052
Email: dbansal@microsoft.com

Chandra Ramachandran
Juniper Networks
Bangalore, India
Email: csekar@juniper.net

Fabio Chiussi
Seattle, Washington 98116
Email: fabiochiussi@gmail.com

Nabil Bitar
Verizon
40 Sylvan Road
Waltham, MA 02145
Email: nabil.bitar@verizon.com

Yakov Rekhter