Layer 3 Multicast Routing on NX-OS

In This Section

- Multicast Overview
- IGMP Multicast Control Plane
- PIM Multicast Control Plane
- Multicast Data Plane

What is Multicast?

- Multicast is data transmission to a group of destinations simultaneously
  - I.e one to many transmission
- As opposed to...
  - Unicast – one to one transmission
  - Broadcast – one to all transmission
  - Anycast – one to nearest transmission

How Multicast Works

- Source application sends multicast with “group” destination address
- Interested receivers “join” group address by signaling routers on LAN
- Routers build loop free “tree” from sender to receivers
- Network segments without receivers will not receive traffic for group

IPv4 Multicast Components

- Multicast can be broken down into three main components
- Group Addressing
  - Layer 3 addressing
  - Layer 2 addressing
- Control Plane
  - IGMP, PIM, MSDP, MBGP
- Data Plane
  - Reverse Path Forwarding (RPF)
  - Multicast Routing Table (MRIB/MFIB)

Multicast Group Addressing

- Multicast "group" is an address agreed upon between the sender and receivers for a particular feed
  - Source sends traffic to destination address of the group
  - Receivers listen for traffic going to group address
- Traffic is always sent ***to*** a group, never ***from***
- Groups use both layer 3 and layer 2 addresses

IPv4 Multicast Addressing

- IPv4 multicast uses Class “D” Addresses
  - 224.0.0.0/4 (224.0.0.0–239.255.255.255)
- Includes reserved ranges
  - Link-local Addresses
    - 224.0.0.0/24 (224.0.0.0 - 224.0.0.255)
  - Source Specific Multicast
    - 232.0.0.0/8 (232.0.0.0 - 232.255.255.255)
  - Administratively Scoped
    - 239.0.0.0/8 (239.0.0.0 - 239.255.255.255)

Layer 2 Multicast Addressing

- IPv4 addresses map to MAC addresses to forward on the LAN
  - Allows L2 switches to forward multicast intelligently
  - E.g. don't forward multicast as broadcast
- MAC address range is 01-00-5E-00-00-00 to 01-00-5E-7F-FF-FF
  - First 25 bits are fixed
  - Last 23 bits are mapped from IPv4 address
- Implies overlap between addresses
  - Last 23 bits must be unique to result in unique layer 2 flow

Layer 2 Multicast Address Conversion

- Conversion shortcut
  - Convert IPv4 2nd octet to binary
  - Set the first bit to 0
  - Convert to hex
  - 3rd and 4th octets convert directly to hex
- Example conversions
  - 224.0.0.1
    - 01-00-5E-00-00-01
  - 230.255.1.2
    - 01-00-5E-7F-01-02
  - 239.127.1.2
    - 01-00-5E-7F-01-02

Multicast Control Plane

- Multicast control plane used to determine…
  - Who is sending traffic and to what group(s)
  - Who is receiving traffic and for what group(s)
  - How traffic should be forwarded when it is received
    - The Multicast "Tree"
- Control plane is built with a combination of
  - Host to Router communication (IGMP)
  - Router to Router communication (PIM and MSDP)

Multicast Control Plane – IGMP

- Internet Group Management Protocol (IGMP)
  - Used for receiver to signal routers on the LAN that it wants traffic for a specific group
- Three versions
  - RFC 1112 - Host Extensions for IP Multicasting
  - RFC 2236 - Internet Group Management Protocol, Version 2
  - RFC 3337 - Internet Group Management Protocol, Version 3

IGMPv1

- Uses two message types to signal group membership
  - Host Membership Query
  - Host Membership Report
- Report used by client to “join” a group
- Query used by router to see if members of the group still exist
  - Essentially an idle timer for the group
- Legacy now, replaced by IGMPv2

IGMPv2

- Enhances IGMPv1 by adding
  - Querier election
    - If multiple routers on the segment, who sends queries?
  - Tunable timers
    - Can speed up query response timeouts
  - Group specific queries
    - Query sent to the group address instead of all multicast hosts
  - Explicit leave
    - Speeds up convergence if no other hosts are joined to that group
- Backwards compatible with IGMPv1

IGMPv3

- Used to support Source Specific Multicast (SSM)
  - IGMPv1/v2 only support group specific joins
    - (*,G) join
  - IGMPv3 supports source specific joins
    - (S,G) join
- Implies IGMPv3 receiver must already know about the sender
  - More details on this later

Next Steps From IGMP

- Router knows that host wants traffic for multicast group “G”
- How does it tell the rest of the network to deliver traffic to it for “G” ?
- Multicast “routing” protocols now take over
  - PIM
  - ***Not*** MBGP or MSDP
    - More on this later…

Multicast Control Plane – PIM

- Protocol Independent Multicast (PIM)
  - Router to router communication used to build loop-free "tree" from sender to receiver(s)
- Considered "protocol independent" because it does not advertise its own topology information
  - Implies IGP already runs in the network to build a loop-free topology
- Two versions and two "modes"
  - PIMv1 & PIMv2
  - Sparse Mode & Dense Mode

PIM Modes

- Dense Mode
  - Considered implicit join
  - All traffic unless you say you don't want it
  - Uses Flood & Prune behavior
- Sparse Mode
  - Considered explicit join
  - No traffic unless you ask for it
  - Uses Rendezvous Point (RP) to process join requests
- PIM modes control how the tree is built, and who receives what traffic
  - More detail later...

PIM Sparse Mode

- RFC 4601 - Protocol Independent Multicast - Sparse Mode (PIM-SM)
- Uses "pull" model or "explicit join"
  - Traffic is not flooded unless you ask for it
- Uses both Shared Trees (RPT) and Shortest Path Trees (SPT)
  - Dense mode uses only shortest path/source trees
  - More scalable than dense mode and usually the better design choice

Shared vs. Source Trees

- Multicast tree determines how traffic is routed from sender to receivers
- Source based trees
  - Uses shortest path from sender to receiver
  - Dense mode or sparse mode
- Shared trees
  - Uses shortest path from sender to Rendezvous Point (RP), then shortest path from RP to receiver
  - Sparse mode only
  - Used to eliminate flooding and pruning and make routing table more scalable

PIM Sparse Mode Operation

- Discover PIM neighbors & elect DR
- Discover RP
- Tell RP about sources
- Tell RP about receivers
- Build shared tree from sender to receivers through RP
- Join shortest path tree
- Leave shared tree
- Multicast table maintenance

Rendezvous Point Overview

- RP is used as a reference point for the root of the shared tree
- RP learns about sources through unicast PIM Register messages
  - Register tells the RP about an (S,G)
- RP learns about receivers through PIM Join messages
  - Tells the RP to add an interface to the OIL for (*,G)
- RP is used to merge the two trees together

Learning the RP's Address

- Without the RP...
  - Sources can't register
  - Joins can't be processed
- All routers must agree on the same RP address on a per-group basis
  - Registers and joins are rejected for invalid RPs
- RP address can be assigned
  - Statically
  - Dynamically
    - Auto-RP
    - BSR

PIM Register Message

- As the root of all shared trees, the RP must know about all sources
- When the first-hop router connected to sender hears traffic, a unicast Register message is sent to the RP
  - If multiple first-hop routers, only the DR registers
- If RP accepts this message, it acknowledges with Register Stop and inserts (S,G) into the table
- At this point only DR and RP know (S,G)

PIM Join Message

- As the root of all shared trees, the RP must also know about all receivers
- When a last-hop router receives an IGMP Report, a PIM Join is generated up the reverse path tree towards the RP
- All routers in the reverse path install (*,G) and forward the Join hop-by-hop to the RP
- At this point the RP and all downstream devices towards the receiver know (*,G)

Merging the Trees

- Once the RP knows about both sender and receiver…
  - RP sends a PIM Join message up reverse path to source
- All routers in the reverse path from the RP to the source install (*,G) with OIL pointing towards RP
- Once (S,G) begins to flow, the tree is built end-to-end through the RP

Joining the SPT

- The shared tree is made up of two Shortest Path Trees
  - SPT from receiver to RP
  - SPT from RP to sender
- SPT from receiver to sender may not be the same as the shared tree
  - Result is that Shared Tree is not optimal forwarding
- To fix this, last-hop router…
  - Joins SPT to source with (S,G) Join
  - Leaves the RPT by sending (*,G) Prune to RP
- Can be modified with **ip pim spt-threshold**

Routing Table Maintenance

- Like PIM Dense Mode, PIM Sparse Mode uses State Refresh to ensure that feeds do not timeout
  - (*,G) join sent to RP or up SPT to refresh the OIL
- Sparse Prune message can be used to speed up state information timeout if IGMP Leave is heard from end host

Multicast Data Plane

- Once the tree from sender to receiver(s) is built, traffic begins to flow
- Before forwarding, Data Plane checks occur
  - Reverse Path Forwarding (RPF) check
    - Was traffic received on the correct interface?
  - Multicast Routing Table (MRIB/MFIB)
    - What interface(s) should I forward the packets out?

The RPF Check

- If PIM does not exchange its own topology, how does it know the network is loop free?
  - Multicast packet comes in, router looks at source IP address and incoming interface
  - Unicast routing table (CEF table) is checked for the reverse path back to source address
- RPF logic is...
  - If incoming multicast interface == outgoing unicast interface, RPF check passed
  - If incoming multicast interface != outgoing unicast interface, RPF check fails and packet is dropped

Why Use RPF?

- Assume that IGP has built a loop-free unicast topology
  - If RPF check passes on multicast packets, we can assume the traffic has not looped
  - If RPF check fails it's possible a loop occurred, so traffic is dropped
- RPF is very conservative, but always loop-free

The Multicast Routing Table

- During exchange of PIM messages, routers learn where sources and receivers exist
  - Interface facing upstream towards source is the "incoming interface"
  - Downstream links to receivers are "outgoing interface list" or OIL
  - Split-horizon like behavior – link cannot be in incoming and OIL at same time
- If RPF check passes…
  - Packets flow from incoming interface to all interfaces in the OIL
- More information at…
  - Multicast Forwarding Information Base Overview
  - Verifying IPv4 Multicast Forwarding Using the MFIB

Recommended Resources

- Books
  - Routing TCP/IP Volume II
  - Developing IP Multicast Networks
  - Interdomain Multicast Routing: Practical Juniper Networks and Cisco Systems Solutions
- Online Resources
  - Multicast Technology Documentation
  - IP Multicast SRND
  - IP Multicast Best Practices for Enterprise Customers