Agenda

• BGP Signaled Multicast Review
  • draft-zzhang-bess-bgp-multicast
  • Presented in 98th IETF

• Controller-based BGP Multicast Signaling
  • draft-zzhang-bess-bgp-multicast-controller

• Summary
Multicast: complexity, fear/dislike, necessity/reality

• Many operators 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 ^
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 soft state and ASM complexities
    • 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
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 peer 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.
BGP hop-by-hop signaled multicast

- Each router independently determines its upstream and send Leaf A-D routes to it
  - Much like PIM/mLDP
- The routes may be reflected by a RR
Controller Based Signaling

- Instead of hop-by-hop signaling initiated from LHRs, an intelligent controller can figure out the entire tree/tunnel and signal to all routers on the tree/tunnel
  - Same Leaf A-D routes as in hop-by-hop case
  - The controller does not have to peer with each router directly – could be via other RRs
- Each router simply sets up forwarding state accordingly
  - No need for PIM/mLDP-like procedures to figure out upstream
  - No need to send message upstream and receive message from downstream
Differences from hop-by-hop case

- A single Leaf A-D route from the controller can signal multiple downstream routers to the same upstream
  - A new Composite Tunnel in Tunnel Encap Attribute (TEA) means traffic is to be sent out of all component tunnels represented by its sub-TLVs
  - Forwarding info used to receive traffic from its upstream is also signaled via the same Leaf A-D route
- Labels could be allocated from:
  1. A controller’s own local label space
  2. A common SRGB
  3. Each router’s SRLB
- With the first two label allocation options, per-tree/direction label could be used
  - Per-tree labels could be use for unidirectional trees
  - Per-<tree, direction> labels could be used for bidirectional trees
  - Neighbor-based RPF is needed in data plane to use per-tree/direction labels
    - <neighbor-identifying label, per-tree/direction label> stack for #2
- In the first option, a controller-identifying label is needed
  - <controller-identifying label, neighbor-identifying label, per-tree/direction label>
  - <controller-identifying label, tree-identifying label>
Consistency

• Multiple controllers could be used
  • Each could calculate and signal independently
  • As long as all the routers on a tree/tunnel choose routes from the same controller it’s fine
    • In labeled case, even if they choose differently, there is no traffic looping
    • In unlabeled case, routers must choose based on controllers’ address to ensure consistency and prevent loops

• Topology change in bidirectional case
  • Since upstream/downstream update their state (per signaling from the controller) independently, transient loops may happen
    • In the unlabeled case, order of updates must be ensured - out of scope
    • In the labeled case, per-tree label cannot be used
      • Per-<tree,direction> label may be fine
  • Not an issue for unidirectional case
Summary

• BGP-signaled multicast could replace PIM/mLDP signaling:
  • Removes PIM refreshes & PIM-ASM complexity
  • Consolidates to BGP signaling
  • draft-zzhang-bess-bgp-multicast

• BGP-controller-signaled multicast is well suited for SR networks
  • Tree calculation delegated to omniscient/omnipotent controllers
    • Based on many factors/constraints/algorithms
  • Router operation is simplified
    • Simple forwarding state programming based on BGP messages
    • Per-tree/direction labels may ease monitoring & troubleshooting
      • May not be supported due to software/hardware capabilities
  • draft-zzhang-bess-bgp-multicast-controller