BGP MultiNextHop Attribute

https://datatracker.ietf.org/doc/html/draft-kaliraj-idr-multinexthop-attribute-09

2023 IETF 117

Kaliraj Vairavakkalai
(on behalf of Co-Authors)

Juniper Networks
July 24, 2023

Agenda

- Changes to the draft – since IDR interim October 2022.
- Recap
  - Background and Problem statement.
  - MultiNextHop Attribute – bird's eye view
- Usecases illustration
  - Signaling WECMP in a scalable manner
  - LB to Multiple CEs in a L3VPN VRF
  - Avoid Label oscillation between Multihomed PEs, Per next hop Label
- Next Steps

Changes to the draft – since IDR interim Oct-2022

- ❑ Removal of Propagation-Scope-Checker.
  - draft-ietf-idr-bgp-attribute-announcement will solve that problem per attribute scope.
  - We plan to introduce a per route propagation scope checker, in a future draft.

- ❑ Removal of ‘Domain Local Preference’
  - draft-uttaro-idr-oad takes up this problem.
  - Better to keep MNH lean, confine to Forwarding Info only.

- ❑ Editorial changes.

- ❑ Added Illustration for a Usecase.
  - Signaling WECMP in a scalable manner
  - LB to Multiple CEs in a L3VPN VRF

- ❑ IANA section review updates.

Background: Expressing nexthops in BGP (Recap)

- What is a nexthop?
  - Instructions on how to forward a payload specified in BGP NLRI.

Nexthop information is extracted from BGP PDU/Route from various portions:

- Endpoint Identifier (Where to forward?)
  - Nexthop attribute (code 3)
  - MP_REACH_NLRI attribute (code 14) : “Network Address of Next Hop”
  - Redirect to IP extended community attribute.
  - Tunnel Encap Attribute.
  - Color-only community attribute.
  - Redirect to VRF extended community attribute.
- Encap to use:
  - MP_REACH_NLRI attribute (code 14) : “Label in NLRI portion”
  - Prefix-SID attribute.
  - Tunnel Encap Attribute.
  - Repair-Label attribute.
  - **Secondary-Label attribute.** (new since idr interim, Oct-2022)
  - **FSv2 Redirect to * actions.**
- Constraints:
  - Color community or Mapping community attribute.
  - Link bandwidth community attribute.

Problems (Recap)

- ❑ Inability to advertise more than one nexthop in a route.
- ❑ Not easily extensible to newer endpoint types, encapsulation types.
- ❑ Addpath unable to express relationship between different nexthops (active/backup, UCMP etc), Scaling heavy.
- ❑ Inability to signal encap-information uniformly across families (e.g. cannot signal Labels for SAFI 1 routes).
- ❑ Inability to signal multiple labels in a route.

  Helpful in some multihomed cases to avoid label oscillation.

- ❑ Semantics of a downstream allocated label is not known to receiver.

  This info may be useful for some scenarios, e.g. network visualization, EPE decisions.

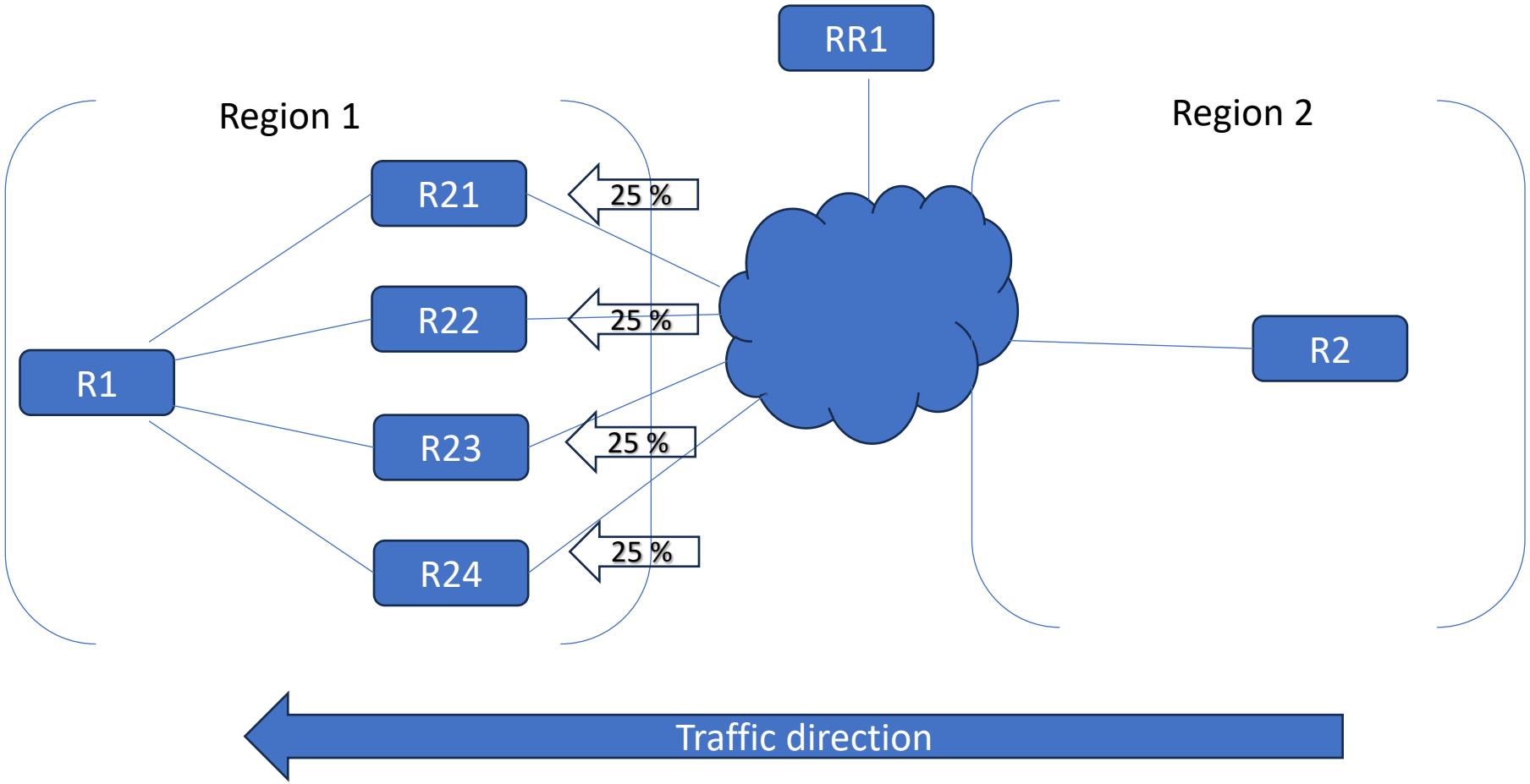
These problems are solved by MultiNexthop Attribute.

MultiNexthop (MNH) attribute – bird's eye view (Recap)

```
MNH Attribute: {
    PrimaryPath {
        [Forwarding Instruction 1],
         ..
        [Forwarding Instruction n]
    }
    BackupPath {
        [Forwarding Instruction 1],
         ..
        [Forwarding Instruction n]
    }
    LabelDescriptor {
        [Forwarding Instruction 1],
         ..
        [Forwarding Instruction n]
    }
}
```

```
Forwarding Instruction : {
    FwdAction, FwdArguments
}
```

Usecase1: Signaling WECMP in a scalable manner



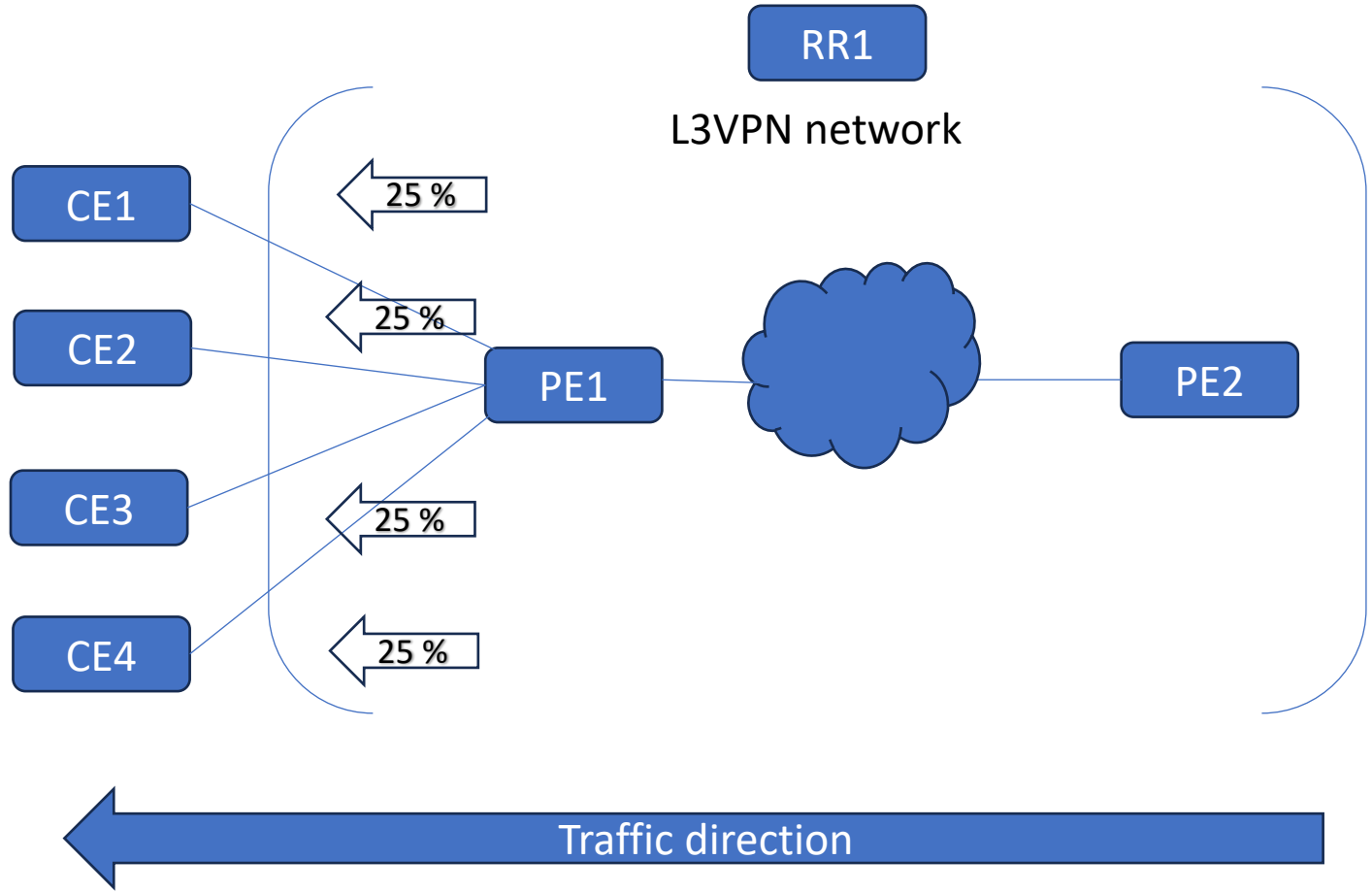
MNH Layout for Usecase1

```
One BGP route with:
MNH Attribute: {
    PrimaryPath {
        [Forward, “R21”, “25%”],
        [Forward, “R22”, “25%”],
        [Forward, “R23”, “25%”],
        [Forward, “R24”, “25%”]
    }
}
```

- Reduces RIB out scale at RR by 4 times.
- Reduces Loc RIB scale at ingress node R2 by 4 times.
- Since 1 route advertisement carries all information instead of 4 addpath advertisements

.

Usecase2: LB to Multiple CEs in a L3VPN VRF



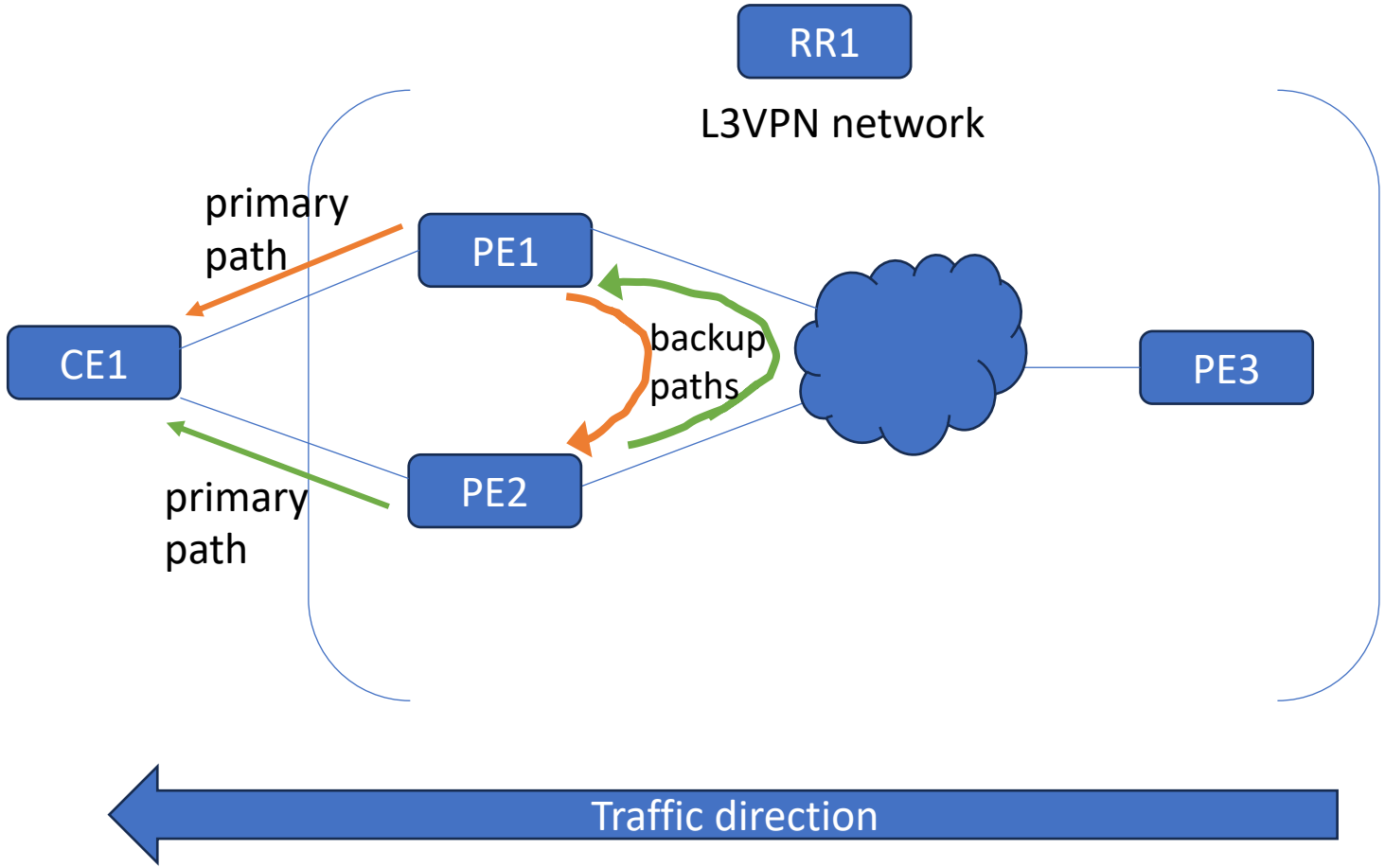
MNH Layout for Usecase2

```
One BGP route with:
MNH Attribute: {
    PrimaryPath {
        [Push “VL_CE1”, “PE1”]
        [Push “VL_CE2”, “PE1”]
        [Push “VL_CE3”, “PE1”]
        [Push “VL_CE4”, “PE1”]
    }
}
```

- Possible to advertise paths to multihomed CEs, without needing additional RDs or Addpath.
- Better LB entropy in network for traffic towards CEs. Without any increase in RIB scale.
- This can be achieved in conjunction with avoiding Label Oscillation too (next usecase)

.

Usecase3: Avoid Label oscillation between Multihomed PEs, Per next hop Label



MNH Layout for Usecase3

PE1 MPLS FIB:

```
VL11: Pop, Fwd to CE1

VL12: Prim {Pop, Fwd to CE1}
      Bkp {BackupPath fm PE2}
```

PE2 MPLS FIB:

```
VL21: Pop, Fwd to CE1

VL22: Prim {Pop, Fwd to CE1}
      Bkp {BackupPath fm PE1}
```

PE1 advertised BGP route:

```
MNH Attribute: {
    PrimaryPath {
        [Push "VL12", "PE1"],
    }
    BackupPath {
        [Push "VL11", "PE1"],
    }
}
```

PE2 advertised BGP route:

```
MNH Attribute: {
    PrimaryPath {
        [Push "VL22", "PE2"],
    }
    BackupPath {
        [Push "VL21", "PE2"],
    }
}
```

- ***Avoids Cyclic dependency between the multihomed PEs.***
- ***Label allocation doesn't depend on other PEs' PrimaryPath.***
- ***BackupPath Label depends on only primary CE paths.***

Next Steps

- ❑ Request for WG Adoption
- ❑ Work on Implementation, attempt to solve these customer use cases.
- ❑ Improve draft by more input from WG. Request more reviews.

Thank you.