BGP Signaled Multicast

draft-zzhang-bess-bgp-multicast-01

Zhaohui (Jeffrey) Zhang, Juniper

Keyur Patel, Arrcus

IJsbrand Wijnands, Cisco

Arkadiy Gulko, Thomson Reuters

98th ITEF, Chicago

Multicast: fear/dislike & necessity

- Many operators, especially DC ones, do not want to burden their infrastructure with multicast trees
  - They can live with ingress replication for multicast traffic
  - They do not like the following aspects of multicast trees
    - Per-tree state
    - PIM soft-state refresh overhead
    - PIM-ASM complexity due to shared-to-source tree switch
    - Yet another protocol to set up the trees

- Nonetheless, some operators have a lot of mission-critical multicast traffic, and still need the efficiency gains of having multicast trees in the infrastructure
  - at least until BIER arrives ᄊ

BGP Signaled Multicast: What & Why

- Use BGP to signal multicast
    - Use as a replacement for PIM
        - (s,g)/(*,g) unidirectional/bidirectional trees
            - Optionally with MPLS data plane
    - Use as a replacement for mLDP
        - Use mLDP FEC (<root, opaque_value>) to identify tree
- Why?
    - Remove PIM-ASM complexities & soft state
        - PIM-Port only removed soft state and deployment has been limited
        - PIM-SSM removes ASM complexities but requires good source discovery methods
    - Consolidate to BGP signaling
        - Single, scalable protocol for unicast/multicast, labeled/unlabeled

How to signal tree/tunnel using BGP

- Use receiver-initiated “joins” - Leaf A-D routes in C-MCAST SAFI
  - Propagated over hop by hop EBGP/IBGP sessions or through RRs
- Each node determines upstream hop by using same RPF procedure as PIM/mLDP
- Leaf A-D routes serve the purpose of PIM Join or mLDP P2MP label mapping
  - NLRI encodes (s,g)/(*,g) or mLDP FEC
  - Route Target identifies Upstream node
  - Routes processed by upstream node and not propagated further
    - A new route with different NLRI is originated for the next node in the tree
  - Tunnel Encapsulation Attribute carries forwarding information
    - In case of labeled tree/tunnel, or
    - If downstream/upstream are not directly connected
  - For MP2MP labeled tunnels, S-PMSI/Leaf A-D routes serve the purpose of mLDP MP2MP-U/MP2MP-D label mappings
- For ASM, source specific trees are set up after source discovery via Source Active (SA) A-D routes, avoiding RP/shared-trees

Signal mLDP Tunnels

- Signal entire or part of an mLDP tunnel using BGP
  - Transition from an existing mLDP deployment w/o changing anything else
  - Reuse mLDP tunnel identification
- Where a P2MP Label Mapping or an MP2MP-D Label Mapping would be sent, a Leaf A-D route is sent instead
  - Unsolicited, but as if an S-PMSI had been received
- Where an MP2MP-U Label Mapping would be sent, an S-PMSI A-D route is sent instead
  - Only MP2MP-D FEC is used
  - Route Targets correspond to the set of downstream neighbors

```
  +-     +-----------------------------------+
  |      |    Route Type - 4 (Leaf A-D)      |
  |      +-----------------------------------+
  |      |     Length (1 octet)              |
  |   +- +-----------------------------------+ --+
  |   |  |Route Type - 0x43 (mLDP S-PMSI A-D)|   |
L | L |  +-----------------------------------+   | S
E | E |  |     Length (1 octet)              |   | |
A | A |  +-----------------------------------+   | P
F | F |  |     P2MP or MP2MP-D FEC           |   | M
  |   |  +        ......                     |   | S
N | R |  |                                   |   | I
L | O |  +                                   +   |
R | U |  |                                   |   | N
I | T |  +                                   +   | L
  | E |  |                                   |   | R
  |   |  +-----------------------------------+   | I
  | K |  |  Upstream Router's IP Address     |   |
  | E |  +-----------------------------------+ --+
  | Y |  |  Originating Router's IP Addrress |
  +-  +- +-----------------------------------+
```

Signal (s,g)/(*,g) Trees

- Where a PIM join would be sent, a Leaf A-D route is sent instead
    - Unsolicited, but as if an S-PMSI had been received
- In case of labeled bidirectional trees, an S-PMSI A-D route is sent to signal the label for upstream path
- (*,g) unidirectional tree allowed when sources can send traffic to root of the tree w/o intersecting the tree
    - Source tree is not used in this case

```
  +-      +-----------------------------------+
  |       |    Route Type - 4 (Leaf A-D)      |
  |       +-----------------------------------+
  |       |     Length (1 octet)              |
  |    +- +-----------------------------------+ --+
  |    |  |    Route Type - 3 (S-PMSI A-D)    |   |
L | L  |  +-----------------------------------+   | S
E | E  |  |     Length (1 octet)              |   | |
A | A  |  +-----------------------------------+   | P
F | F  |  | Multicast Source Length (1 octet) |   | M
  |    |  +-----------------------------------+   | S
N | R  |  |  Multicast Source (variable)      |   | I
L | O  |  +-----------------------------------+   |
R | U  |  |  Multicast Group Length (1 octet) |   | N
I | T  |  +-----------------------------------+   | L
  | E  |  |  Multicast Group   (variable)     |   | R
  |    |  +-----------------------------------+   | I
  | K  |  |  Upstream Router's IP Address     |   |
  | E  |  +-----------------------------------+ --+
  | Y  |  |  Originating Router's IP Addrress |
  +-   +- +-----------------------------------+
```

Source Discovery for ASM

- First Hop Routers (FHRs) advertise SA routes
  - Upon receiving locally originated traffic
- Last Hop Routers (LHRs) receive SA routes and join source specific trees
- Similar to MSDP method, but:
  - Extended from among RPs to among FHRs and LHRs
  - With BGP advantages:
    - No periodical refreshing
    - No RPF checks for SA propagation
    - RRs and Route Target Constrain (RTC) can be used to avoid flooding SA routes
      - FHRs attach a RT that encodes the group address and advertise to RRs
      - LHRs advertise RT Membership NLRIs that encode the above mentioned RT for groups that they're interested in
      - SAs are only advertised to interested LHRs due to the RTC mechanism

Incremental Transition

- For mLDP or PIM-SSM replacement, transition can independently happen at any node
  - If the upstream neighbor can support BGP multicast signaling, then use it
- For PIM-ASM replacement, first upgrade the RPs so that they can advertise SA routes. After that each node can independently transition
  - If an upgraded node receives (*,g) PIM join, and its upstream supports BGP multicast signaling, it behaves as if it were a LHR
    - Terminate (*,g) join
    - Send RT Membership NRLI corresponding to the group
    - Establish source trees after receiving corresponding SA routes.

Next steps

- Add details like handling of neighbors not directly connected
- Seek comments and feedback