BGP-LU for HSDN Label Distribution
draft-fang-idr-bgplu-for-hsdn-01

Luyuan Fang, Chandra Ramachandran, Fabio Chiussi, Yakov Rekhter

IDR meeting, IETF 93

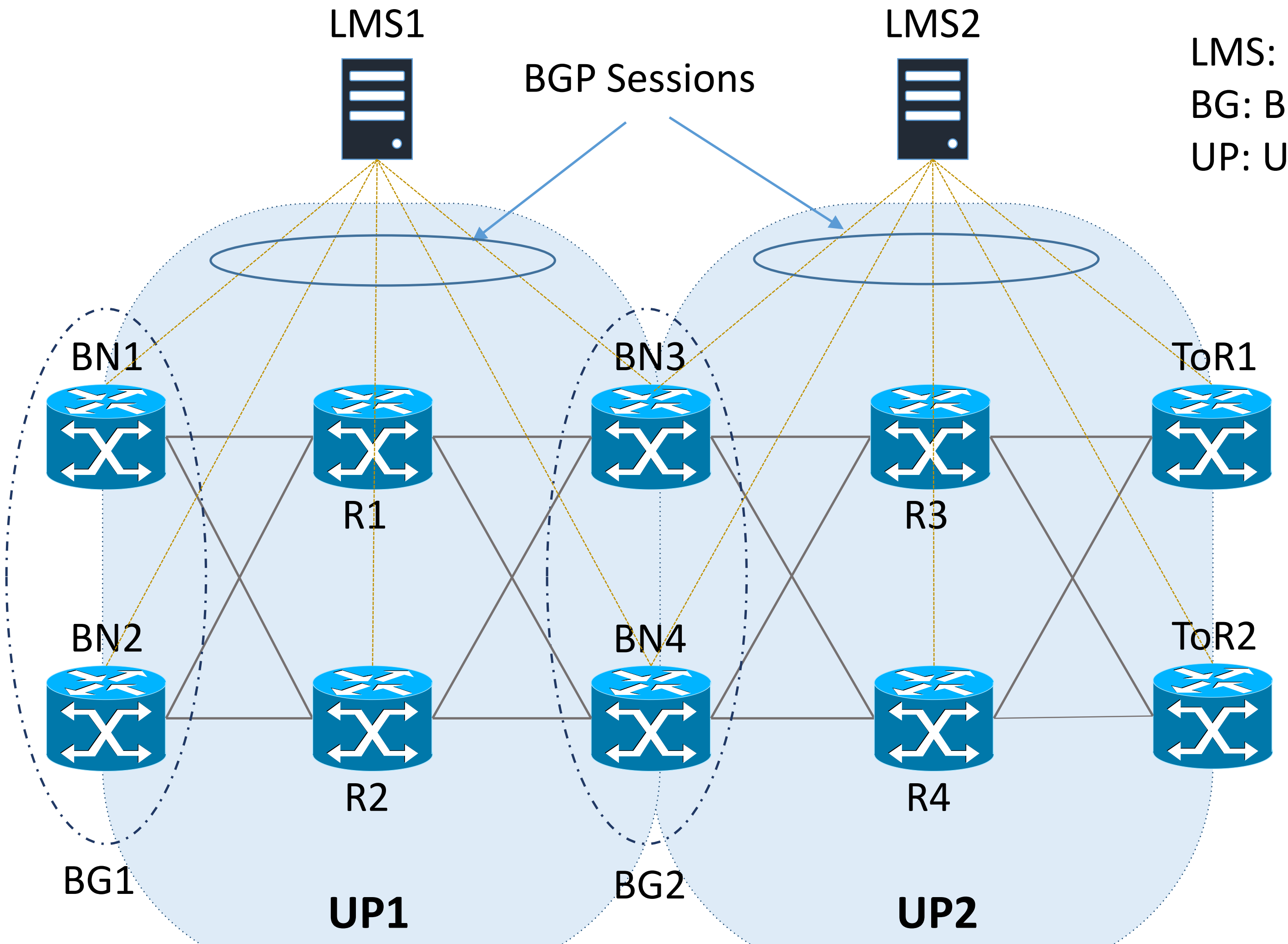
July 20, 2015, Prague

Summary of changes from 00 version

- Broaden applicability of the proposal
    - Any partitioned underlay network using BGP-LU label stacks
- Modified procedures so that they are applicable for partitions running either eBGP or IGP
    - Previous version only covered IGP
- Introduced a new extended community type so that the procedures can be used to:
    - Enable border nodes to unambiguously signal the remote BGP speaker(s) that new BGP-LU procedures requesting partition-unique label(s) should be executed
    - Enable Border Node and BGP speaking Label Mapping Server to scope the label request and the response to a unique partition

Example Topology

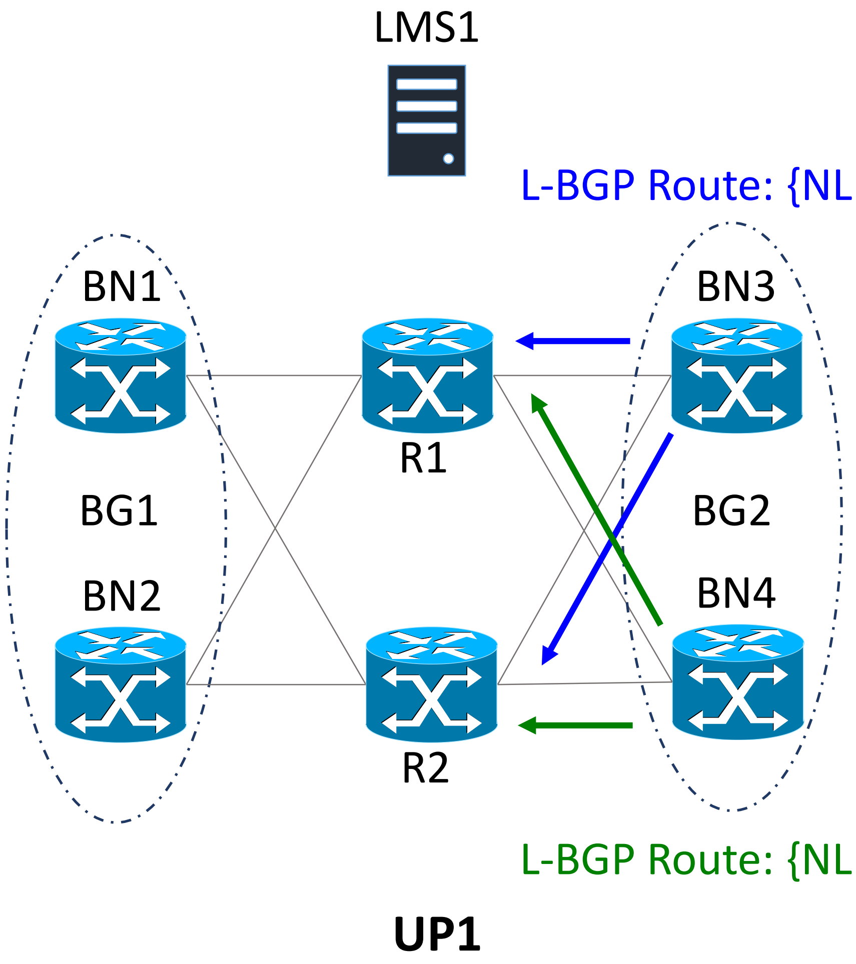
- An abstract model of DC in two level hierarchy



LMS: Label Mapping Server
BG: Border Group
UP: Underlay Partition

BN3, BN4, ToR1 and ToR2 have BGP peering with LMS2
BN1, BN2, BN3 and BN4 have BGP peering with LMS1

BNs of UP1 learn routes to BN3 and BN4



L-BGP Route: {NLRI = BN3-Loopback:CL3, NH = BN3, Com = BG2}

L-BGP Route: {NLRI = BN4-Loopback:CL4, NH = BN4, Com = BG2}

- CL3/CL4 is NULL label (PHP) if same LFIB is used for UP1 and UP2. Otherwise locally allocated Context Label is used
- BN3 and BN4 append a BGP community (value is BG2) when they advertise themselves in UP1

BNs of UP1 learn route to BN3 and BN4 (2)

L-BGP Routes (View of BN1):

{NLRI = BN3-Loopback:L113, NH = R1, Com = BG2}

{NLRI = BN4-Loopback:L114, NH = R1, Com = BG2}

{NLRI = BN3-Loopback:L123, NH = R2, Com = BG2}

{NLRI = BN4-Loopback:L124, NH = R2, Com = BG2}

- Interior routers in UP1 (R1 and R2) do not modify or remove border-group community in the L-BGP route
- When BN1 and BN2 receive the L-BGP routes for BN3 and BN4, they can conclude that BN3 and BN4 belong to BG2 group.

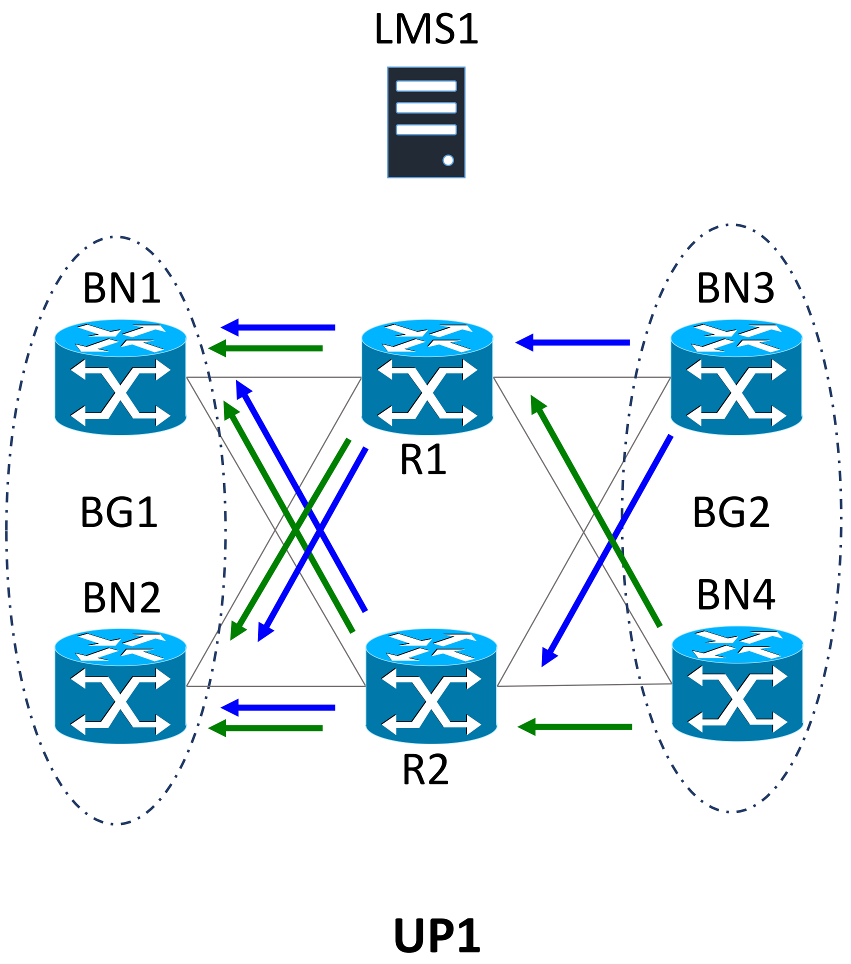
L-BGP Routes (View of BN2):

{NLRI = BN3-Loopback:L113, NH = R1, Com = BG2}

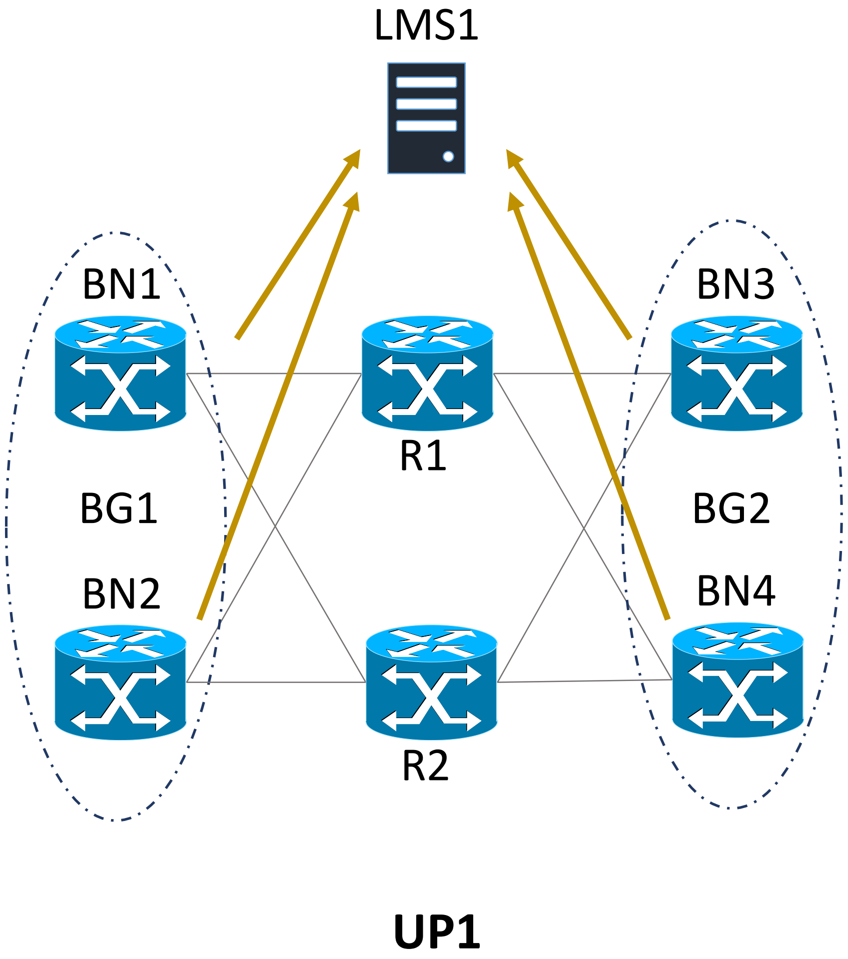
{NLRI = BN4-Loopback:L113, NH = R1, Com = BG2}

{NLRI = BN3-Loopback:L123, NH = R2, Com = BG2}

{NLRI = BN4-Loopback:L124, NH = R2, Com = BG2}



All BNs in UP1 Advertise Themselves to LMS1



L-BGP Routes (View of LMS1):

{NLRI = BN1-Loopback:CL1, NH = BN1, Com = BG1}
{NLRI = BN2-Loopback:CL2, NH = BN2, Com = BG1}
{NLRI = BN3-Loopback:CL3, NH = BN3, Com = BG2}
{NLRI = BN4-Loopback:CL4, NH = BN4, Com = BG2}

- LMS may not run regular BGP decision processes to compute routes
- LMS learns the group membership of BN3 and BN4 from the L-BGP advertisement

Partition labels – new procedures

- In the example so far, BN1 and BN2 have learnt BN3 and BN4 using normal BGP-LU procedures
- What is new?
  - BN1 and BN2 are configured to be partition border nodes for UP1 (the partition represented in the BGP extended community value)
  - When BN1 and BN2 learn a destination (BN3 or BN4) through L-BGP from BGP peers (R1 and R2) that belong to UP1 partition, then BN1 and BN2 do not allocate a label from platform label space and do not re-advertise
  - Instead, BN1 and BN2 “learn” the label for the destination (BN3 or BN4) in “partition label space” from the Label Mapping Server (LMS) through the new procedures specified in the draft

BN1 learns partition label for BN3 and BN4

IP Routes (from BN1 to LMS1):

{NLRI = BN3-Loopback, NH = BN1, Com = BG1, **Ext-com = R:UP1-context**}

{NLRI = BN4-Loopback, NH = BN1, Com = BG1, **Ext-com = R:UP1-context**}

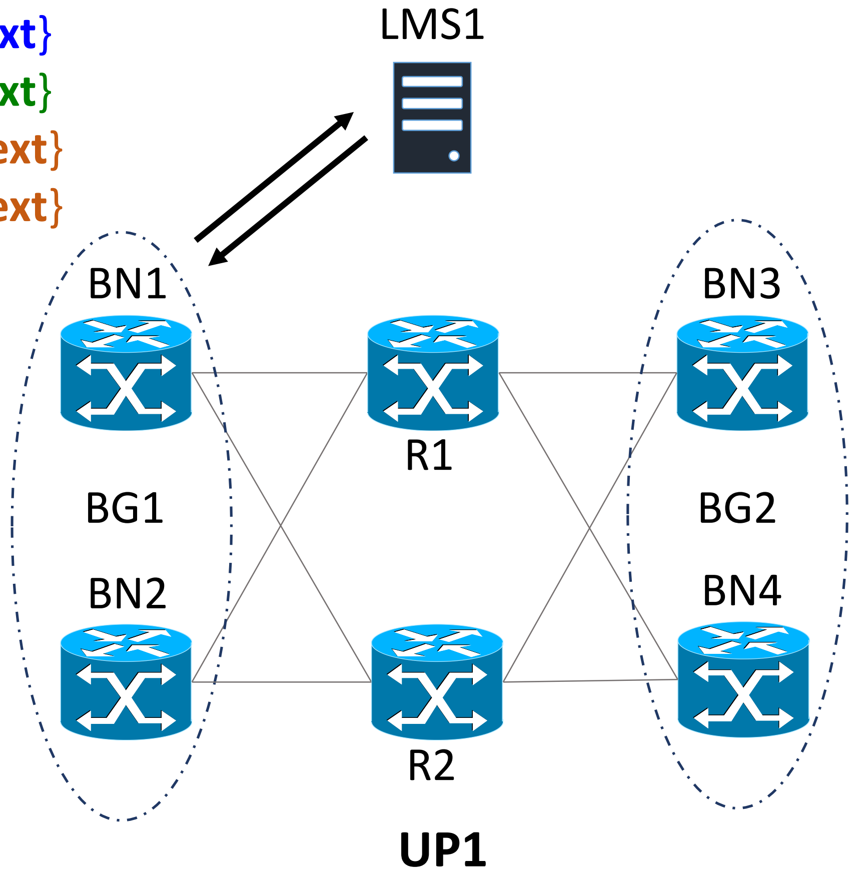
L-BGP Route (from LMS1 to BN1):

{NLRI = BN3-Loopback:PL13, NH = BN1, Com = BG1, **Ext-com = 0:UP1-context**}

{NLRI = BN4-Loopback:PL14, NH = BN1, Com = BG1, **Ext-com = 0:UP1-context**}

{NLRI = BN3-Loopback:PLG2, NH = BN1, Com = BG1, **Ext-com = G:UP1-context**}

{NLRI = BN4-Loopback:PLG2, NH = BN1, Com = BG1, **Ext-com = G:UP1-context**}



PLG2: Partition Label assigned for Border Node Group (BG2)

Ext-com:

- R: Request
- G: Group

BN2 learns partition label for BN3 and BN4

IP Routes:

{NLRI = BN3-Loopback, NH = BN2, Com = BG1, Ext-com = R:UP1-context}

{NLRI = BN4-Loopback, NH = BN2, Com = BG1, Ext-com = R:UP1-context}

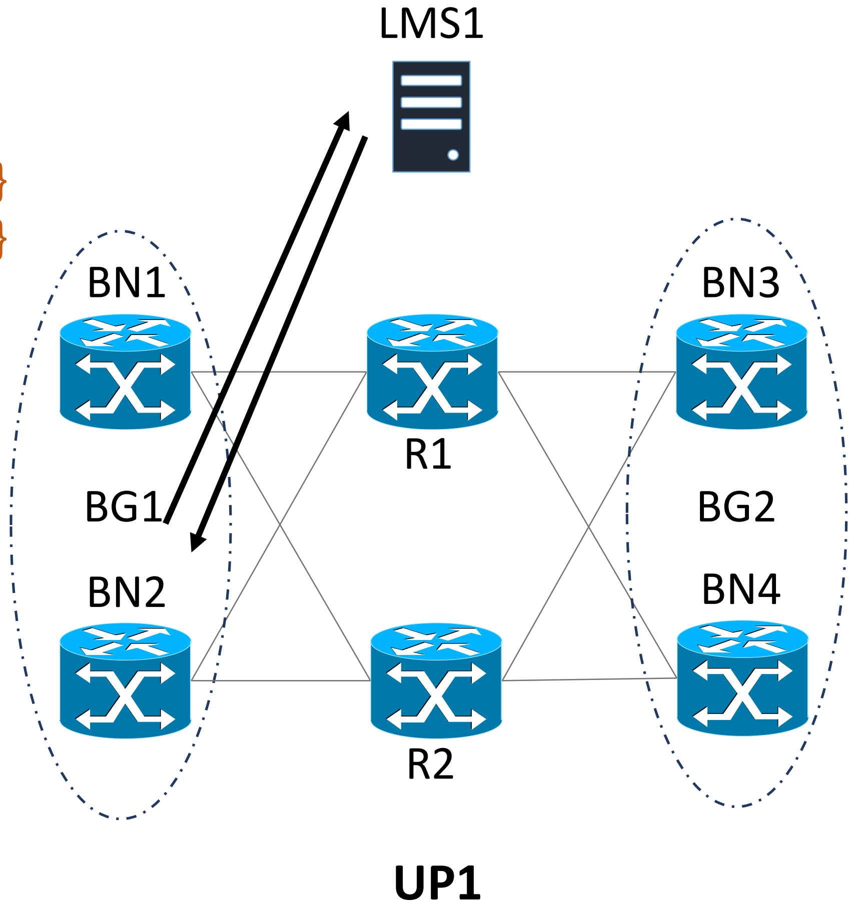
L-BGP Route:

{NLRI = BN3-Loopback:PL23, NH = BN2, Com = BG1, Ext-com = 0:UP1-context}

{NLRI = BN4-Loopback:PL24, NH = BN2, Com = BG1, Ext-com = 0:UP1-context}

{NLRI = BN3-Loopback:PLG2, NH = BN2, Com = BG1, Ext-com = G:UP1-context}

{NLRI = BN4-Loopback:PLG2, NH = BN2, Com = BG1, Ext-com = G:UP1-context}



All BNs in UP2 Advertise Themselves to LMS2

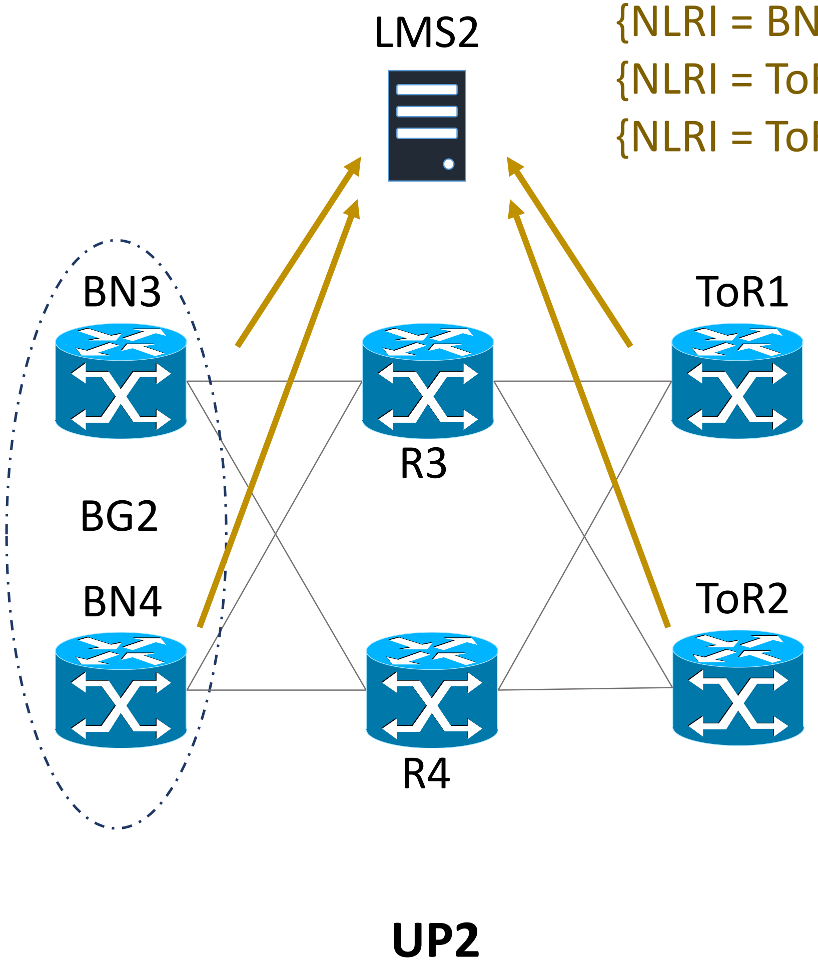
L-BGP Routes (View of LMS2):

{NLRI = BN3-Loopback:NULL, NH = BN3, Com = BG2}

{NLRI = BN4-Loopback:NULL, NH = BN4, Com = BG2}

{NLRI = ToR1-Loopback:NULL, NH = ToR1}

{NLRI = ToR2-Loopback:NULL, NH = ToR2}



BNs of UP2 learn route to ToR1 and ToR2

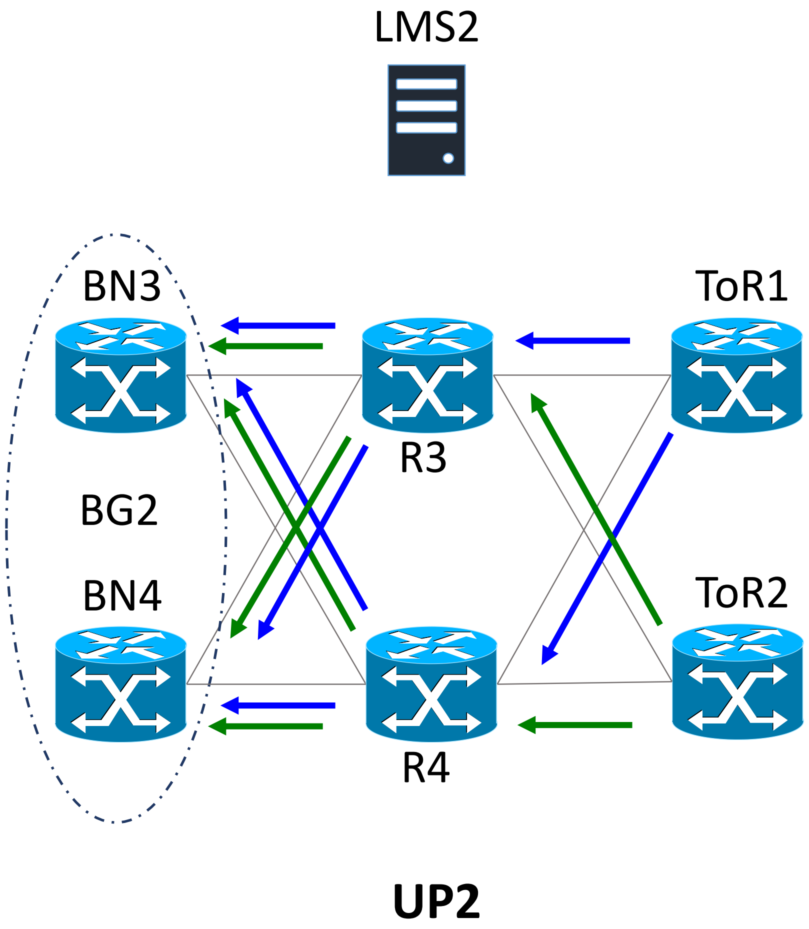
L-BGP Routes (View of BN3):

{NLRI = ToR1-Loopback:L231, NH = R3}
{NLRI = ToR2-Loopback:L232, NH = R3}
{NLRI = ToR1-Loopback:L241, NH = R4}
{NLRI = ToR2-Loopback:L242, NH = R4}

ToR1 and ToR2 do not belong to any Border Groups in this example

L-BGP Routes (View of BN4):

{NLRI = BN3-Loopback:L231, NH = R3}
{NLRI = BN4-Loopback:L232, NH = R3}
{NLRI = BN3-Loopback:L241, NH = R4}
{NLRI = BN4-Loopback:L242, NH = R4}



BN3 learns partition label for ToR1 and ToR2

IP Routes (From BN3 to LMS2):

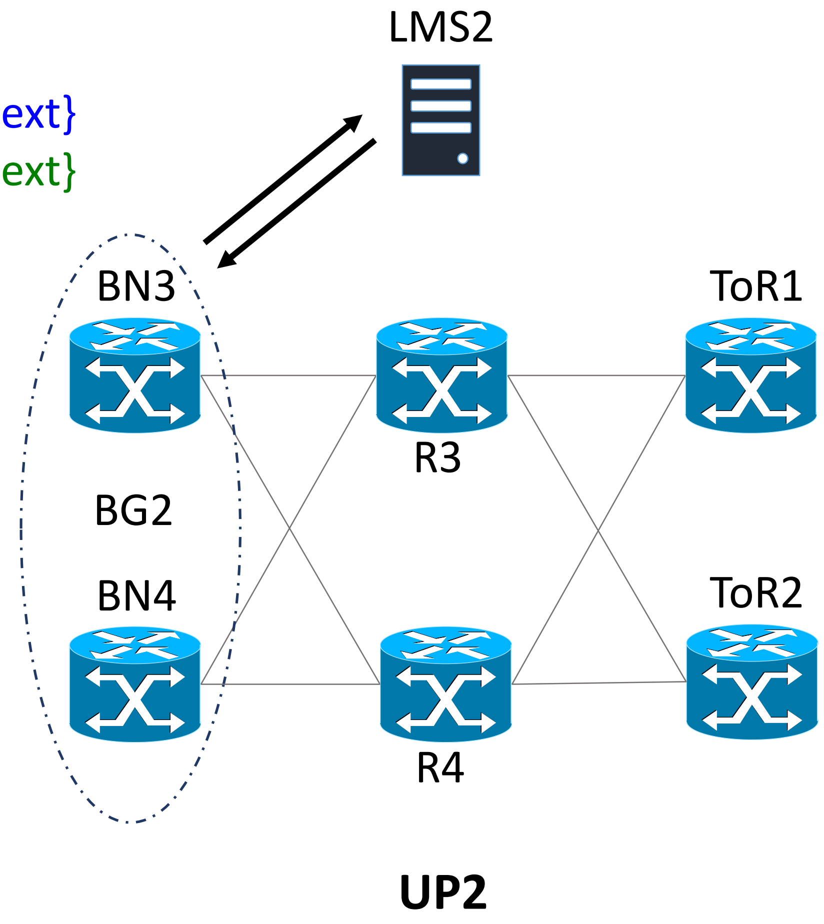
{NLRI = ToR1-Loopback, NH = BN3, Com = BG2, Ext-com = R:UP2-context}

{NLRI = ToR2-Loopback, NH = BN3, Com = BG2, Ext-com = R:UP2-context}

L-BGP Route (from LMS2 to BN3):

{NLRI = ToR1-Loopback:PL21, NH = BN3, Com = BG2, Ext-com = 0:UP1-context}

{NLRI = ToR2-Loopback:PL22, NH = BN3, Com = BG2, Ext-com = 0:UP1-context}



Summary and Next steps

- Summary:
  - Partitioning is a key aspect for scaling
  - BGP is natural glue to connect the partitions
  - New extended community allows to support underlay partition in an efficient and clean way, similar as L3VPN, and supports brownfield deployment well
  - BGP is used as protocol to request and learn the operator assigned labels
  - The procedure defined here can be used for any partition technology
- Next Steps
  - Gather feedback and welcome contributions from the working group
  - Asking for working group adoption after further revision