BGP MultiNexthop Attribute


IETF IDR 115

Kaliraj Vairavakkalai

Juniper Networks
Nov 07, 2022
Agenda

• Background
• Problem statement.
• MULTI_NEXT_HOP Attribute
  -- Propagation Scope.
  – layout and organization
  -- Error handling.
• Use cases discussed inline..
  – a uniform API to receiver’s FIB.
  -- DOMAIN_LOCAL_PREF.
  – Label oscillation avoidance.
Background: Expressing nexthops in BGP.

- 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.

- Constraints:
  - Color community or Mapping community attribute.
  - Link bandwidth community attribute.
Background: expressing multiplicity.

- **Addpath**
  - Advertise multiple paths for one prefix each with its own nexthop (previous slide).
  - Increased RIB scale. Specifically RIB-out.
  - Unspecified for most of the mechanisms carrying endpoint-identifier in previous slide. Works for Nexthop attribute (code 3) and MP_REACH_NLRI attribute (code 14)

- **Multipath, PIC**
  - Info from Multiple routes is consumed in conjunction with local config

**Observation:**
These mechanisms have organically grown over the period, and information is spread across:

- Different portions of a BGP route (NLRI, and different attributes)
- Local configuration.
Problems.

• Inability to advertise more than one nexthop in a route.

• Not easily extensible to newer endpoint types, encapsulation types.

• Even with addpath, inability to express relationship between the different route nexthops (active/backup, UCMP etc).

  *These properties are important to use BGP as an API to receiver’s FIB for both IP and MPLS routes.*

• Inability to signal encaps-information uniformly for different address families (e.g. cannot signal Labels for SAFI 1 routes).

  *Being able to do so can confine service routes to the edge, and make the core light weight. Extending the principles of BGP free core.*
Problems (contd).

• Inability to express 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._

A problem slightly unrelated to nexthops:

• Local-preference is designed to be used in one administrative domain (AS, Confed) but doesn’t work for option-C domains, because it consists of multiple AS, even though a single admin control.

• Lack of Scoping control for attribute advertisement within option-C domain scope.

These problems are attempted to be solved by MultiNexthop Attribute. 
Lets see how..
MultiNexthop Attribute (MULTI_NEXT_HOP)

• MNH is an Optional Nontransitive attribute.

• Usage negotiated with a new BGP capability.

• TLVized format extensible for newer endpoint types, encapsulation types, forwarding actions, argument types.

• Can carry 1 or more nexthop instructions.

• Can be used to enable BGP based API to the receiver’s FIB. For IP or Upstream allocated MPLS routes.
MultiNexthop attribute – bird’s eye view.

- Propagation Scope checker controls attribute propagation scope.
- Nexthop Forwarding Information TLV: The Nexthop.
- Fowarding Instruction TLV: The Nexthop Leg/Element.
**Propagation Scope.**

- NonTransitive, will not unintentionally leak to Internet.

- Carries Advertising PNH (BGP Protocol Nexthop), which can be used to know if MNH is valid, added by the router who rewrote nexthop.
  
  *Q: Do we need this anymore, since the propagation scope is made conservative?*

- Even amongst speakers that understand MNH, advertisement is controlled by a “Propagation scope checker” (PSC).
  
  - PSC flag I: When Set allow advertisement to IBGP peers.
  - PSC flag C: When Set allow advertisement to Confed-EBGP.
  - PSC flag E: When Set allow advertisement to EBGP peers in Allowed-AS list.
  - PSC Allowed-AS list: list of (4 octect) AS numbers that are under same administrative control.

  This enables DOMAIN_LOCAL_PREF to be used in option-C domain scope.
**MNH TLV**

Types:

- 1: Upstream signaled primary forwarding path.
- 2: Upstream signaled backup forwarding path (to avoid label oscillation problem)
- 4: Downstream signaled Label Descriptor.

All above Types contain Nexthop Forwarding Information TLV.

- 3: Domain Local Preference (DOMAIN_LOCAL_PREF)

This Type contains Domain Local Preference (4 byte value). It is to be used during Path Selection in place of LOCAL_PREF attribute, within an option-C domain.

- Unknown types: are propagated, if MNH is propagated.

*Q: Perhaps add indication like Partial bit?*
Nexthop Forwaring Information TLV

• This TLV describes a Nexthop.

• It contains
  • Num Nexthops: Number of Nexthop Leg Elements.
  • one or more Nexthop Leg elements (Forwarding Instruction TLVs)
Forwarding Instruction TLV

- This TLV describes a Nexthop Leg.
- It comprises of:
  - FwdAction.
    - Forward
    - Pop-And-Forward
    - Swap
    - Push
    - Pop-and-Lookup
    - Replicate
  - One or more Arguments (Forwarding Argument TLV)
Forwarding Argument TLV (1/3)

• Endpoint Identifier:
  • IPv4 Address,
  • IPv6 Address,
  • MPLS Label (Upstream allocated or global scope),
  • Fwd Context RD, identifies a receiver on the receiving node.
  • Fwd Context RT, identifies a receiver on the receiving node.
Forwarding Argument TLV (2/3)

• Path Constraints:
  • Proximity check
    • S bit: Restrict to Singlehop path
    • M bit: Expect Multihop path.
    • When both S and M bits are set, M bit behavior takes precedence.
    • When both Clear, proximity derived from peer type (EBGP is singlehop, IBGP is multihop)
  • Transport Class ID (Color)
  • Load balance factor (for UCMP)
Forwarding Argument TLV (3/3)

- Payload encapsulation info signaling
  - MPLS Label Info (contains ELC as flag)
  - SR MPLS label Index Info
  - SRv6 SID info

- Endpoint attributes advertisement
  - Available Bandwidth (8 octets, bits per sec)
Error handling

• Follows the 'Attribute discard' approach described in [RFC7606]
• Try to deal gracefully with errors, as much as possible.
• Unknown TLVs are ignored, gracefully. With enough diagnostic data.
• For a ‘FwdAction’, if extraneous arguments are ignored. If minimum required arguments not available, then the Fwd-Instruction-TLV is ignored.
• If Num-Nexthops in NFI TLV is not acceptable to receiver, he ignores the MNH attribute. Attribute discard approach.
• More details in Section 6 of the draft.
References:

Thank you.