BGP MultiNextHop Attribute

https://datatracker.ietf.org/doc/html/draft-kaliraj-idr-multinexthop-attribute-10

2023 IETF 118

Kaliraj Vairavakkalai
(on behalf of Co-Authors)

Juniper Networks

Nov 10, 2023

Agenda

- Background and Problem statement (recap).
- MultiNextHop Attribute – bird’s eye view (recap)
- Changes to the draft – since IDR 117.
- Rethink whether MNH capability negotiation is needed.
- Usecase illustration
  - 4PE/6PE - Signal MPLS Label for SAFI 1 routes

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
}
```

Changes to the draft – since IDR 117

- ❑ Moved some usecases from draft bgp-ct to this draft.
  - Signaling Intent over PE-CE Attachment Circuit
    - Using DSCP in MultiNexthop Attribute
    - MPLS-enabled CE

- ❑ Added Illustration for a new Usecase.
  - 4PE – Signaling MPLS Label for IPv4 Unicast Routes (explained in this session)

Rethink: whether MNH Capability is needed.

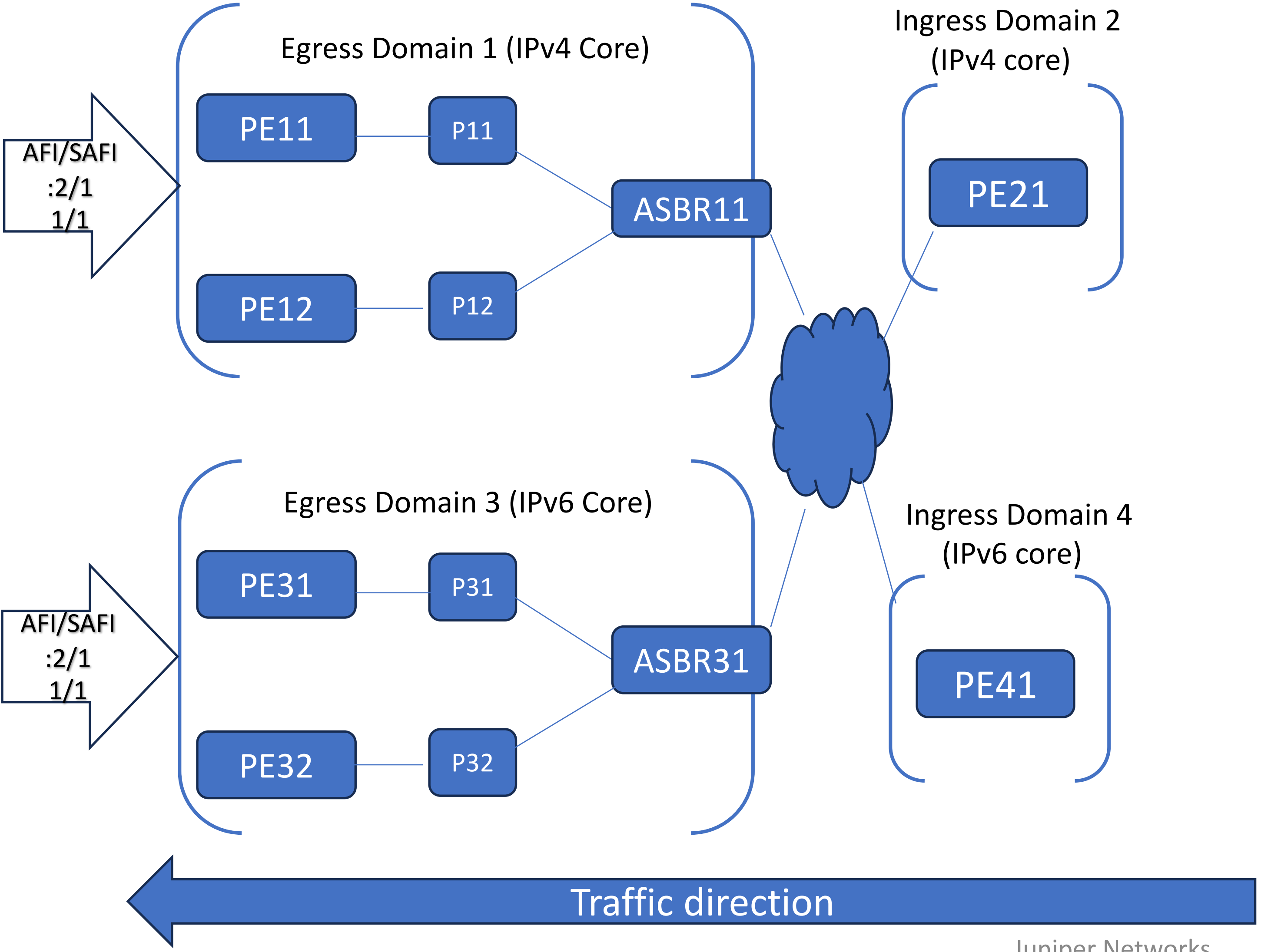
❑ Being a negotiated Open Capability causes BGP session flap whenever config changes.

❑ Optional Non-Transitive attribute stops propagation as unrecognized attribute.

❑ Adding Receive side rule may be enough to stop unintended propagation across supported node also.

> If the MNH attribute is received on a BGP session where MNH support is not enabled, the attribute MUST be treated as Unrecognized non-transitive attribute. This rule provides additional protection against unintended propagation of this attribute, when both BGP speakers understand MNH but receiver has not enabled the support. A RFC3392 Capability is not used for this purpose, because it would cause BGP session reset whenever MNH support config is changed.

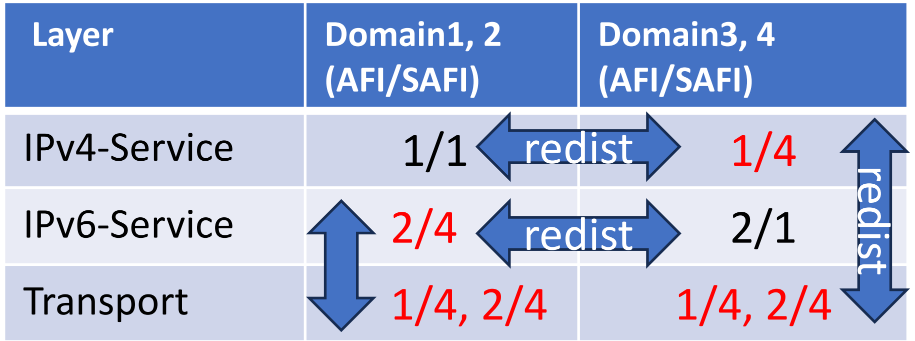
Usecase: 4PE/6PE - Signaling MPLS Label for SAFI 1 Routes



Approach with MNH
(no cross family redistribution)

Layer	Domain1, 2 (AFI/SAFI)	Domain3,4 (AFI/SAFI)
IPv4-Service	1/1	1/1 + MNH
IPv6-Service	2/1 + MNH	2/1
Transport	1/4, 2/4	1/4 , 2/4

Approach with Overloading SAFI 4
(redistribution across families all layers : risky)



Layer	Domain1, 2 (AFI/SAFI)	Domain3, 4 (AFI/SAFI)
IPv4-Service	1/1	1/4
IPv6-Service	2/4	2/1
Transport	1/4, 2/4	1/4, 2/4

redist
redist
redist

MNH Layout for 4PE Usecase (IPv6 core)

```
AFI/SAFI 1/1 BGP route with:
  MNH Attribute: {
      PrimaryPath {
           [Forward, “::ffff:1.1.1.1”, “Label 0”],
      }
   }
```

❑ “Explicit NULL” Label Signaled using MNH on a AFI/SAFI : 1/1 route only by Egress-PEs who’s PHP-nodes need it.

❑ Consistent Service layer address family (AFI/SAFI : 1/1) across the network. No redistribution between AFs needed, which is error prone and risky.

❑ Consistent Transport layer address family (AFI/SAFI : 2/4), that can span across IPv4 domain over (IPv4-MPLS tunnels using IPv4-mapped-IPv6 nexthops) as-well as pure-IPv6 domain (over IPv6-MPLS tunnels).

MNH Layout for 4PE Usecase (IPv6 core)

```
AFI/SAFI 1/1 BGP route with:
  MNH Attribute: {
      PrimaryPath {
          [Forward, “2::2”, “Label 0”],
      }
   }
```

□ “Explicit NULL” Label Signaled using MNH on a AFI/SAFI : 1/1 route only by Egress-PEs who’s PHP-nodes need it.

□ Consistent Service layer address family (AFI/SAFI : 1/1) across the network. No redistribution between AFs needed, which is error prone and risky.

□ Consistent Transport layer address family (AFI/SAFI : 2/4), that can span across IPv4 domain over (IPv4-MPLS tunnels using IPv4-mapped-IPv6 nexthops) as-well as pure-IPv6 domain (over IPv6-MPLS tunnels).

Next Steps

❑ WG Adoption

❑ Work on Implementation.

❑ Improve draft by more input from WG. Request more reviews.

Thank you.