

# Multicast in SR Networks

NANOG 81, February-2021

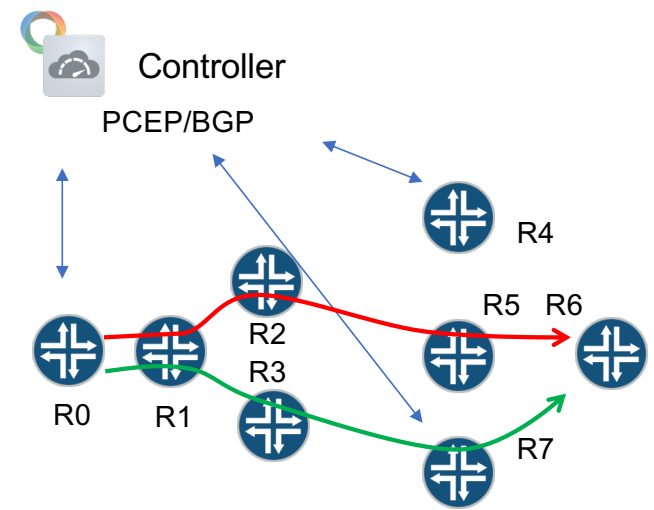
Jeffrey Zhang, Juniper DE

# Agenda

- SR principles and Multicast Options
- Controller Signaled P2MP
  - PCEP/BGP-signaled SR-P2MP
  - BGP-signaled mLDP
- BGP Signaled Multicast
  - Both IP multicast and P2MP
  - With or without controllers
- E2E Inter-region Multicast
- Multicast with Classful Transport

# Segment Routing principles

1. No per-flow/tunnel state inside network
  - Packets have embedded segment list for traffic steering
2. Optional/preferred use of controllers
  - To instruct ingress to embed segment list in packets for per-flow/tunnel traffic steering



R2 R5 payload

Flow 1 packets

R3 R7 payload

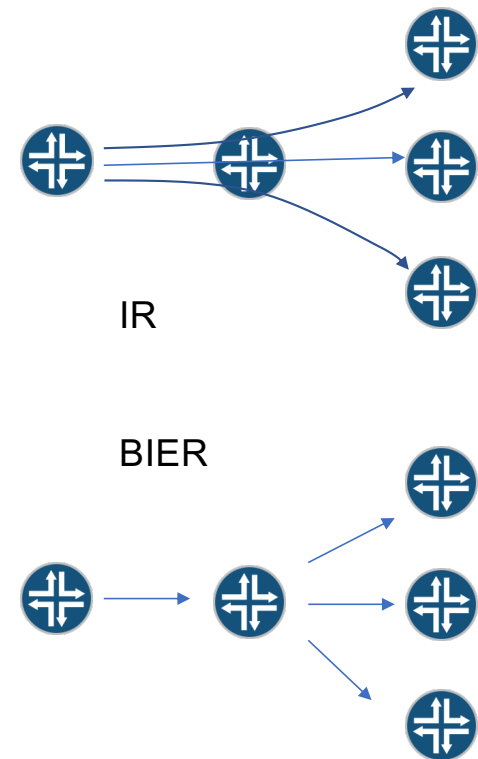
Flow 2 packets

# Multicast per SR principle #1

## 1. No per-tree state inside network

- Ingress Replication (IR)
  - Inefficient but applicable for certain use cases
- BIT Indexed Explicit Replication (BIER)
  - Packets carrying a BitString indicating the targeted edge routers
    - With BIER-TE, the BitString can also specify transit routers
  - **Best** multicast technology though with a new forwarding plane
    - Not covered in this presentation

*Both per SR principle yet independent of SR*



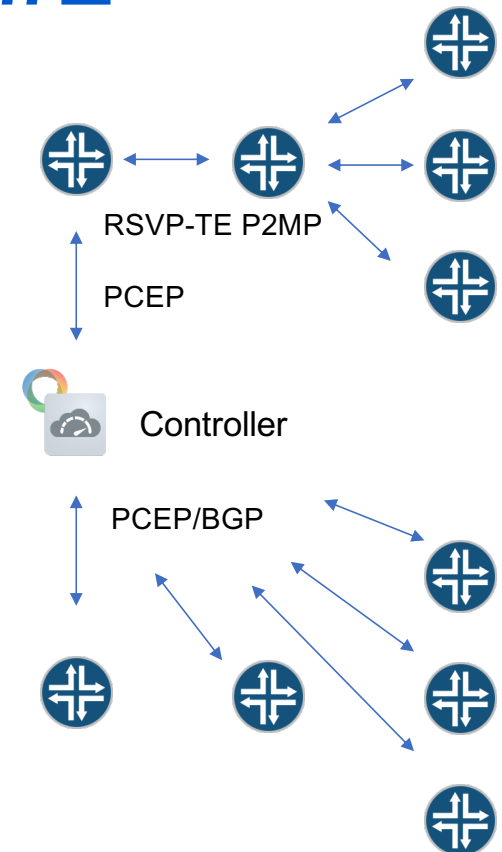
# Multicast per SR principle #2

## 2. Use of controllers

- Controller calculated RSVP-TE P2MP
  - Signaled by RSVP-TE from ingress
- Controller calculated and signaled:
  - SR-P2MP (aka tree-sid)
  - mLDP (signaled via BGP)

## All have identical forwarding plane

- As with legacy mLDP/RSVP-TE P2MP
- Per-tunnel state inside the network
- Label in -> replicated label out



# Multicast Options in SR Networks

- BIER
  - If you care about effective replication with no-state inside the network, and,
  - Most routers support BIER
- Traditional Multicast (PIM/P2MP/IR)
  - If it works well for you
    - You don't need controller, and,
    - You don't mind running PIM/mLDP/RSVP in your SR network for multicast
      - Perfectly ok to run PIM/mLDP/RSVP for multicast while running SR unicast
- Controller Signaled Multicast
  - If you need controller-calculated trees, and/or,
  - You want to remove PIM/mLDP/RSVP
  - Note that you will still have per-tree/tunnel state inside the network

# Agenda

- SR principles and Multicast Options
- **Controller Signaled P2MP**
  - PCEP/BGP-signaled SR-P2MP
  - BGP-signaled mLDP
- BGP Signaled Multicast
  - Both IP multicast and P2MP
  - With or without controllers
- E2E Inter-region Multicast
- Multicast with Classful Transport

# SR-P2MP

- Previously known as Tree-SID
  - Being specified in IETF SPRING/PIM/BESS/PCEP WGs
- Controller signals per <tree, node> Replication Segments to each tree node:
  - Forwarding state identification
    - <root, tree-id, candidate-path, targeted-node>
  - Forwarding information
    - incoming label, outgoing label and branches
- PCEP/BGP-MCAST/BGP-SRTE signaling
  - This presentation focuses on BGP-MCAST



# Controller/BGP Signaled mLDP

- Labeled forwarding just like SR-P2MP and legacy mLDP
- mLDP FEC as tree identification in control plane
  - This is the only relevance to mLDP
    - LDP signaling not used
  - Flexible/extensible identification due to opaque structure
  - Easy transition from existing mLDP deployment, e.g. BGP-MVPN with mLDP
    - no change on MVPN part; just mLDP tunnels signaling changed to BGP
- Signaling via BGP-MCAST
  - From controllers, or
  - Hop-by-hop from leaves towards root

# Agenda

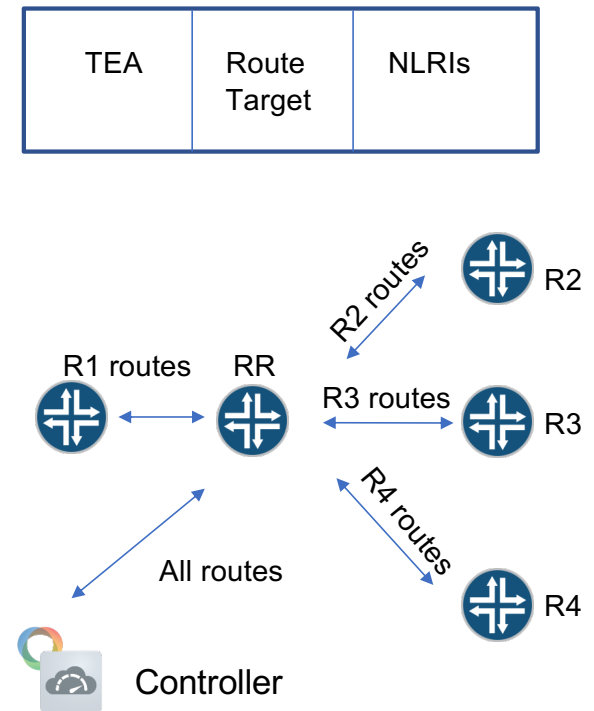
- SR principles and Multicast Options
- Controller Signaled P2MP
  - PCEP/BGP-signaled SR-P2MP
  - BGP-signaled mLDP
- **BGP Signaled Multicast**
  - Both IP multicast and P2MP
  - With or without controllers
- E2E Inter-region Multicast
- Multicast with Classful Transport

# BGP-MCAST Signaling

- Using BGP MCAST-TREE SAFI to signal for:
  - SR-P2MP, mLDP
  - IP Multicast
    - forwarding and tree identification by (source, group)
  - Any potential future types – just using different NLRI types
    - E.g. tree identification by label directly
- Signaling from controllers
  - draft-ietf-bess-bgp-multicast-controller
- Hop-by-hop leaf → root signaling
  - draft-ietf-bess-bgp-multicast

# Essence of BGP Signaling From Controllers

- NLRI encodes the following
  - Tree identification (with different NLRI route types)
    - IP Multicast: source, group
    - SR-P2MP: Candidate Path (via RD), root-ID, tree-ID
    - mLDP: mLDP FEC
  - Targeted Router
- Tunnel Encapsulation Attribute (TEA) encodes forwarding information
  - TEA encodes a list of “tunnels”
  - A “tunnel” identifies the upstream or a downstream replication branch
- A Route Target controls the propagation and importation of the route



# Tunnel Encapsulation Attribute

- Encodes a list of tunnels (maybe of different types)
- Already specified for unicast
  - Traffic ECMP'ed out of one of the tunnels
- Extended for BGP-MCAST
  - Traffic replicated out of all the “tunnels”
  - A tunnel can be of type “AnyEncap”
    - Any way of getting to a downstream node
      - Native, MPLS, GRE, SR Path, whatever
  - A tunnel may have the following:
    - “RPF” sub-tlv, indicating it is for receiving traffic from upstream
    - “Tree Label” sub-tlv, for incoming/outgoing tree-label
    - “Endpoint address”, identifying downstream node/link
    - Maybe other information for more complicated scenarios

Tunnel Encap Attribute

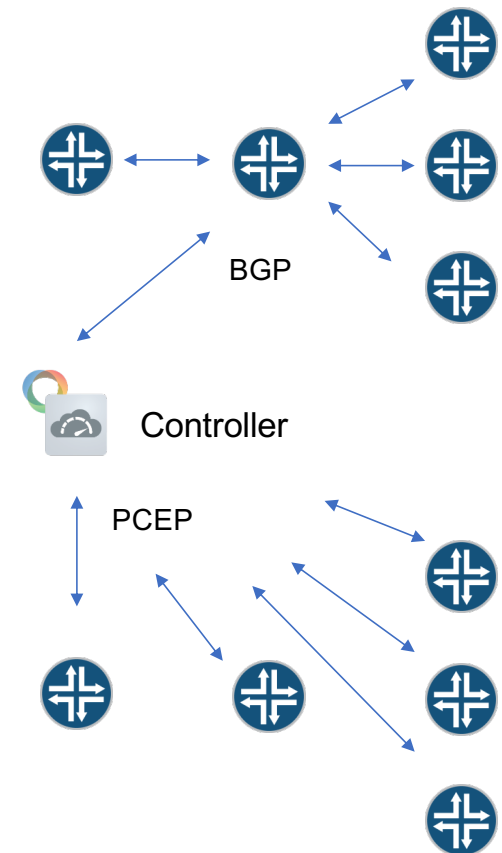
RPF (upstream) 10.1.1.1 Tree Label 100	tunnel1
10.2.1.1 Tree Label 100	tunnel2
10.3.1.1 Tree Label 100	tunnel3
10.4.1.1 Tree Label 200	tunnel4

# SR-MPLS vs. SRv6

- SR/mLDP-P2MP works with both MPLS and SRv6
  - Minimum differences/extensions in TEA for SRv6
- MPLS forwarding plane
  - Forwarding on a branch uses a <transport labels, tree label> stack
    - Transport labels get the packets to the downstream node
      - Explicitly encoded in a TEA tunnel or derived from the tunnel endpoint
    - Transport labels may be empty
      - Upstream and downstream nodes connected directly or via a non-MPLS tunnel
    - Transport labels can be for an SR path
    - Transport label can also represent a p2mp (sub-)tree
- SRv6 forwarding plane
  - Above mentioned label stack becomes an IPv6 address
    - Locator part represents the downstream - corresponding to the transport labels
    - Func/Arg part represents the tree - corresponding to the tree label

# Why BGP-MCAST Is the Best Option

- Single session from controller to **one** of the BGP speakers in the network
  - Vs. one PCEP session to **every** tree node
- Great coverage and extensibility
  - Same procedure for both underlay tunnels and overlay multicast
    - IP multicast, SR-P2MP, mLDP ...
    - BGP-MVPN replacement in certain scenarios
  - Support bidirectional trees
  - Hop-by-hop or Controller-driven
  - E2E inter-region support
  - Integration with classful transport

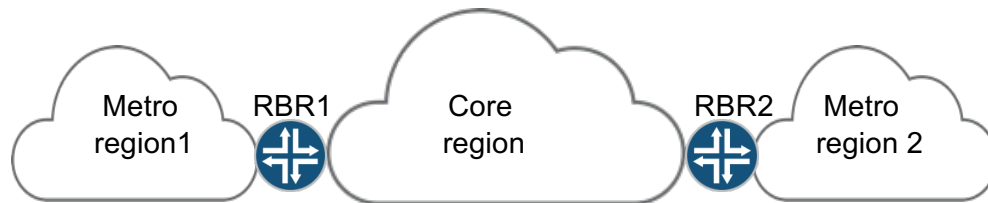


# Agenda

- SR principles and Multicast Options
- Controller Signaled P2MP
  - PCEP/BGP-signaled SR-P2MP
  - BGP-signaled mLDP
- BGP Signaled Multicast
  - Both IP multicast and P2MP
  - With or without controllers
- E2E Inter-region Multicast
- Multicast with Classful Transport



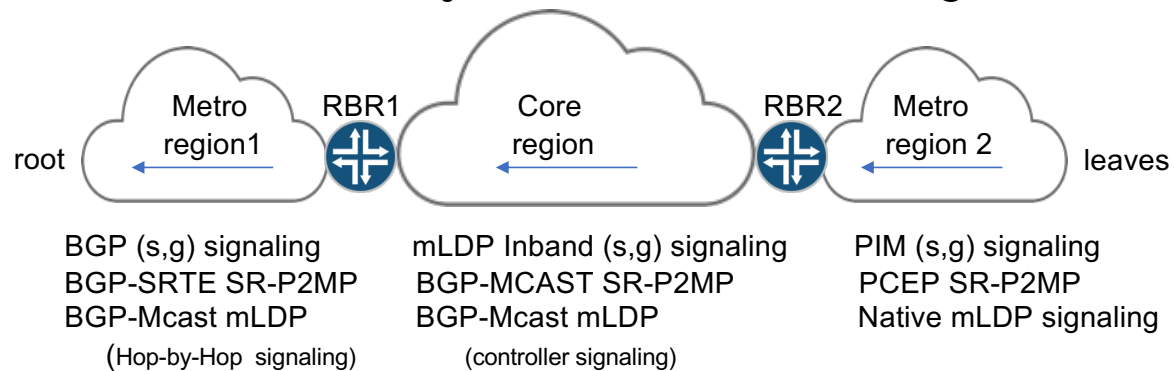
# Inter-region Multicast



- draft-ietf-bess-bgp-multicast, Section 1.2.6
- An E2E IP multicast tree or P2MP tunnel can span multiple regions
  - A region is an AS or an IGP area
  - Different signaling can be used in different regions
- Inband signaling across a region
  - Internal routers in a region maintain state per E2E tree/tunnel
- Overlay signaling over a region
  - Internal routers do not keep state for E2E trees

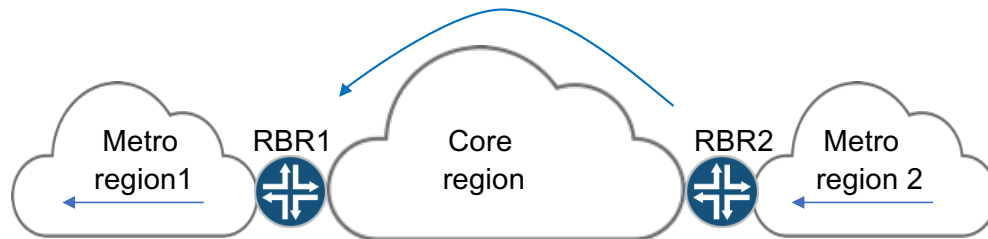
# Inband Signaling across a Region

- Different methods may be used in different regions



- In case of hop-by-hop signaling:
  - A router needs to determine its Reverse Path Forwarding (RPF) neighbor
    - The next-hop neighbor towards the tree root
  - If internal routers don't have route to the root, RPF is towards the upstream RBR instead
    - Encoded as PIM RPF Vector, mLDP Recursive FEC, or BGP-MCAST RPF Address EC

# Overlay Signaling over a Region

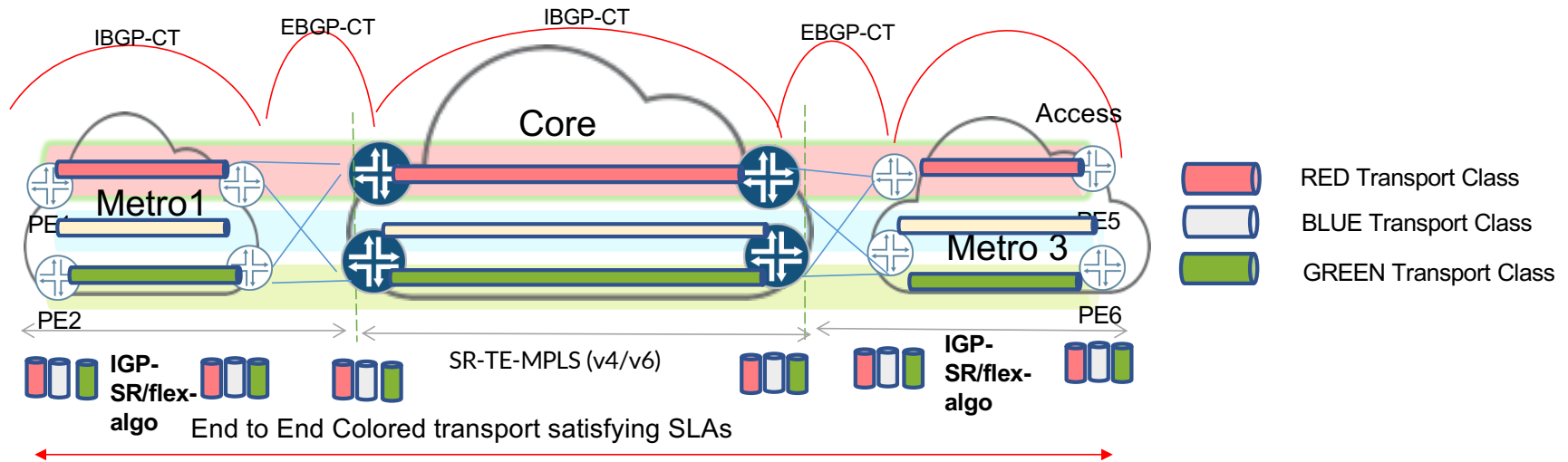


- Internal routers do not keep state for E2E trees
- RBR2 signals to RBR1 directly
  - BGP-MVPN [RFC6514]
  - mLDP over targeted sessions [RFC7060], BGP-MCAST
    - Actually no difference between overlay and inband signaling – it's just whether upstream/downstream nodes are directly connected or not
- RBR1 tunnels traffic to RBR2 – via Ingress Replication or P2MP

# Agenda

- SR principles and Multicast Options
- Controller Signaled P2MP
  - PCEP/BGP-signaled SR-P2MP
  - BGP-signaled mLDP
- BGP Signaled Multicast
  - Both IP multicast and P2MP
  - With or without controllers
- E2E Inter-region Multicast
- Multicast with Classful Transport

# BGP Classful Transport



- Underlay routes classified into Transport Classes (TCs)
  - Advertised via Classful Transport SAFI, with Transport Class Route Target
    - The Route Target specifies the TC and controls route propagation and import
- Service/overlay routes carry a mapping community
  - To map to the TC used to resolve Protocol NH

# Multicast with Classful Transport

- A multicast tree/tunnel may be:
  - an underlay one for a particular TC, or,
  - an overlay one using a particular TC
- Either way, the BGP-MCAST signaling may carry a mapping community for the TC, which affects:
  - path/tunnel selection between an upstream and its downstream nodes
  - upstream node selection for a downstream node

# Summary

- Various options for multicast in SR networks
  - Per SR principles or not
  - Per deployment considerations
- BGP-MCAST is the best none-BIER, non-traditional option
  - Single Session from controller to one of the BGP speakers
  - Great coverage and extensibility
  - Additional benefit for BGP-MCAST signaled mLDP
    - Simplifies transition from existing mLDP deployment
    - Flexibility and extensibility due to opaque nature of mLDP FEC
- E2E inter-region multicast
  - Different signaling methods in different regions
- BGP-MCAST signaling easily integrates with classful transport



# Thank you

February-2021

