Methods for efficient Multicast Delivery in MPLS networks

Praveen Muley

Email : praveen.muley@alcatel-lucent.com

Pradeep Jain

Email : pradeep.jain@alcatel-lucent.com

Agenda

1. Evolution of Multicast Technologies
2. Brief overview of multicast tunnels
3. Deployment Considerations for multicast traffic
   - Minimizing packet replication
   - S2L Path Vs. Global tree re-optimization (RSVP-P2MP)
   - Flexibility in mapping multicast channels to P2MP LSP
   - Resilience of multicast source and tree
4. Conclusion

1 Evolution of Multicast Technologies

Multicast using PIM

- Use of PIM in core
  - Maintains (S,G or per Channel) state in core .
  - Failure detection and convergence based on IGP.
    - Needs to rebuild the trees again in case of failure.
  - Lack of traffic Engineering capabilities.

- Choice of tunnel
  - No tunneling. Traditional multicast IP forwarding.

Multicast VPNs using PIM

- draft-rosen-vpn-mcast-XX introduced M-VPN
  - At least one multicast tree per M-VPN in the core. Provides no option to aggregate multiple M-VPN into a reduced amount of core multicast trees in order to reduce the amount of multicast state in the core of the network.
  - PE to PE protocol exchanges are only described using PIM. Does not allow the usage of BGP to line up the Core signaling between unicast VPNs and multicast VPNs
  - Inter-AS M-VPN deployments require a full mesh multicast core tree between all the PE(s) of all the AS(s) that are involved in the M-VPN. Does not allow for segmented inter-AS trees to provide more scalable inter-AS M-VPN deployments.
  - Failure detection and convergence based on IGP.
    - Needs to rebuild the trees again in case of failure.
  - Lack of traffic Engineering capabilities.

- Choice of tunnel
  - GRE is the only option described to encapsulate Multicast VPN traffic.
  - Provides no option to allow the usage of a multicast MPLS data-plane. As such MPLS unicast VPNs and multicast VPNs are using a different data-plane technologies.

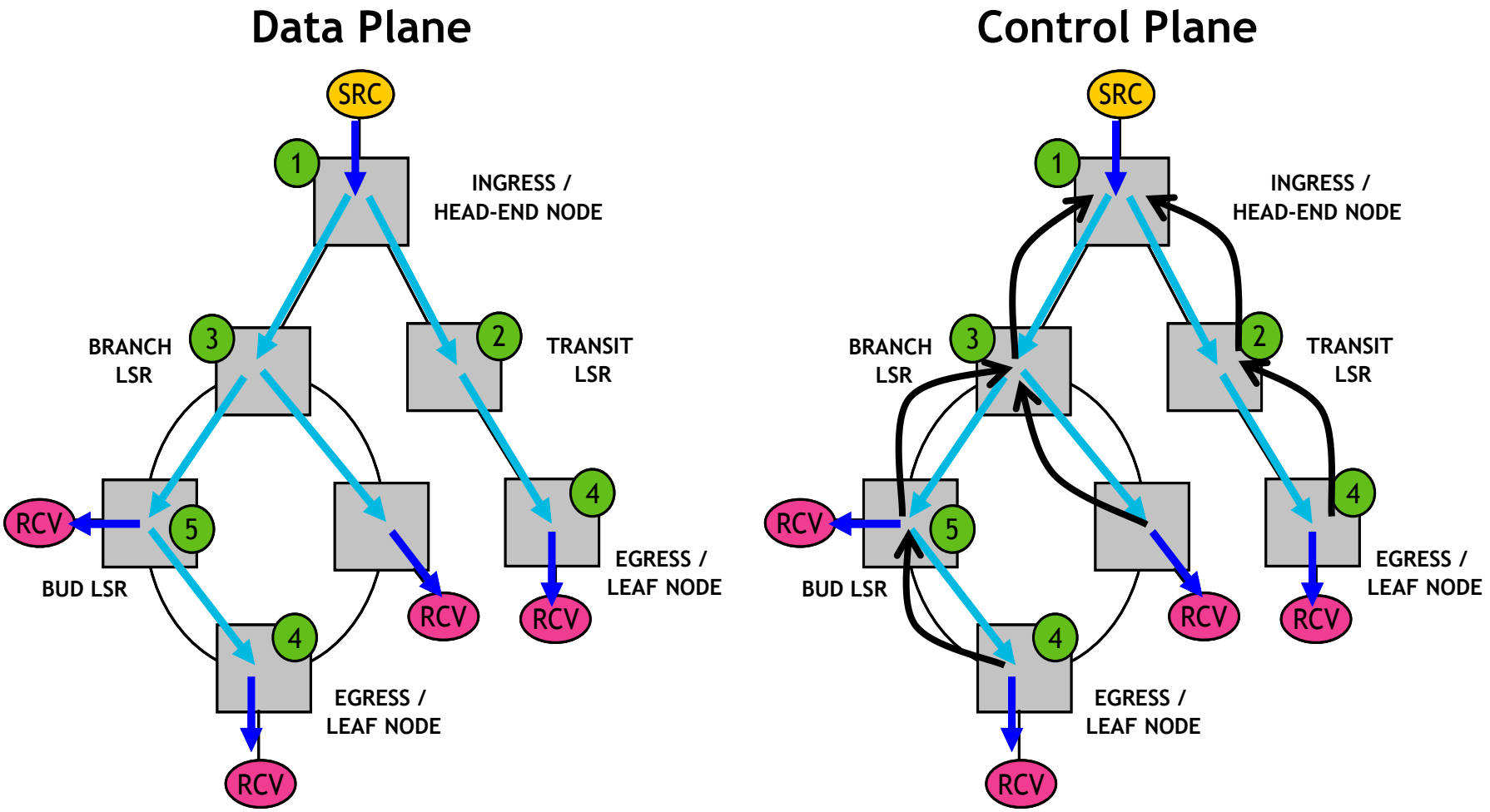
Multicast technology choices in MPLS

- Global context Vs. VPNs
  - Static mapping to tunnels (per Source or per interface).
  - Use of Multicast-VPN aka M-VPN (similar to 2547 VPNs) .
    - BGP based signaling
    - Aggregated
    - Non-Aggregated
    - decouples the procedures for exchanging routing information from the procedures for transmitting data traffic.
    - I-PMSI and S-PMSI can be 2 different tunneling technology.

- Choice of tunnel
  - M-LDP (multi-point LDP)
  - RSVP-P2MP (RSVP Point to multi-point)

2 Overview of Multicast Tunnels

LDP P2MP data plane vs. control plane views



LDP P2MP LSP

LDP P2MP LSP is modeled as a single merged tree in the control plane

LSP setup is initiated by the leaf node towards the root node. Control plane is an extension of LDP P2P signaling. The leaf node sends the LABEL map towards the root node and merges with the tree at a branch node LSR along the path towards the root node.

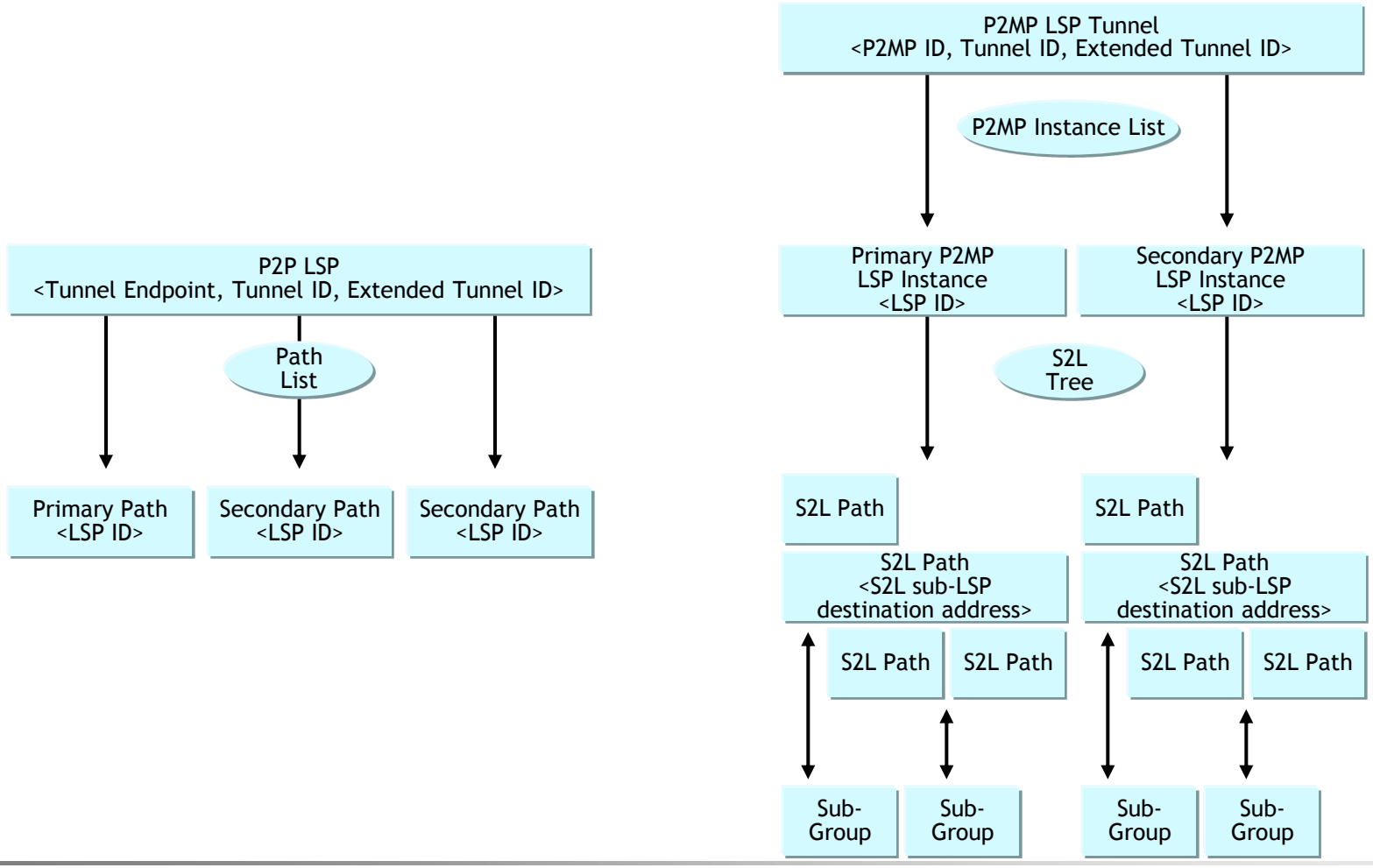
A P2MP LSP is identified by <Generic 32 bit Identifier>

- Each client application is assigned a unique ID for each P2MP tree on the root node. In a dynamic application like MVPN, head-end informs all leaf nodes (via BGP MVPN signaling) to initiate a P2MP LSP towards it, along with the allocated LDP P2MP ID. All static applications must configure the P2MP tree using LDP P2MP ID that must match on head-end and leaf node.
- Proposal at IETF allows the identification of each LDP P2MP LSP based on client application (Dynamic, Static, MVPN etc.)

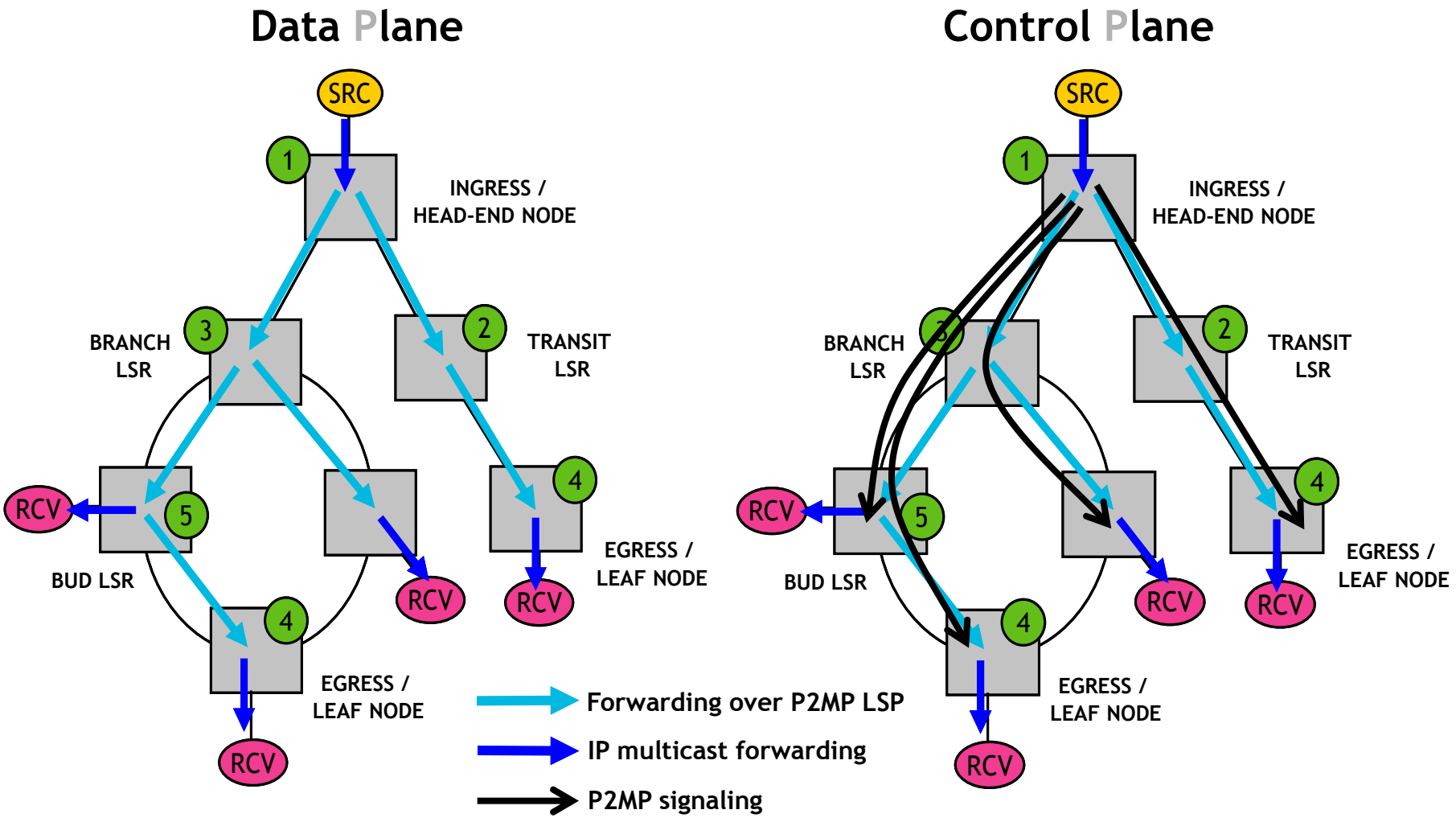
LDP P2MP (continued)

- Benefits
  - PIM free core.
  - No per (S,G) state maintenance in core of network.
  - MPLS data forwarding plane.
- Open issues
  - Failure detection and convergence based on IGP.
    - Needs to rebuild the trees again in case of failure.
  - Lack of traffic Engineering capabilities.
    - No Fast Reroute capabilities (FRR).
    - FRR work in progress. Loop Free Alternates (LFA) not yet deployed.
    - No constraints based paths.

RSVP P2MP LSP versus P2P LSP models



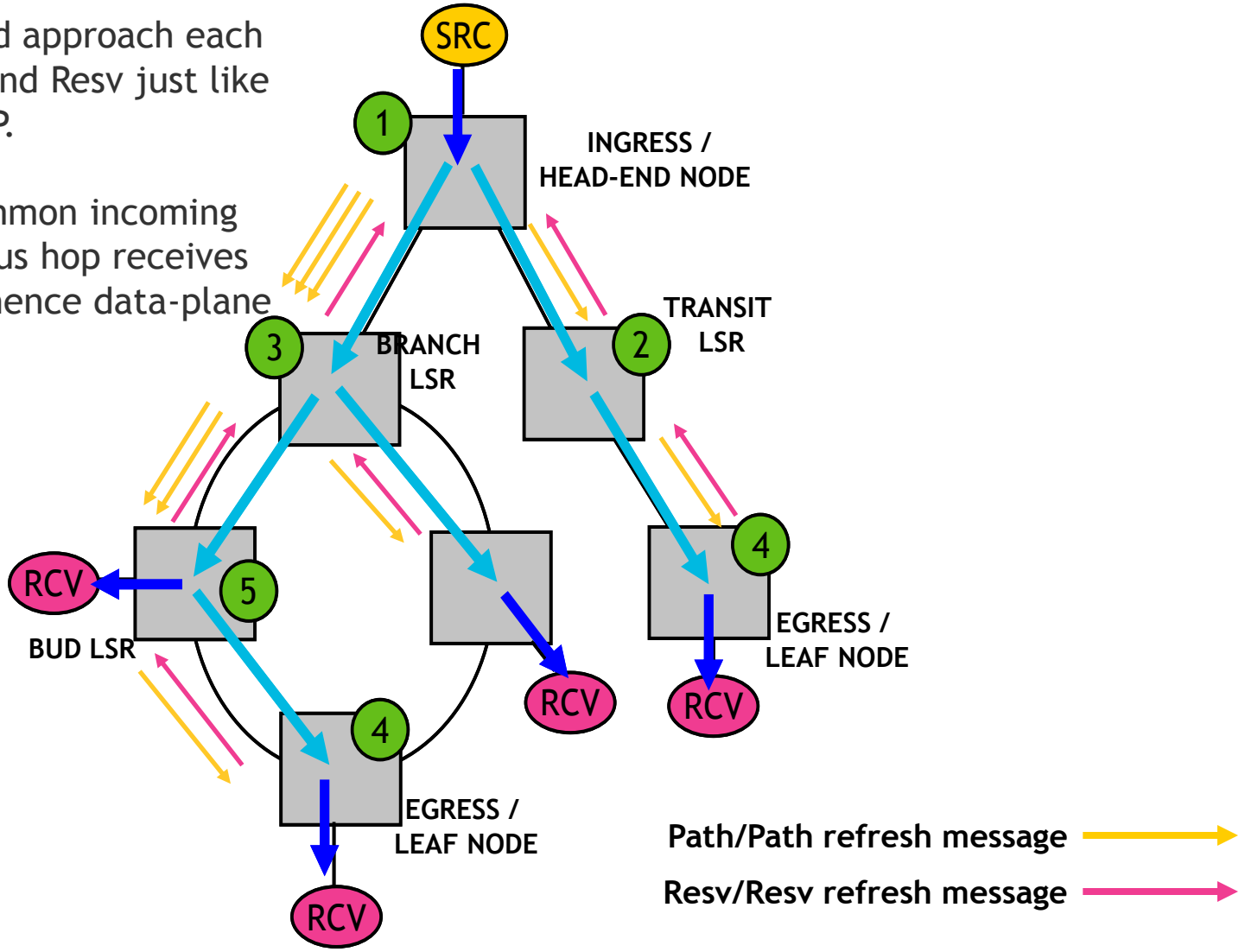
P2MP RSVP Data plane vs. Control plane view



RSVP Path and Resv state refresh in the de-aggregated method

- In de-aggregated approach each S2L signals Path and Resv just like point-to-point LSP.

- S2Ls having common incoming interface/ previous hop receives same label. And hence data-plane merges.



RSVP P2MP

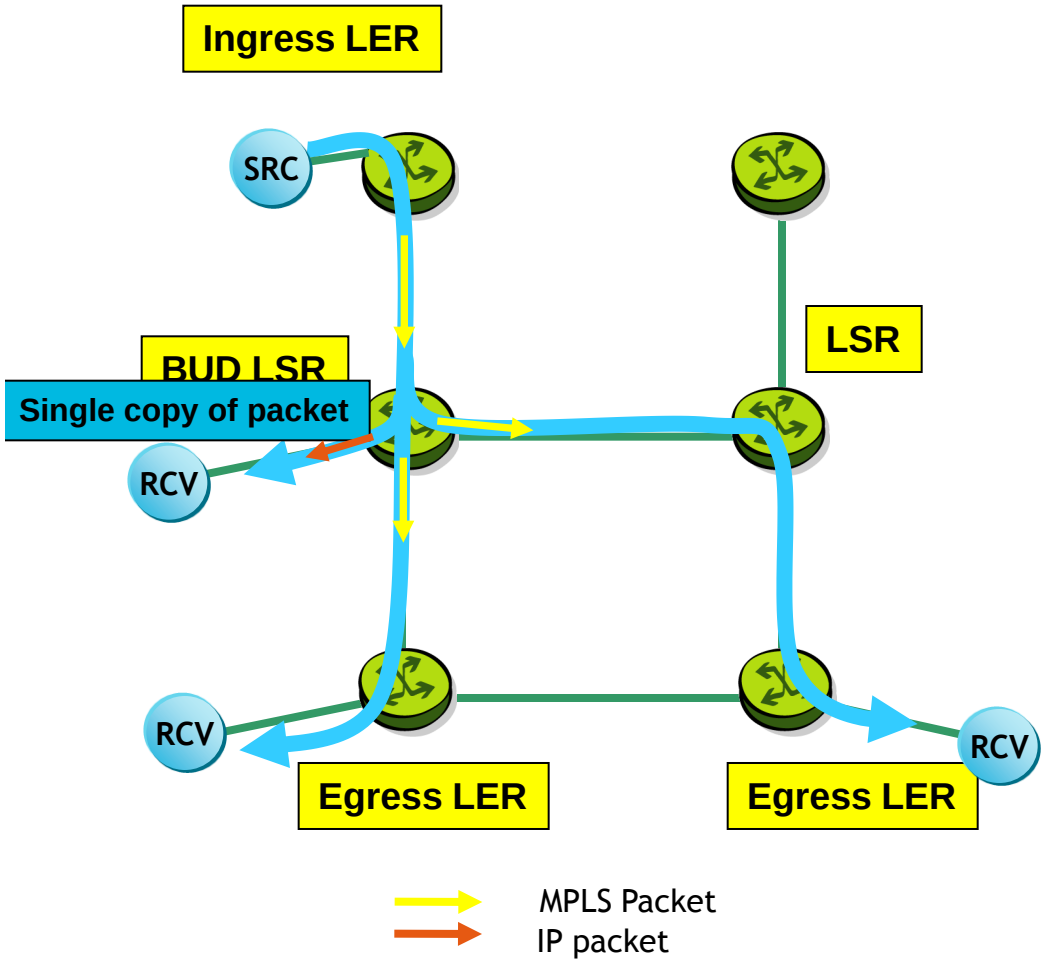
- Benefits
  - PIM free core.
  - No per (S,G) state maintenance in core of network.
  - MPLS data forwarding plane.
  - Failure detection and convergence sub 50 msecs.
    - Using FRR (Facility byPass preferred method).
  - Support of traffic Engineering capabilities.
    - Constraints based paths (Strict/loose paths)
    - Constraints such as admin groups/ SRLG.

3 Deployment considerations for Multicast traffic

Minimize Packet Replication

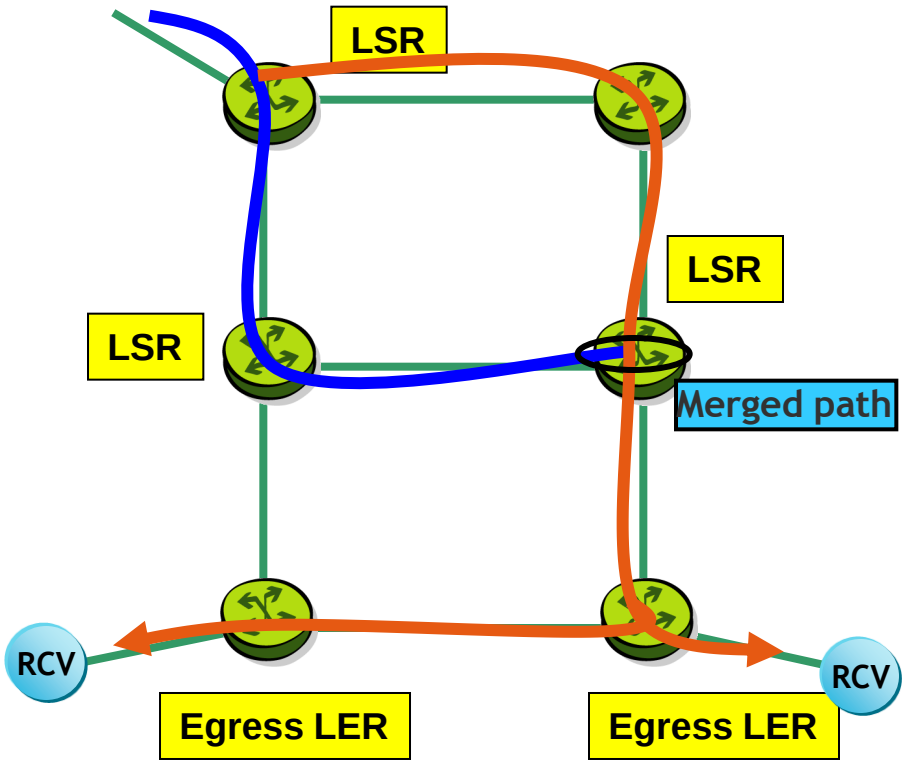
Minimizing Packet Replication at BUD LSR

- BUD LSR performs both MPLS and IP packet replication
- Use of two different labels from upstream LSR to BUD LSR causes double packet replication
- BUD LSR must perform simultaneous pop and swap operation on the packet in the data-path to avoid double bandwidth usage



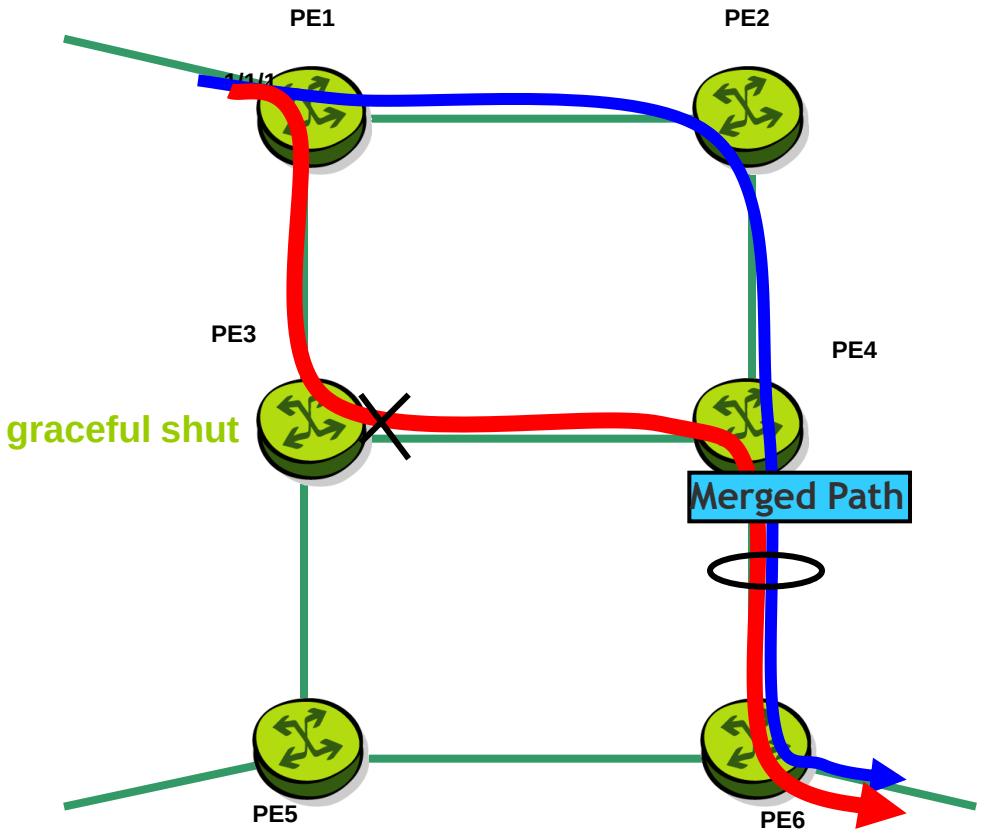
Minimizing Packet Replication due to Remerge – Case 1

- paths of two different S2Ls of the same P2MP LSP instance have ILMs on different incoming ports but go out on the same NHLFE.
- Blue ILM must be blocked from forwarding in the data-path at re-merge node



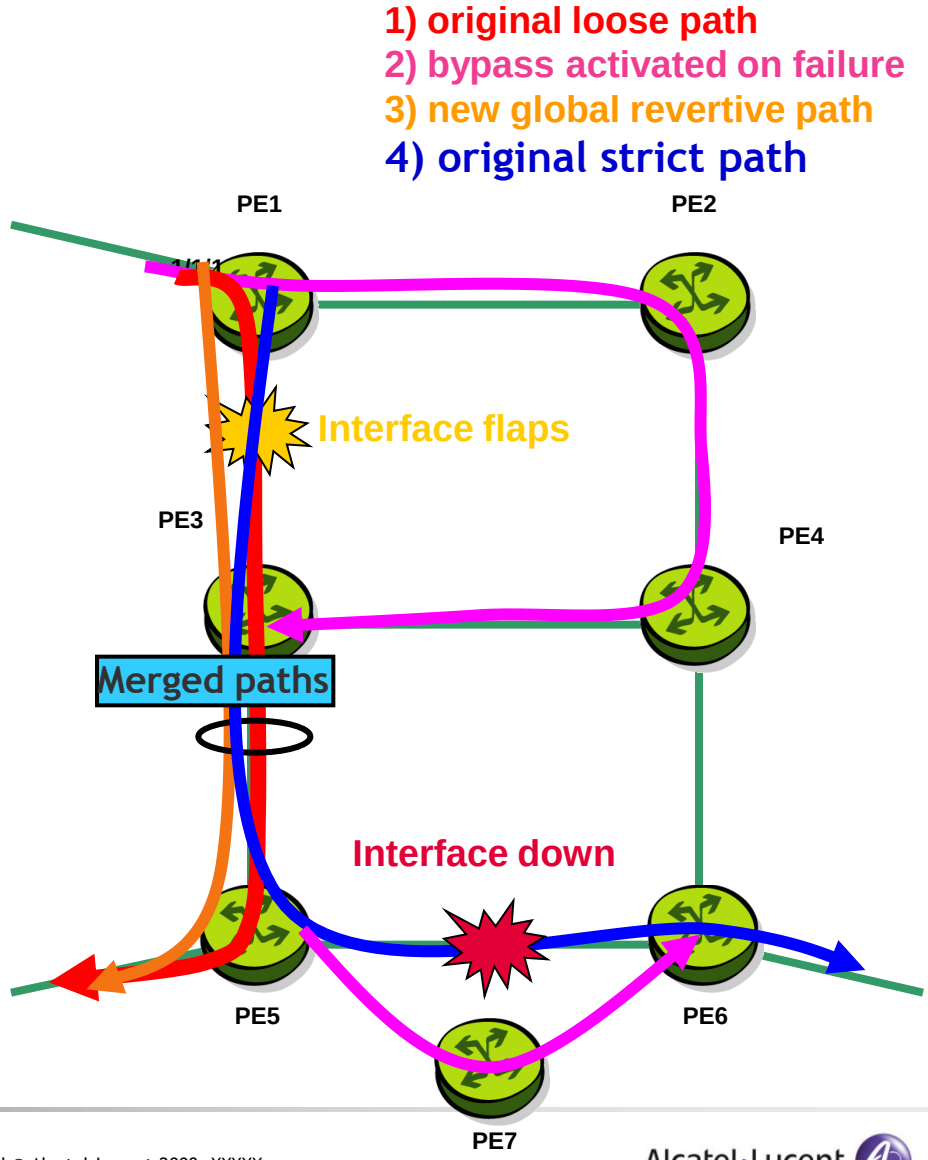
Minimizing Packet Replication due to Remerge – Case 2

- Two paths of same S2L have ILMs on different ports but go out on the same NHLFE
- Occurs during MBB of S2L path, e.g., graceful shutdown of TE link PE3-PE4
- Blue ILM must be blocked from forwarding in the data-path at re-merge node until original RED path torn down



Minimizing Packet Replication due to Remerge – Case 3

- New Global revertive path of an S2L arrives on the same incoming interface as the original path and must merge with active bypass path
- PE3 must signal PE1 a different label from the original S2L path or it cannot re-merge and duplication would occur all the way to egress PE
- If blue S2L has strict path and Global revertive MBB fails due to a double failure
- Blue S2L path forwards on two active bypass LSPs
- If PE3 does not signal PE1 a different label for orange S2L path, duplication of traffic from PE1 to PE5 is permanent until link PE5-PE6 comes back up



S2L Path Vs. Global tree optimization

Reroute choices

- Individual S2L re-optimization required in case of MBB
  - MBB due to FRR backup path active
    - Bandwidth usage is duplicated on a P2P bypass LSP and thus can't wait for global tree to re-signal.
  - MBB due to other operation such as TE graceful shutdown.
- Global tree re-optimization needed
  - On a regular basis to optimize entire multicast tree
- Switch to a backup P2MP LSP tree during failure
  - Useful when a failure is at a node closer to the root of the tree.
  - When multiple leaf nodes are impacted by the failure
  - Requires re-signaling and moving multicast streams to new set of S2L paths.
    - new tree must have at least as much coverage as the existing one before switching.
    - Node must maintain state for both trees until new tree has same coverage.
    - Not recommended under single or a few S2L path failures.

Flexibility in mapping multicast channels to P2MP LSPs

Mapping multicast channels to P2MP LSP

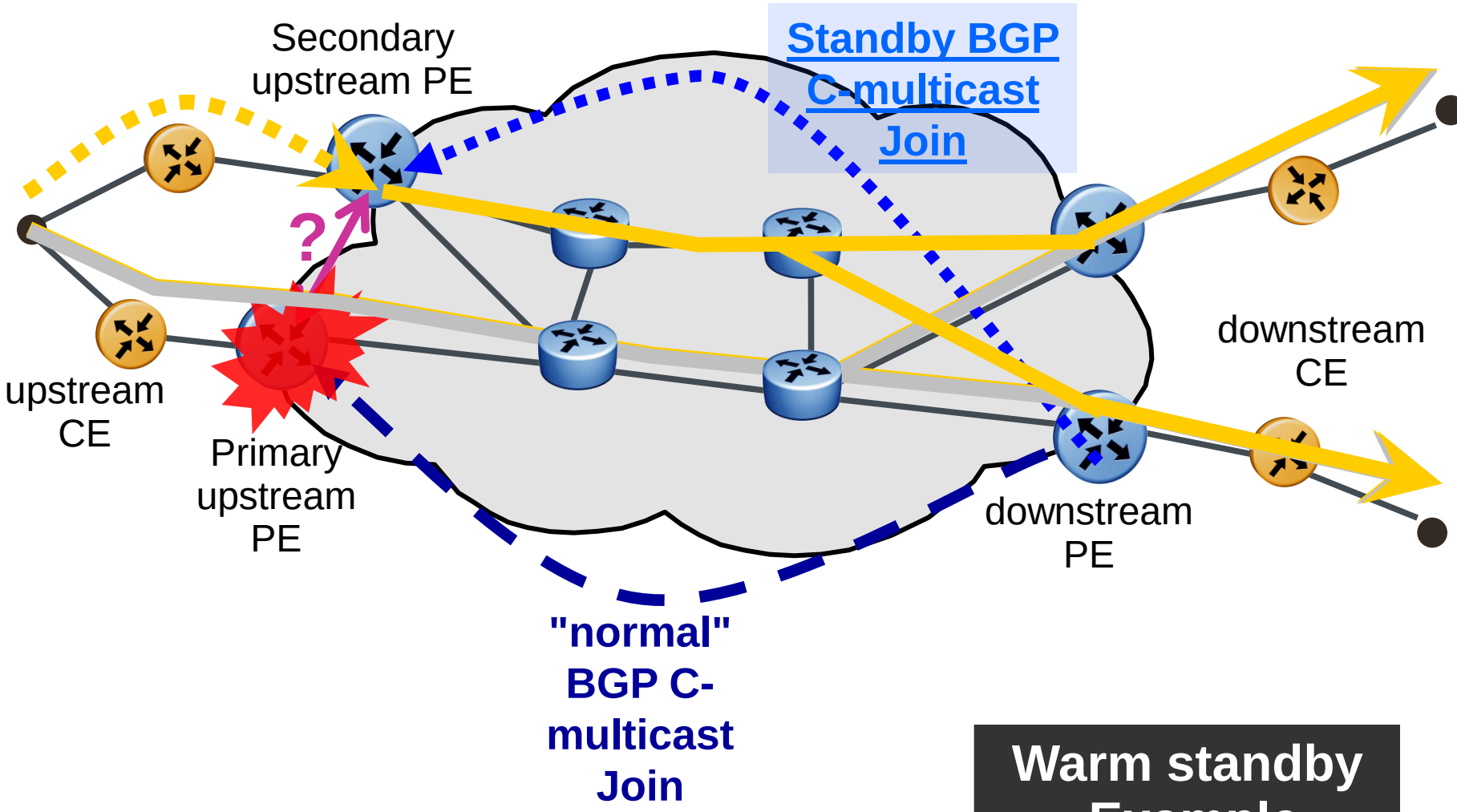
- Flexible mapping of <S,G> from any IP interface to any P2MP LSP
  - Useful for specifying different bandwidth for video channels (e.g., SD versus HD)
  - Provides means of avoiding fate sharing among types of video channels.
  - Allows egress LER of P2MP LSP to also operate as PIM Designated Router for downstream receivers

Resilience of Multicast source and Tree

Different protection levels

- Cold standby
  - Backup PE waits for the failure before joining toward the CE
- Warm standby
  - Backup PE is ready to send traffic when failure occurs ("pre-joined" toward the CE)
- Hot standby
  - Backup PE sends the traffic before the failure occurs
  - Downstream PE switch to backup tunnel based on VPN unicast routes
- Hot leaf standby
  - Backup PE sends the traffic before the failure occurs
  - Downstream PE switch to backup tunnel based on tunnel status

Use of M-VPN in conjunction with RSVP-P2MP – Warm standby



Use of M-VPN in conjunction with RSVP-P2MP – Warm standby

- **Standby BGP C-Multicast route**
- Idea: prepare the backup PE so that it is prepared for a failure of the primary PE
- How ?
  - Besides advertising a normal (C-S,C-G) C-multicast Tree Join route to the nominal upstream PE, downstream PEs advertise a Standby C-multicast Tree Join route to the backup upstream PE
  - The backup upstream PE prepares for a possible failure (e.g. by joining the source)
  - The backup upstream PE monitors the reachability of C-S through the nominal PE
  - On failure, traffic is forwarded by backup PE
- Failure detection can be done, for instance
  - based on P2MP OAM.
  - Based on unicast VPN reachability to C-S
- **Key : Avoid signaling at failure time**

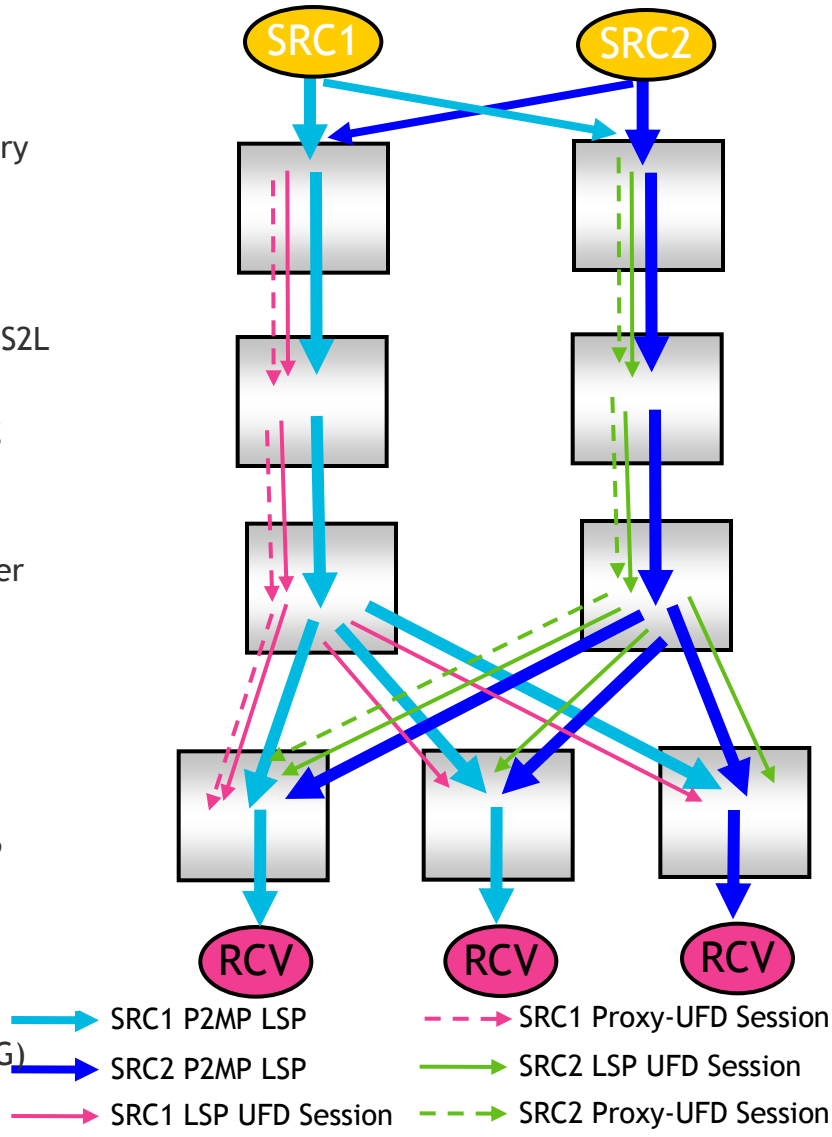
Use of M-VPN in conjunction with RSVP-P2MP – HOT Leaf Standby

- **UMH selection based on tunnel status for MVPN fast-failover**
- Reminder:
  - "UMH Selection" designate how a downstream PE determines the PE from which it will receive a said multicast flow
  - "UMH Selection" is done based on VPN unicast routing information
  - (similar to PIM RPF)
- Idea:
  - Make "UMH Selection" fail-over to a backup PE as soon as the P-tunnel is down, without waiting for unicast VPN convergence
  - Different possible ways to detecting that a tunnel is down:
    - P2MP OAM (Multipoint BFD)
    - Traffic counters
    - P-Tunnel signaling (RSVP-TE PathTear)
    - IGP tunnel root tracking
    - …
- **Key: avoid waiting for unicast convergence.**

Example of Resilient Solution

- Primary and backup source forward over two path disjoint P2MP LSPs terminating on egress LER
  - Egress LER receives duplicate multicast streams on primary and secondary P2MP LSP S2Ls
- Ingress LER establishes Unidirectional Forwarding Detection (UFD) session to egress LER
  - one for the LSP to detect node failure – link failure is via S2L sub-LSP FRR
  - optionally one proxy-UFD for each multicast source being tracked by the ingress LER
    - Source tracking via routing table, BFD, etc…
- If egress LER misses a number of successive UFD packets over the primary LSP S2L,
  - It declares the primary P2MP LSP S2L and/or the specific multicast source as down.
  - Moves the receiving of the affected <S,G> records to the secondary LSP S2L.
  - Reverts back if a failure is detected on the secondary LSP S2L.

  Egress also moves record of specific source
  - ufd message with explicit source down message
  - Stream quality monitoring using differential rates per (S,G) from primary and secondary S2L



4 Conclusion

Concluding Remarks

- Multicast using RSVP P2MP LSP and M-VPN offers operators a number of distinct properties
  - simplifies configuration by operating a PIM-free core
  - Allows a fine grained design of the multicast trees
  - Placement of branching points
  - Ability to add constraints to tree calculation.
  - Utilize the bandwidth in core optimally by re-optimizing the LSP periodically.
  - Provides FRR protection of paths.
  - Dynamic establishment of I and S-PMSI based on auto-discovery based on M-VPN provisioning.
  - A combination of aggregated I and S-PMSI for multiple M-VPNs can be most optimal solution
- Operators should take into account the above mentioned deployment considerations while deploying multicast.

www.alcatel-lucent.com