BGP-LU for HSDN Label Distribution

draft-fang-idr-bgplu-for-hsdn-00

Luyuan Fang, lufang@microsoft.com

Chandra Ramachandran, csekar@juniper.net

Fabio Chiussi, fchiussi@cisco.com

Yakov Rekhter

IDR meeting, IETF 92

March 24, 2014, Dallas, TX

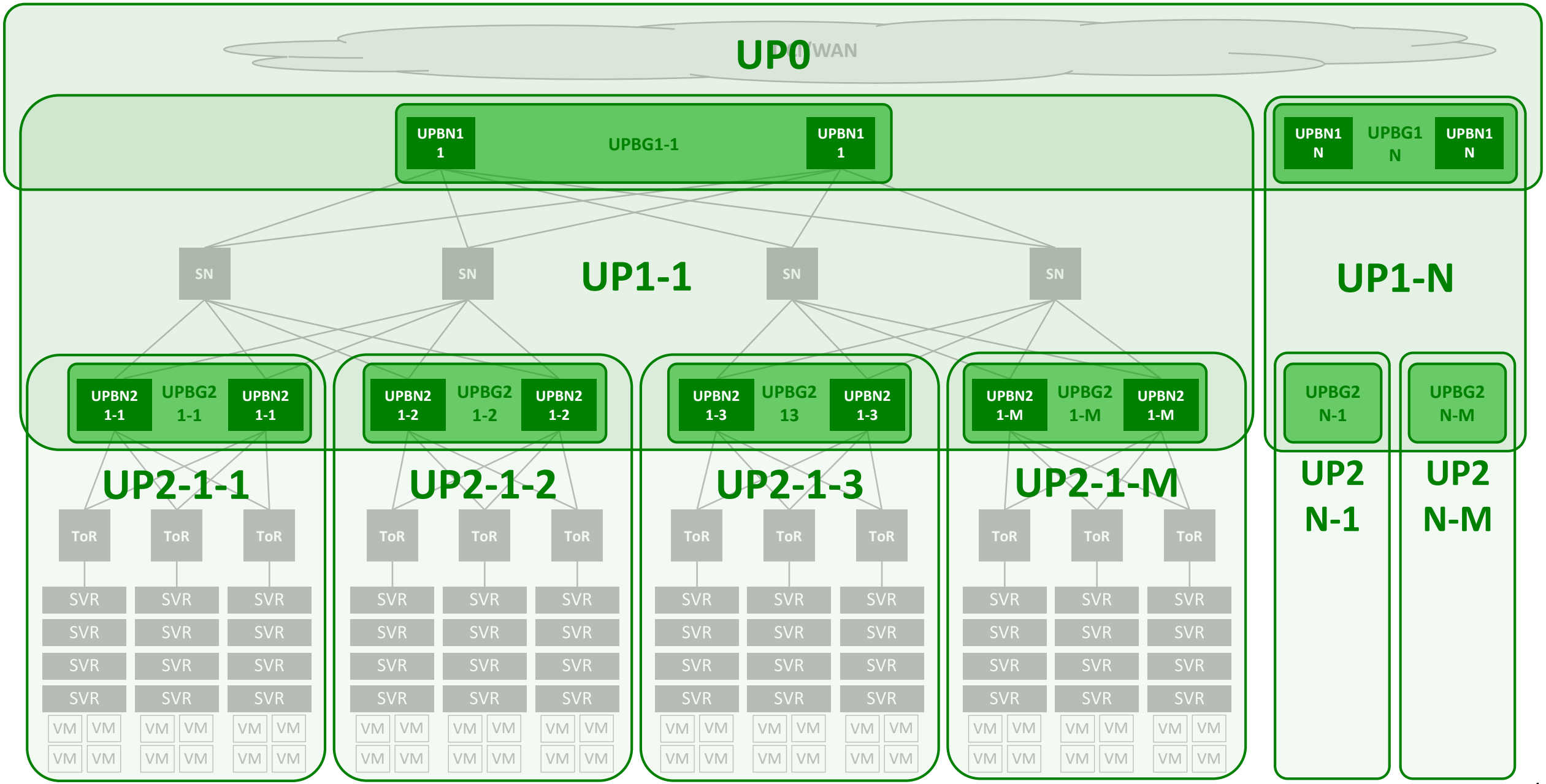
Purpose of the draft

Use BGP Labeled Unicast (BGP-LU), with modified BGP Route Reflector (RR) operation, as one of the options, for label distribution in the Hierarchical SDN (HSDN) (draft-fang-mpls-hsdn-for-hsdc-01) control plane (hybrid approach) for the hyper-scale Data Center (DC) and cloud networks.

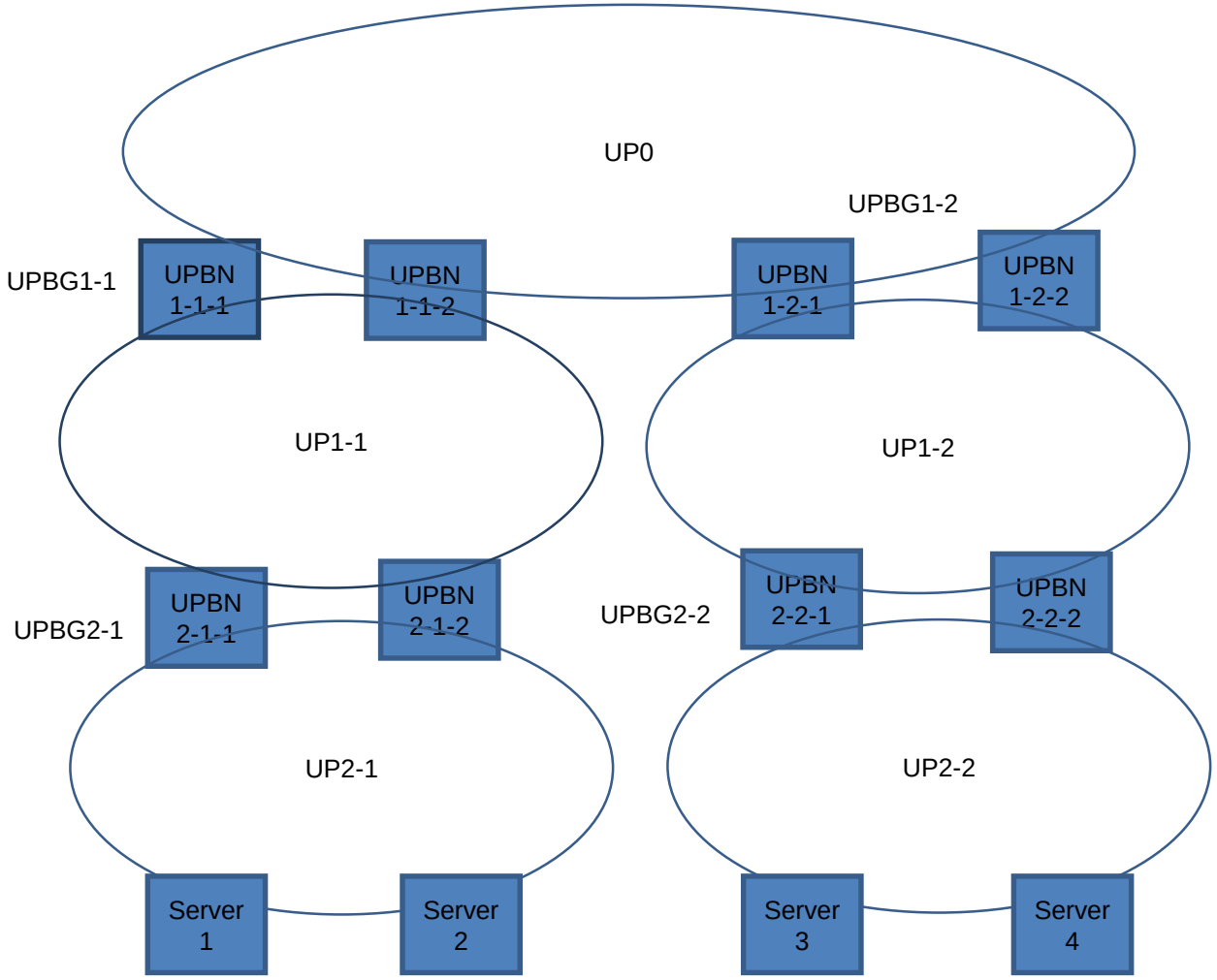
Terminoglogy

- UP: Underlay Partition.
- UPBN: Underlay Partition Border Node.
- UNBG: Underlay Partition Border Group.
- RR: BGP Route Reflector.
- BGP Peer Group: Collection of BGP peers for which a set of policies are applied on a BGP speaker.
- Label Mapping Server: A node present in each Underlay Partition that allocates labels for destinations in the partition.
- Label Mapping RR (LM-RR): A modified or customized BGP RR that uses BGP-LU to advertise label bindings for destinations in UP. LM-RR is an implementation of Label Mapping Server that uses BGP-LU to advertise the labels for partition destinations.
- Peer Community: An IP based extended community carried in BGP update that represents UPBG of a partition.
- Route Resolver: A single or a collection of entities that provides the MPLS label stack to reach a destination underlay end device.

Reference Model of HSDN: Hiarchical Partition with UPBNs and UPBGs

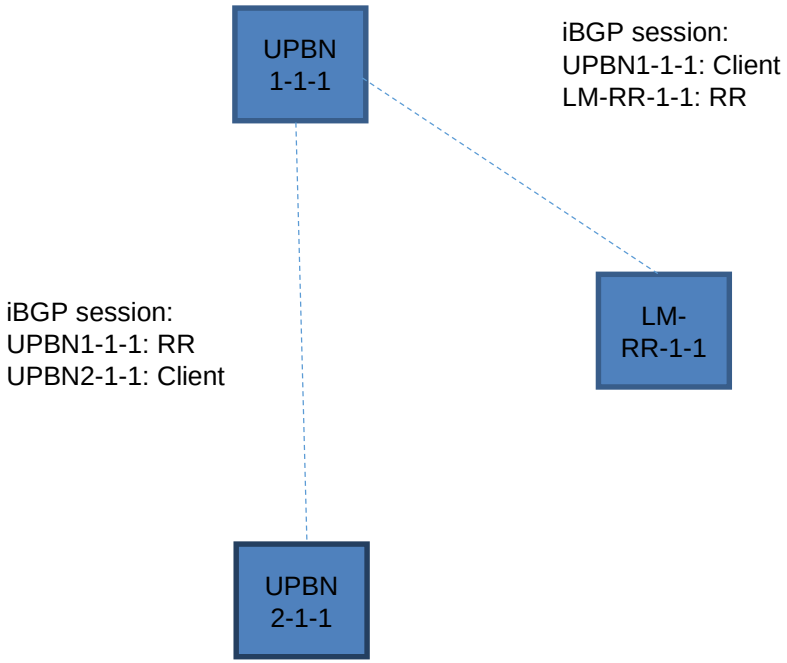


HSDN Example Topolgy



UPs running IGP

iBGP Sessions in UP1-1

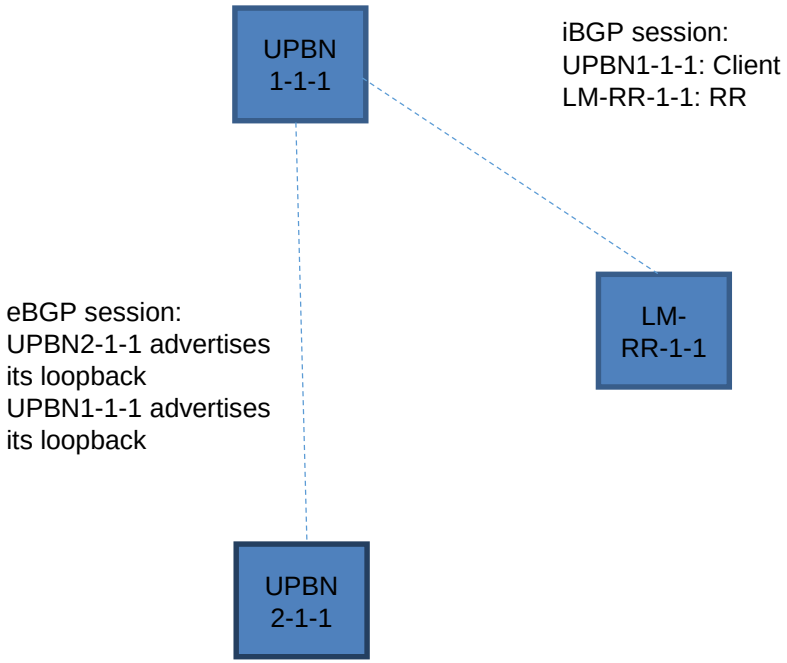


Note:

- On UPBN1-1-1, iBGP session with UPBN2-1-1 and iBGP session with LM-RR-1-1 belong to different peer-group
- On UPBN1-1-1, IGP for UP1-1 and UP0 are different instances and routes are not leaked between the instances

UPs running eBGP

eBGP Sessions in UP1-1



Note:

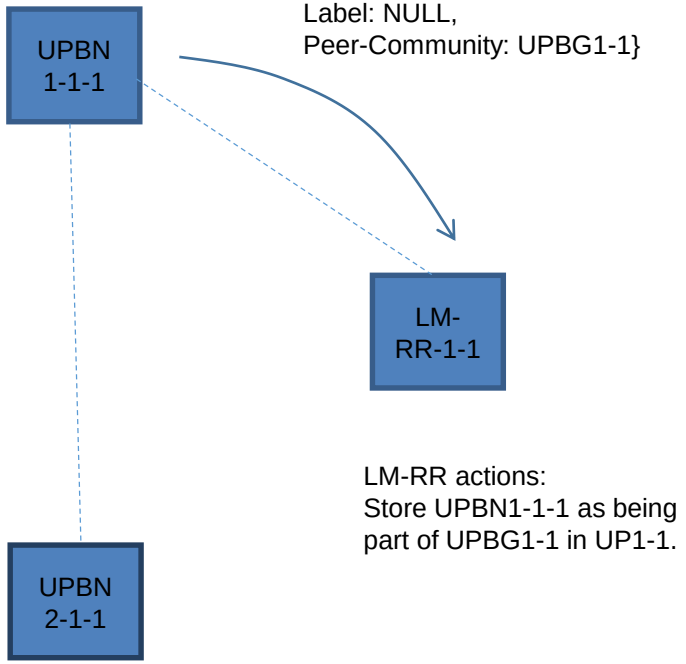
- UPBN1-1-1 and UPBN2-1-1 are in different AS
- If UPBN1-1-1 and UPBN2-1-1 are not directly connected, then there will be eBGP peerings such that each intermediate node in UP belongs to a different AS

BGP-LU procedures:

- The procedures described in the following slides are applicable for UPs running IGP or eBGP.

Step 1: UPBN1-1-1 originates self route

UPBN1-1-1 actions:
As I am UPBN of UP1-1, originate route to the peer-group with LM-RR



Route:
NLRI: UPBN1-1-1,
NH: UPBN1-1-1,
Label: NULL,
Peer-Community: UPBG1-1}

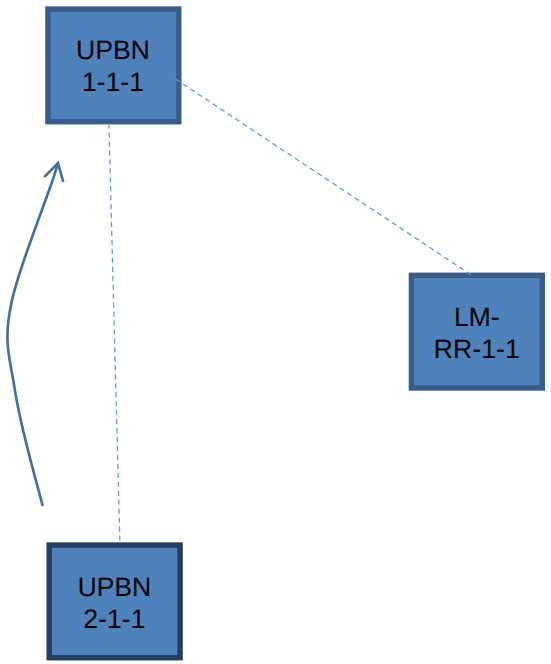
LM-RR actions:
Store UPBN1-1-1 as being part of UPBG1-1 in UP1-1.

Note:

- Peer-Community is a new extended community
- Each UPBG will have a unique Peer-Community value
- LM-RR may act as “vanilla” BGP-RR for labeled-unicast routes

Step 2: UPBN2-1-1 originates self route

UPBN1-1-1 actions:
As I am UPBN of UP1-1 and I have received route from peer-group of UP1-1, then do not perform normal BGP-LU actions

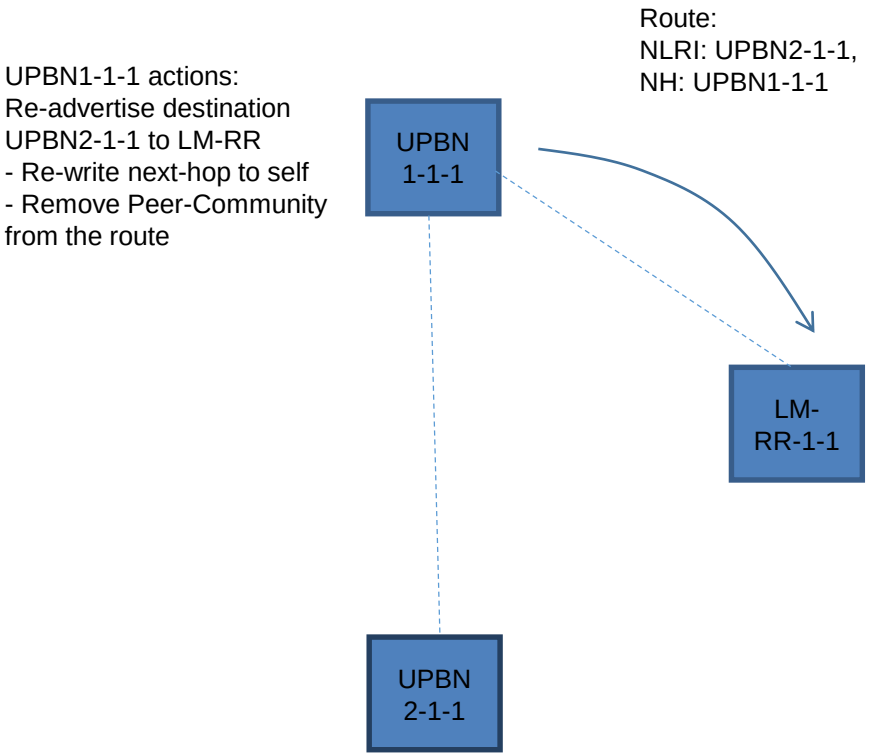


Route:
NLRI: UPBN2-1-1,
NH: UPBN2-1-1,
Label: NULL,
Peer-Community: UPBG2-1}

Note:

- UPBN2-1-2 will also advertise labeled-unicast route for itself to UPBG1-1 with Peer-Community UPBG2-1
- UPBN2-1-1 will also have iBGP session with UPBN1-1-2
- UPBNs of UP1-1 are RRs and destinations of UP1-1 are clients.

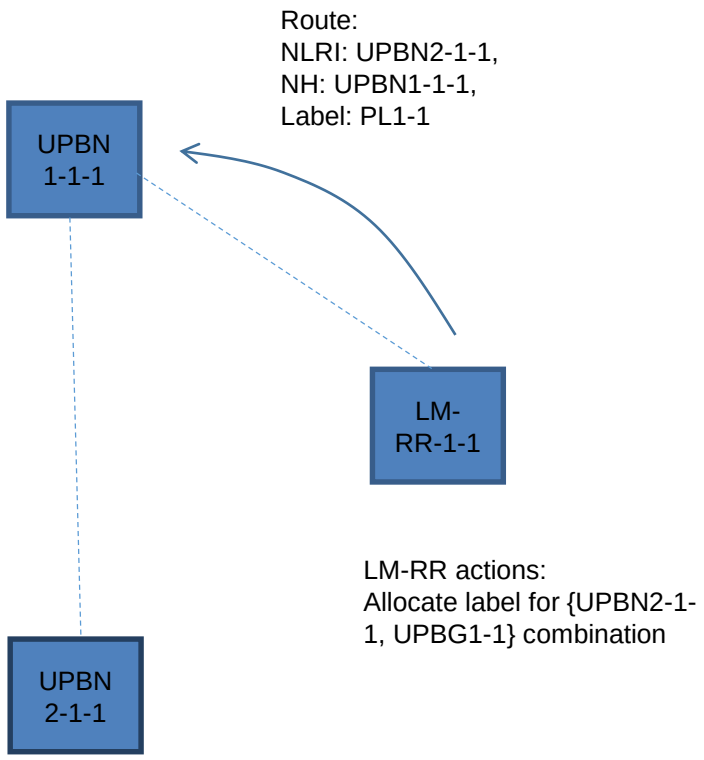
Step 3: UPBN1-1-1 re-advertises to LM-RR



Note:

- UPBN1-1-1 converts labeled-unicast route to inet family
- One can think of this as special UPBN action where unlike normal RFC3107 receiver, UPBN cannot allocate a label by itself and so it internally copies such “un-allocated” destinations to a special TIB called “LM-TIB”
- Any route in “LM-TIB” leads to route origination to iBGP peer-group with LM

Step 4: LM-RR allocates & advertises label

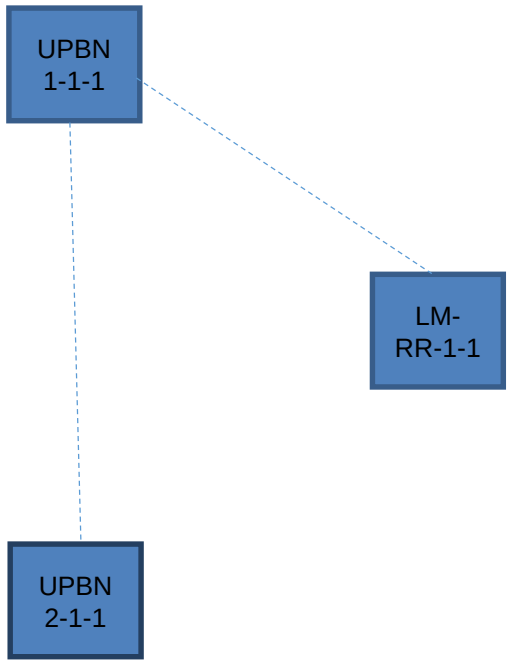


Note:

- LM-RR converts inet route to labeled-unicast family
- One can think of this as special LM-RR action where unlike normal RR, LM-RR cannot reflect “vanilla” IP routes
- LM-RR can be thought of as originating labeled-unicast route for each inet destination learnt where the label is allocated for each destination per UPBG (present in “LM-TIB”

Step 5: UPBN1-1-1 installs label in LFIB

UPBN1-1-1 actions:
- Install PL1-1 in LFIB
- Resolve UPBN2-1-1 using any LSP within UP1-1 to the destination



Note:
- One can think of this as special UPBN action where it internally copies the labeled-unicast route from LM-RR to “LM-TIB” making the destination UPBN2-1-1 as “allocated”
- Any “allocated” route in “LM-TIB” leads installation of label in LFIB

Summary so far…

- UPBN1-1-1 learns destination in its ‘own’ UP i.e. UPBN2-1-1
- UPBN1-1-1 places all these routes as “vanilla” IP destinations in its ‘LM-TIB’
  - This automatically triggers UPBN function resulting in advertisement to LM-RR peer group
- LM-RR learns and places these routes in its local ‘LM-TIB’
  - This automatically triggers LM function resulting in (a) labels allocated (or picked from static configuration) for “vanilla” IP destinations, and (b) origination of corresponding L-BGP route for the IP destinations
- UPBN1-1-1 learns L-BGP route from LM-RR and places these routes also in ‘LM-TIB’
  - Addition of L-BGP routes results in UPBN1-1-1 installing the labels in LFIB with forwarding action necessary to reach corresponding destination

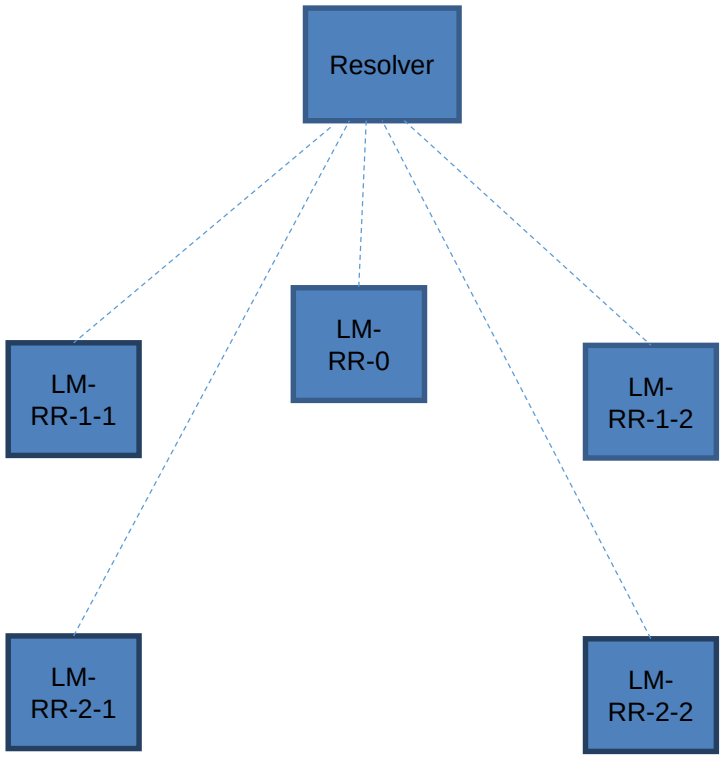
Forwarding packets ‘up’ the hierarchy (1)

- Slides so far only focused on UP1-1 destination i.e. how UPBN1-1-1 forwards packets from UP0 to UP2-1-1
- How does UP1-1 forward packets from UP1-1 to UP0 destination?
- Solution: Statically configure labels corresponding to UP0 destinations i.e. UPBNs of L1 partitions on all LM-RRs
- But, how is it different from static configuration of labels on all routers?
  - It does not require static configuration on all routers, but only on much fewer LM-RRs!

Forwarding packets ‘up’ the hierarchy (2)

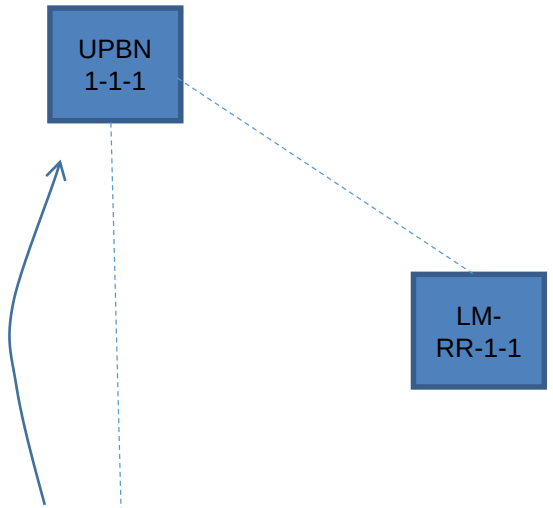
- Is it possible to completely avoid globally unique labels for higher level UPBNs too?
- A possible solution: In addition to providing label stack to (Source) Server, Resolver also provides “forwarding next-hop”.
  - Resolver provides one or more UPBN(s) connected to UP0 as forwarding next-hop(s)
  - (Source) Server or the ToR to which Server is connected to can learn these UPBNs connected to UP0 using “vanilla” BGP-LU!
  - Note that advertising routes “down” the hierarchy does not increase LFIB size and so acceptable (whereas advertising routes “up” the hierarchy is not).

Why Label Mapping RR? (1)



- LM-RRs apart from reflecting L-BGP routes in ‘LM-TIB’ to UPBN can also reflect them to Resolver (which is another BGP speaker)
  - Configure policy on Resolver so that routes from one LM-RR is not advertised to other LM-RRs
- What does this achieve?
  - Given an end-destination (or destination Server), Resolver can now use recursive route resolution using the L-BGP routes to determine the label stack to reach the end-destination
  - No other protocol is required
- What if number of BGP sessions on Resolver becomes a problem?
  - Let Resolver only speaks to LM-RR-0 and LM-RRs of level 1
  - LM-RRs at level 1 only speak to their corresponding child LM-RRs, and so on…

Why Label Mapping RR? (2)



Route:
NLRI: UPBN2-1-1,
NH: UPBN2-1-1,
Label: CL2,
Peer-Community: UPBG2-1}

UPBN2-1-1 actions:
As I am UPBN of UP2-1, I create a context LFIB for UP2-1 and point nexthop of CL2 to point to the context LFIB

- UPBNs by policy can place all routes learnt from LM-RR in a context LFIB and advertise a label that points to the context LFIB when they advertise themselves to parent UP
- For example, UPBN2-1-1 advertises CL2 to parent UPBN1-1-1 so that packets to UP2-1 destinations are looked up in separate context

Next Steps

- Initial draft, feedback is much appreciated
- Add procedure for label distribution for HSDN TE tunnels