Multicast in MPLS/VPLS Networks

An IP/MPLS Forum Sponsored Tutorial

Dr. Yakov Rekhter
IP/MPLS Forum Ambassador
Juniper Fellow
Juniper Networks

Multicast in MPLS/VPLS Networks Tutorial Agenda

- Introduction to Multicast with MPLS
- Multicast in BGP/MPLS VPNs (2547 VPNs)
- Multicast in VPLS Networks

Multicast in MPLS/VPLS Networks Tutorial Contributors

- Matthew Bocci – Alcatel-Lucent
- Rao Cherukuri – Juniper Networks
- Arnold Jansen – Alcatel-Lucent
- Kireeti Kompella – Juniper Networks
- Yakov Rekhter – Juniper Networks

Introduction to the IP/MPLS Forum

- IP/MPLS Forum is an international, industry-wide, non-profit association of service providers, equipment vendors, testing centers and enterprise users
  - Created with the name change of the MFA Forum (Oct 2007) to reflect renewed focus on driving global industry adoption of IP/MPLS solutions in the market, by focusing on standards initiatives for IP/MPLS such as inter carrier interconnect (ICI), mobile wireless backhaul, and security.
- Objectives: Unify service providers, suppliers and end users on common vision of IP/MPLS based solutions

Awareness

- Promote global awareness of the benefits of IP/MPLS
- Empower the telecom industry to migrate from legacy technologies to IP/MPLS-based next generation networking

Migration

- Guide the telecom end user to make the leap from legacy technologies to IP/MPLS-based services

Systems-Level Solutions

- Drive implementation of standards for IP/MPLS based solutions
- Validate implementations and advance interoperability of standardized IP/MPLS based solutions

- Deliverables: Technical Specifications, Test Plans, Technical Tutorials, Collateral

Introduction to the IP/MPLS Forum

- **Current Work Items**
  - Framework and Reference Architecture for MPLS in Mobile Backhaul Networks
  - MPLS Inter-Carrier Interconnect
  - Packet Based GMPLS Client to Network Interconnect
  - Generic Connection Admission Control (GCAC) Requirements for IP/MPLS Networks
  - Layer 2 VPNs using BGP for Auto-discovery & Signaling (BGP L2 VPN)
  - MPLS Over Aggregated Interface
  - Voice Trunking format over MPLS
  - TDM Transport over MPLS using AAL1

  The Forum is also planning several industry-driven future Work Items.
- **Service Provider Council**
- **Public Interoperability Events**
- **Technical Tutorials -** to broaden the understanding of the technology and benefits of the solutions
- Next meeting: June 24-26, Vancouver, Canada
- Please join us!
  - **To join the Forum contact Alysia Johnson, Executive Director**

    E-Mail: ajohnson@ipmplsforum.org

    Phone: 510 492-4057

Technical Tutorials

Introduction to MPLS	**½ and full day**
MPLS L2/L3 VPNs	**½ day**
MPLS VPN Security	**½ day**
Traffic Engineering	**½ day**
GMPLS	**½ day**
Migrating Legacy Services to MPLS	**½ day**
MPLS OAM	**½ day**
Voice over MPLS	**½ day**
Multi-service Interworking over MPLS	**½ day**
Multicast in MPLS/VPLS Networks	**½ day**
IP/MPLS in the Mobile RAN	**½ day**
MPLS Inter-Carrier Interconnect	**½ day**

New tutorials based upon demand

Section 1

Introduction to Multicast with MPLS

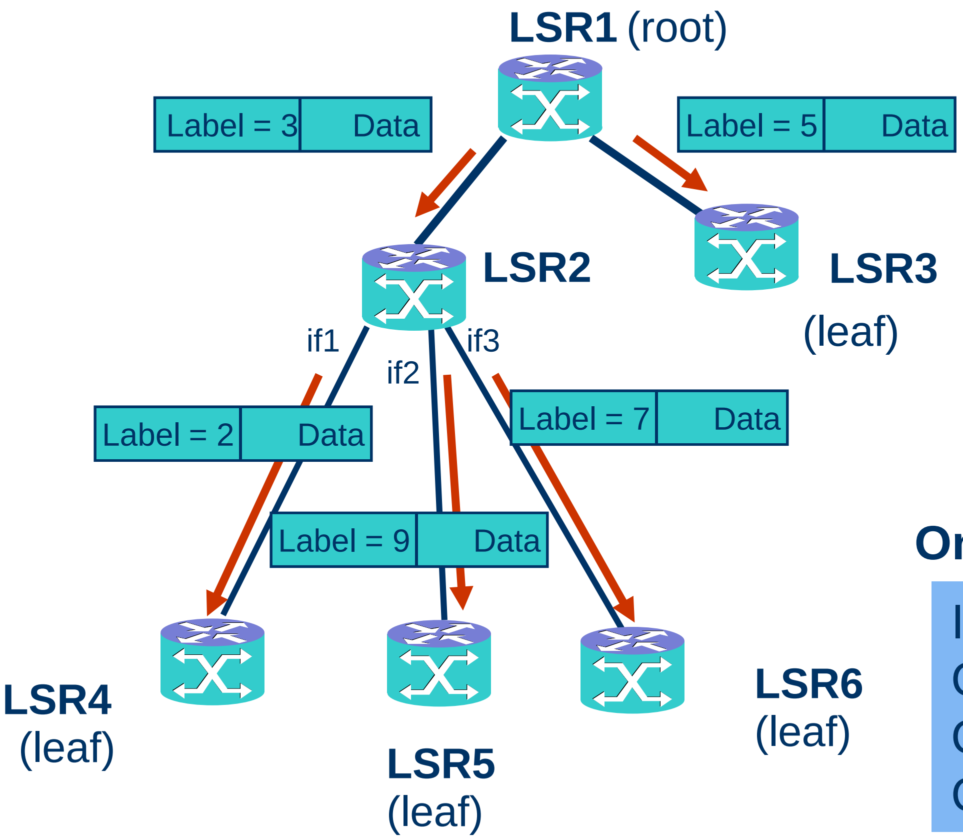
Agenda

- P2MP LSP
  - Setting up P2MP LSP with LDP
  - Setting up P2MP LSP with RSVP-TE
- Upstream labels with P2MP LSPs
  - P2MP LSP over LAN
  - Distribution of upstream labels
  - P2MP LSP Hierarchy
- Usage of P2MP LSPs

Point-to-multipoint Label Switched Path (aka P2MP LSP)

- P2MP LSP is the fundamental construct to support multicast with MPLS
  - Just like point-to-point and multipoint-to-point LSPs are fundamental constructs to support unicast with MPLS
- Each P2MP LSP has a single root, but multiple leaves
  - Although "multiple" may be just one
- All the leaves have to know the identity (IP address) of the root
- Root may, or may not know the identity (IP addresses) of all the leaves
- "Next Hop Label Forwarding Entry" (NHLFE) maintained by each node along a P2MP LSP:
  - Specifies a set of next hops
    - Although the set contains only a single member
  - Semantics: the packet is to be replicated, and a copy of the packet sent to each of the specified next hops
    - Each copy carries its own distinct label
- For a given P2MP LSP its root and all the leaves must agree on the FEC that corresponds to the P2MP LSP

P2MP LSPs – example (data plane)



On LSR2:

Incoming label 3
Outgoing label 2, outgoing interface if1
Outgoing label 9, outgoing interface if2
Outgoing label 7, outgoing interface if3

Constructing P2MP LSP with LDP

Step 1: All leaf nodes find the identity of the LSP

- By means outside of LDP
  - By an application that uses P2MP LSP with LDP
- The identity includes the address of the root node

Step 2: Each leaf node initiates P2MP LSP setup by sending LDP Label Mapping message towards the root

- Sent only to the (upstream) LSR that is on the path to the root
  - Using unicast route towards the root
- Label Mapping message carries the identity of the LSP
  - Encoded as P2MP FEC (Forwarding Equivalence Class) element – see next slide

Step 3: Each intermediate node along the path from a leaf to the root propagates LDP Label Mapping towards the root

- Sent only to the (upstream) LSR that is on the path to the root
  - Using unicast route towards the root

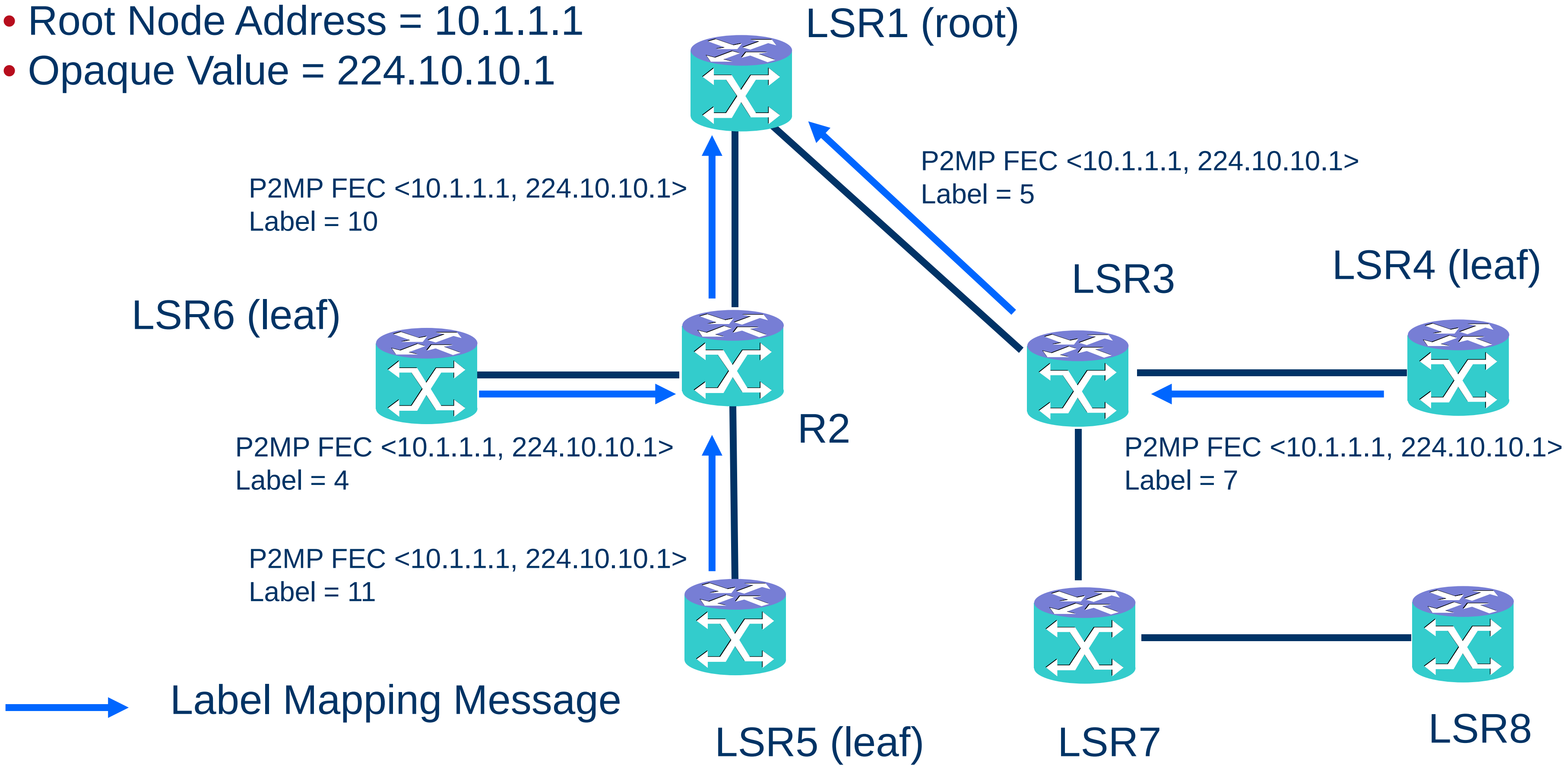
Constructing P2MP LSP with LDP: P2MP FEC Element

- FEC – Forwarding Equivalence Class
- All leaf nodes of a given P2MP LSP have to use the same P2MP FEC Element for the LSP
  - Accomplished by means outside of LDP
    - By an application that uses P2MP LSP with LDP
- P2MP FEC Element carries:
  - Root Node Address
  - Opaque Value
    - Has to be unique at least within the context of the root node
- P2MP FEC Element forms a globally unique identifier for a P2MP LSP
- P2MP FEC Element need not directly specify which packets are mapped into the P2MP LSP
  - Indirect mapping (e.g., via auto-discovery in Multicast VPN)
- P2MP FEC Element is carried in LDP Label Mapping and Label Withdraw messages
  - Creates label binding for a given P2MP FEC Element (for a given P2MP LSP)

LDP P2MP signaling: example

P2MP FEC Element :

- Root Node Address = 10.1.1.1
- Opaque Value = 224.10.10.1



LSR1 (root)

P2MP FEC <10.1.1.1, 224.10.10.1>
Label = 10

P2MP FEC <10.1.1.1, 224.10.10.1>
Label = 5

LSR3

LSR4 (leaf)

LSR6 (leaf)

R2

P2MP FEC <10.1.1.1, 224.10.10.1>
Label = 4

P2MP FEC <10.1.1.1, 224.10.10.1>
Label = 7

P2MP FEC <10.1.1.1, 224.10.10.1>
Label = 11

Label Mapping Message

LSR5 (leaf)

LSR7

LSR8

Constructing P2MP LSP with RSVP-TE

Step 1: Root finds IP addresses of all the leaf nodes

- By means outside of RSVP-TE
  - By an application that uses P2MP LSP with RSVP-TE

Step 2: Root computes paths from itself to all the leaf nodes

- Either constrained shortest path first (CSPF), or (approximate) constrained minimum cost tree (aka Steiner tree)
  - Supports the same Traffic Engineering Constraints as point-to-point LSP with RSVP-TE
- May also be precomputed by off-line tools

Step 3: Root uses RSVP-TE to set up P2MP LSP

- Establishes label forwarding state
- May involve resource reservations
- See the following slides for more details on this…

RSVP-TE signaling for P2MP LSP

- RSVP-TE for P2MP LSPs as an extension to RSVP-TE for point-to-point LSPs
  - Minimize changes to the existing RSVP-TE
- Building blocks:
  - P2MP Tunnel
  - P2MP LSP
    - One or more per P2MP Tunnel
  - S2L sub-LSP
    - One or more per P2MP LSP
- Path, Resv messages

RSVP-TE P2MP building blocks (1): P2MP Tunnel

- P2MP Tunnel is identified by the P2MP SESSION object which includes:
  - Extended Tunnel ID: IPv4/IPv6 address of the root node of the tunnel
  - P2MP ID: Logical 32 bit identifier of the P2MP tunnel
    - Unique within the context of the root node
  - Tunnel ID: 16 bit identifier
    - Unique within the context of the root node
  - <Extended Tunnel ID, P2MP ID, Tunnel ID> triplet forms a globally unique identifier for a P2MP Tunnel
- P2MP Tunnel comprises of one or more P2MP LSPs
  - All such P2MP LSPs have the same root node and the same set of leaf nodes

RSVP-TE P2MP building blocks (2): P2MP LSP

- P2MP LSP is a specific instance of a P2MP Tunnel
- P2MP LSP is identified by a combination of P2MP SESSION and P2MP SENDER_TEMPLATE objects
- P2MP SENDER_TEMPLATE object includes:
  - IPv4/IPv6 Tunnel Sender Address : address of the root node of the LSP
    - The same as Extended Tunnel ID in the P2MP SESSION object
  - LSP ID - identifies a specific instance of a P2MP Tunnel
- P2MP LSP comprises of multiple S2L sub-LSPs

RSVP-TE P2MP building blocks (3): S2L Sub-LSP

- P2MP LSP comprises of multiple S2L sub-LSPs
  - One per each leaf node
- S2L sub-LSP is an LSP from the root node to a particular leaf node
- S2L sub-LSP is represented by:
  - S2L_SUB_LSP object:
    - Identifies a particular S2L sub-LSP
    - Leaf node (unicast) destination address
  - ERO or sub-ERO object:
    - Represents the explicit route from the root to the leaf
    - May be compressed, if multiple S2L sub-LSPs are carried in the same Path message

RSVP-TE P2MP signaling: Path and Resv messages

- One P2MP LSP can be signaled using one or more Path/Resv message
- Each such Path/Resv message can signal one or more S2L sub-LSPs
- Limiting cases:
  - A separate Path/Resv message for each S2L sub-LSP of a given P2MP LSP
  - A single Path/Resv message for all S2L sub-LSPs of a given P2MP LSP

RSVP-TE P2MP signaling: example

P2MP Tunnel:
Root: R1
Leaves: {R4, R5, R6}
ID: 224.10.10.1

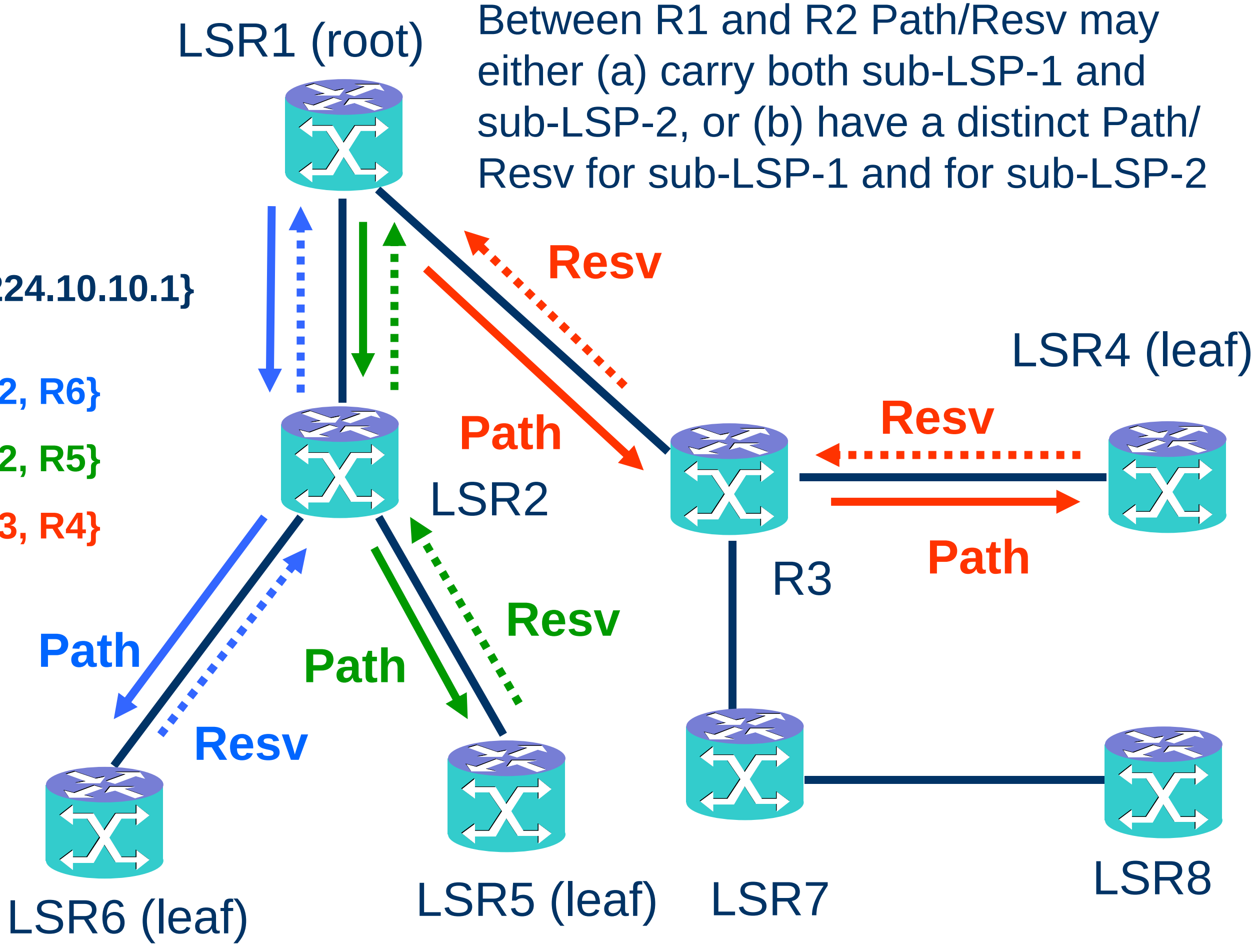
P2MP LSP:
P2MP Tunnel {Root R1, ID 224.10.10.1}
LSP-ID: 23

S2L sub-LSP-1: Leaf R6, ERO {R2, R6}

S2L sub-LSP-2: Leaf R5, ERO {R2, R5}

S2L sub-LSP-3: Leaf R4, ERO {R3, R4}

Between R1 and R2 Path/Resv may either (a) carry both sub-LSP-1 and sub-LSP-2, or (b) have a distinct Path/Resv for sub-LSP-1 and for sub-LSP-2



Agenda

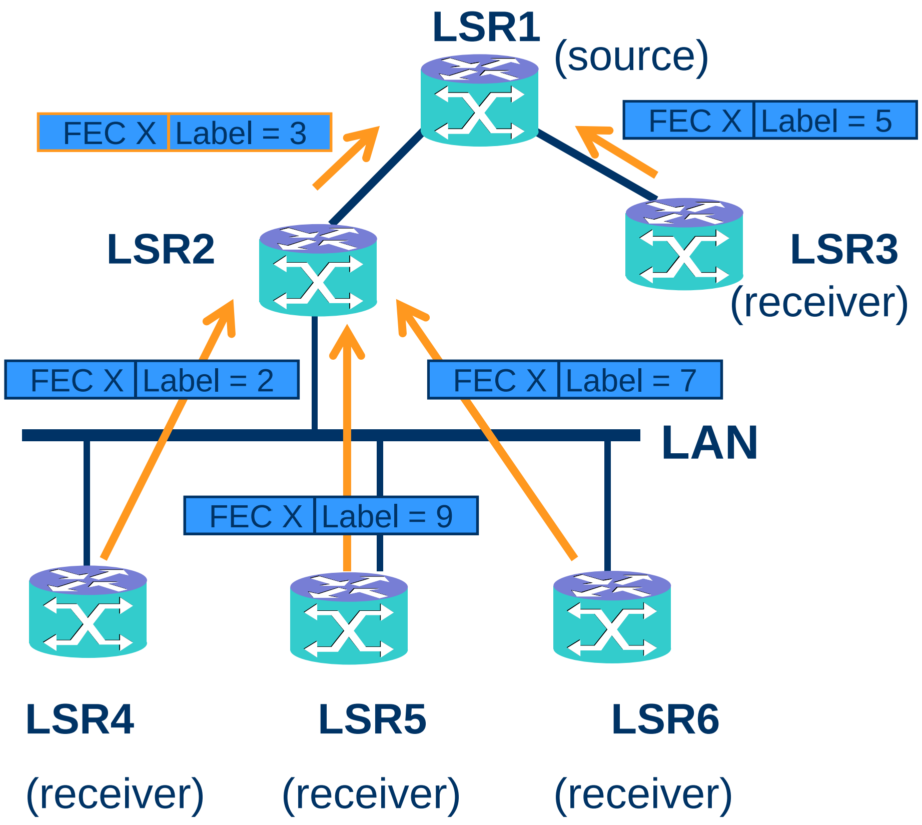
- P2MP LSP
  - Setting up P2MP LSP with LDP
  - Setting up P2MP LSP with RSVP-TE
- ⇨ Upstream labels with P2MP LSPs
  - P2MP LSP over LAN
  - Distribution of upstream labels
  - P2MP LSP Hierarchy
- Usage of P2MP LSPs

Upstream Labels – why ?

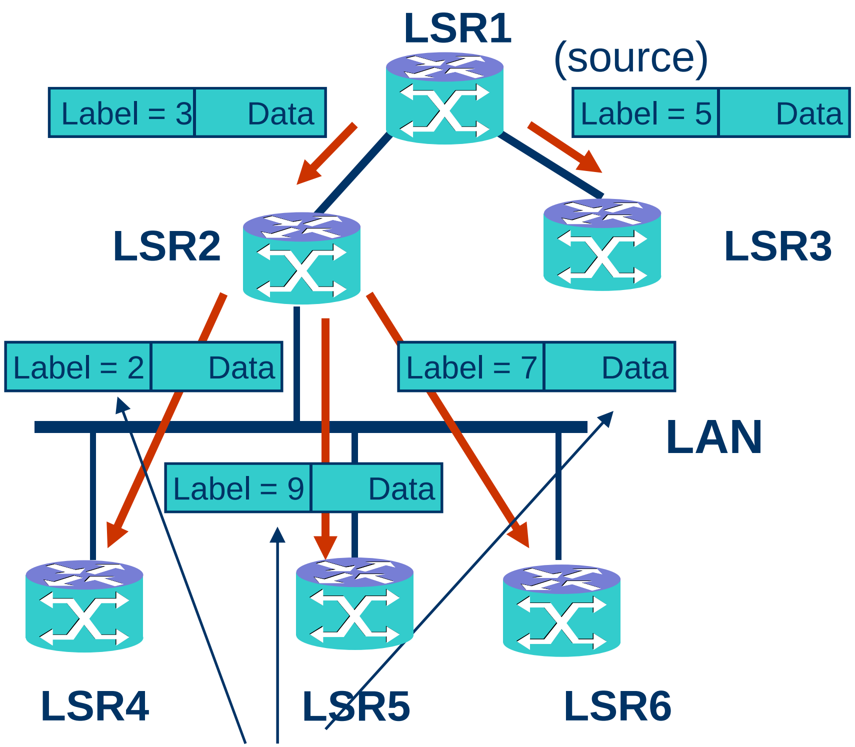
- P2MP LSP over LAN
  - How to send only a single copy of a packet when a P2MP LSP traverses a LAN ?
- P2MP LSP Hierarchy
  - How to nest multiple P2MP LSPs inside a single P2MP LSP ?

P2MP LSP over LAN with downstream-assigned label – what is the problem ?

Control Plane (Label Distribution)



Data Plane (Data Forwarding)



**3 copies of the same packet !!!
(each with its own label)**

P2MP LSP over LAN with downstream-assigned labels

- P2MP LSP with downstream-assigned labels assumes <u>each downstream router assigns its own label for the LSP</u>
  - The upstream router uses this label when sending the data to the downstream router
- On a point-to-point links: there is only one downstream router
  - The upstream router sends a single copy of a packet
- On a LAN: there may be <u>several downstream routers</u>, <u>each assigning its own label for the LSP</u>
  - The upstream router sends multiple copies of a packet over the same LAN, one per downstream router
  - A copy sent to a particular downstream router carries the label assigned by that router
- Downstream-assigned labels result in a suboptimal behavior for P2MP LSPs traversing LANs
  - Suboptimal bandwidth use – multiple copies of the same packet traversing LAN
  - Suboptimal router use – replication to send multiple copies of the same packet

P2MP LSP over LAN – upstream-assigned label

To send only a single copy of a packet on a LAN requires the upstream router and all of the downstream routers to “agree” on a common label that should be used for a given P2MP LSP

Question 1: Who is to assign the label ?

Answer 1: Upstream LSR

Question 2: How to distribute the label ?

Answer 2: Using extensions to LDP and RSVP-TE

Question 3: How to make this label unambiguous ?

Answer 3: Using context specific label space

Upstream-assigned vs. downstream-assigned labels

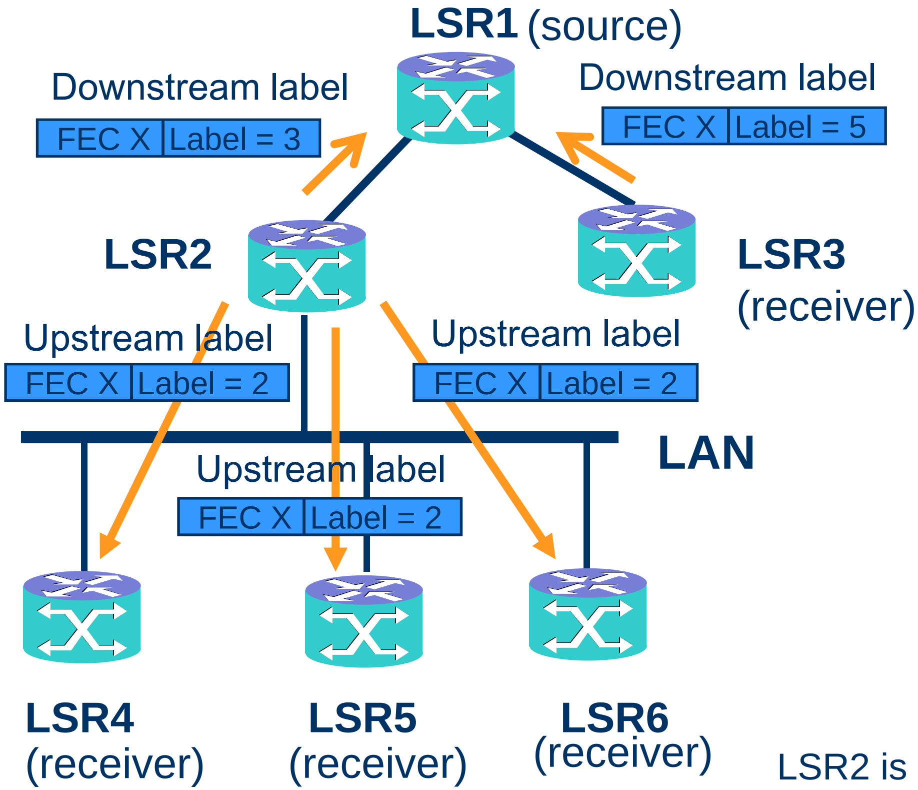
Before Router-upstream can send an MPLS packet to Router-downstream with label L at the top of the label stack, Router-upstream and Router-downstream must agree on the Forwarding Equivalence Class (FEC) which is bound to L.

- **Downstream-assigned labels: if the L to FEC binding is first made by Router-downstream, and then advertised by Router-downstream to Router-upstream, it is an "*downstream-assigned*" binding/label**
  - **Advertisement is via any label distribution protocol (e.g., LDP, RSVP-TE)**

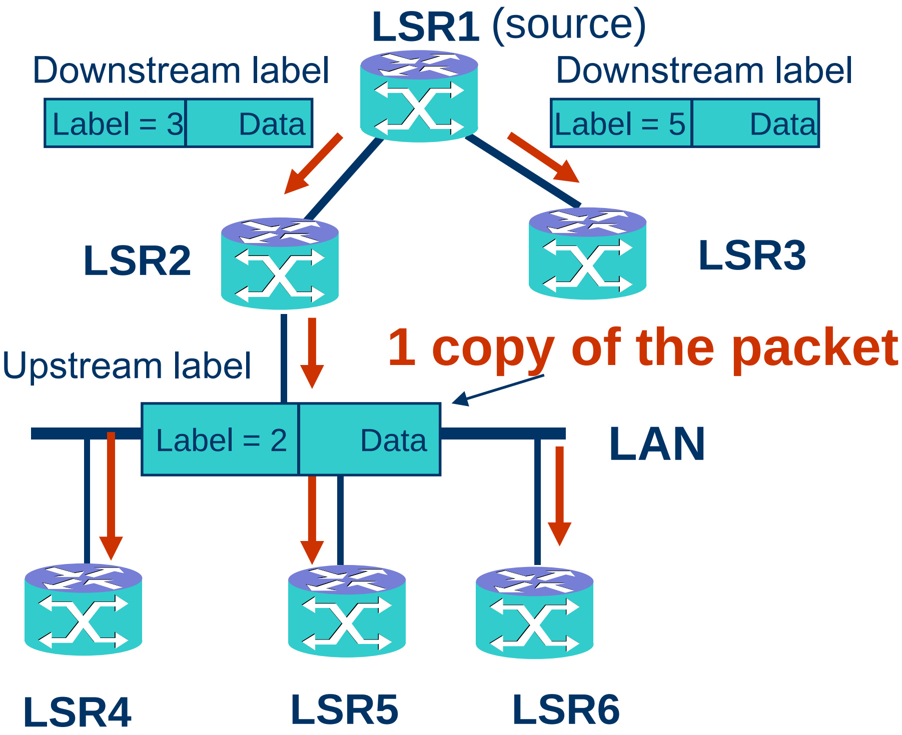
- **Upstream-assigned labels: if the L to FEC binding is first made by Router-upstream and then advertised by Router-upstream to Router-downstream, it is an "*upstream-assigned*" binding/label**
  - **Advertisement is via any label distribution protocol (e.g., LDP, RSVP-TE), but requires new extension**

P2MP LSP over LAN with upstream-assigned label

Control Plane (Label Distribution)



Data Plane (Data Forwarding)

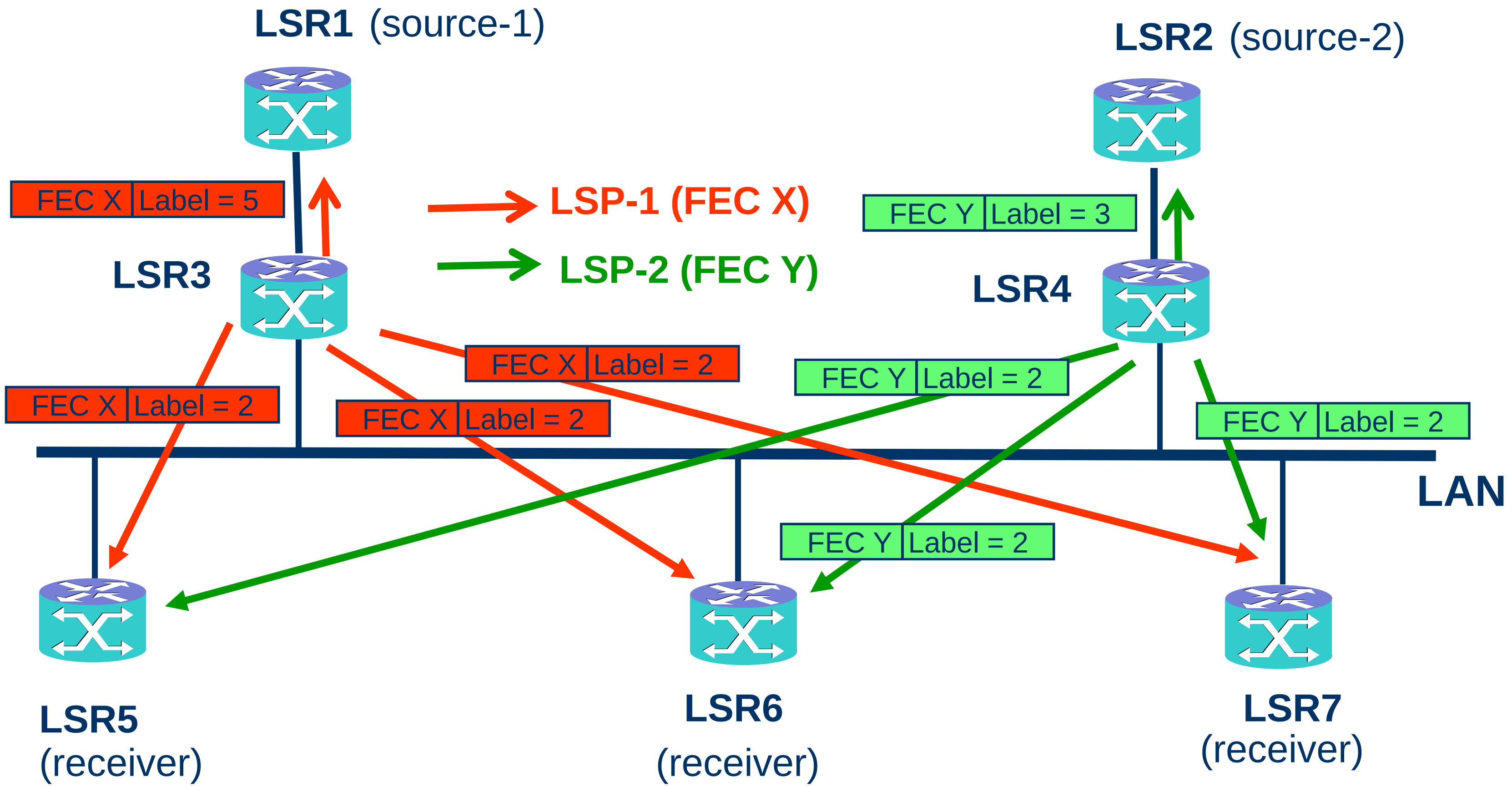


LSR2 is upstream with respect to LSR4, LSR5, and LSR6
LSR4, LSR5, and LSR6 are downstream with respect to LSR2

P2MP LSP over LAN with upstream-assigned label

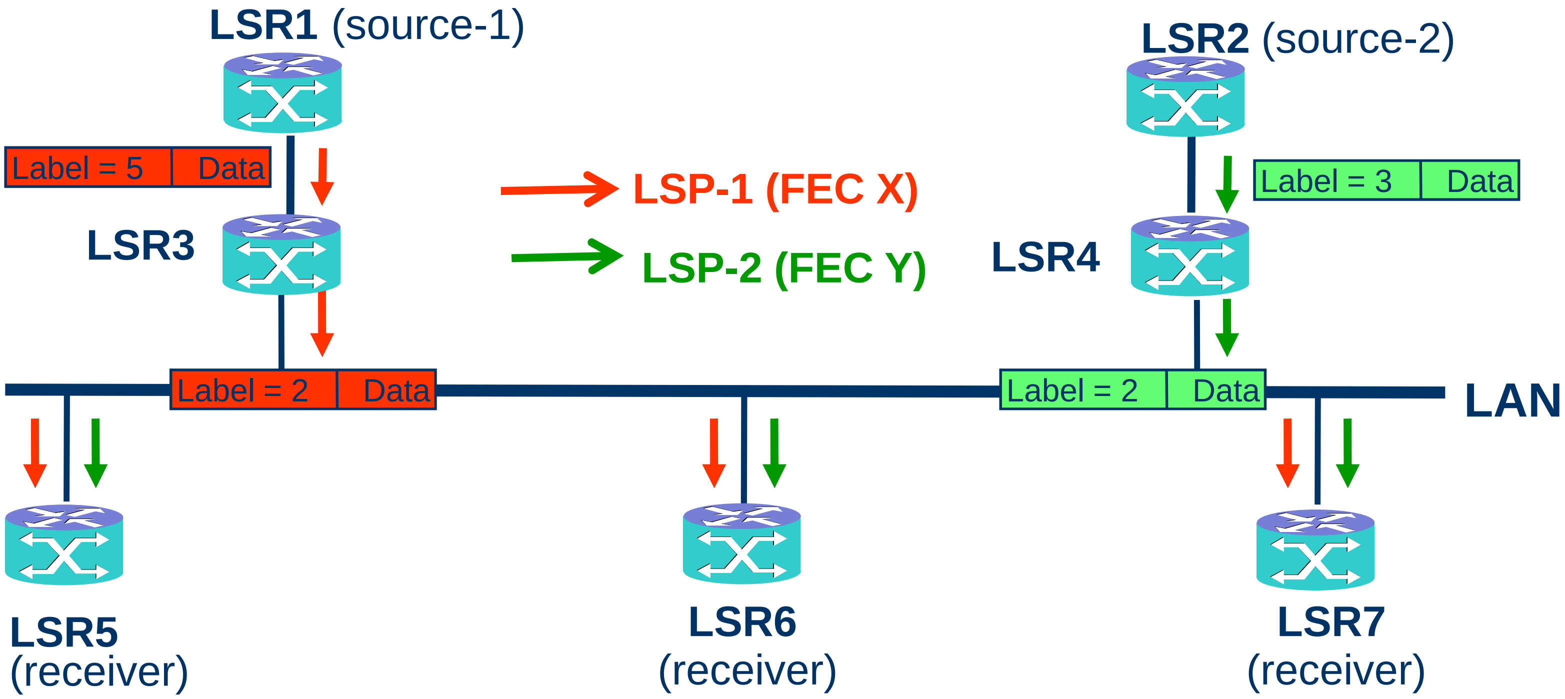
- The upstream router assigns the label
  - Upstream-assigned label
- The upstream router distributes FEC to label binding to the downstream routers on a LAN
- The upstream router sends a single copy of a packet over LAN
  - Destination MAC Address is set to 01-00-5e-8x-xx-xx
    - "x-xx-xx" is the twenty-bit MPLS label value
  - All the downstream routers on a LAN receive the packet
    - As the Destination MAC address is multicast

How to make upstream labels unambiguous ? (control plane)



LSR3 and LSR4 do not coordinate assignment of their upstream labels with each other

How to make upstream labels unambiguous ? (data plane)



Question: How can LSR5, LSR6, LSR7 distinguish between **LSP-1** and **LSP-2** (as both have the same label) ?

Answer: By interpreting incoming label within the context of the upstream router - using the **context specific label space**

Context specific label space

- Context specific label space identifies a particular label space used for MPLS lookup of incoming packets - an MPLS label is looked up in a particular context-specific label space
- Per-platform label space: lookup within the context of the label space associated with the receiving LSR

old

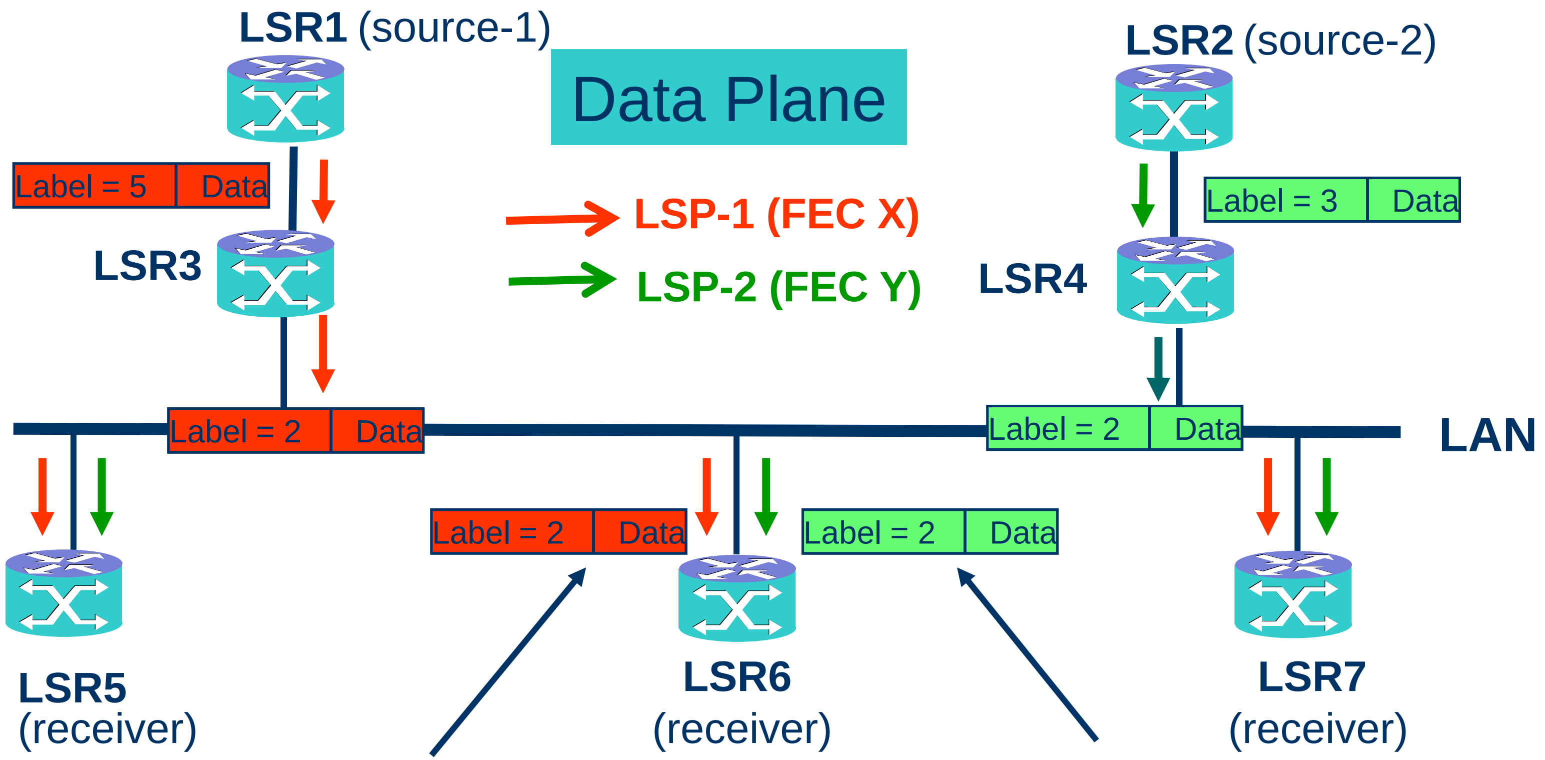
- Per-interface label space: lookup within the context of the label space associated with a particular interface on the receiving LSR

old

- Per-neighbor label space: lookup within the context of the label space associated with a particular neighbor
  - For upstream-assigned labels the context is associated with the sender (upstream neighbor)

new

Making upstream labels unambiguous: per neighbor label space



Label lookup on LSR6 is within the context of the LSR3 label space (sender label space)

Label lookup on LSR6 is within the context of the LSR4 label space (sender label space)

Receiving LSR needs the ability to identify a particular sender LSR on a LAN

Distribution of upstream labels

RSVP-TE:

- RSVP Hello message extensions: CAPABILITY object indicates support for upstream labels
  - Implies support for context specific label spaces
- Path message extensions: UPSTREAM_ASSIGNED_LABEL object – carries the (upstream) label
  - From upstream LSR towards downstream LSR

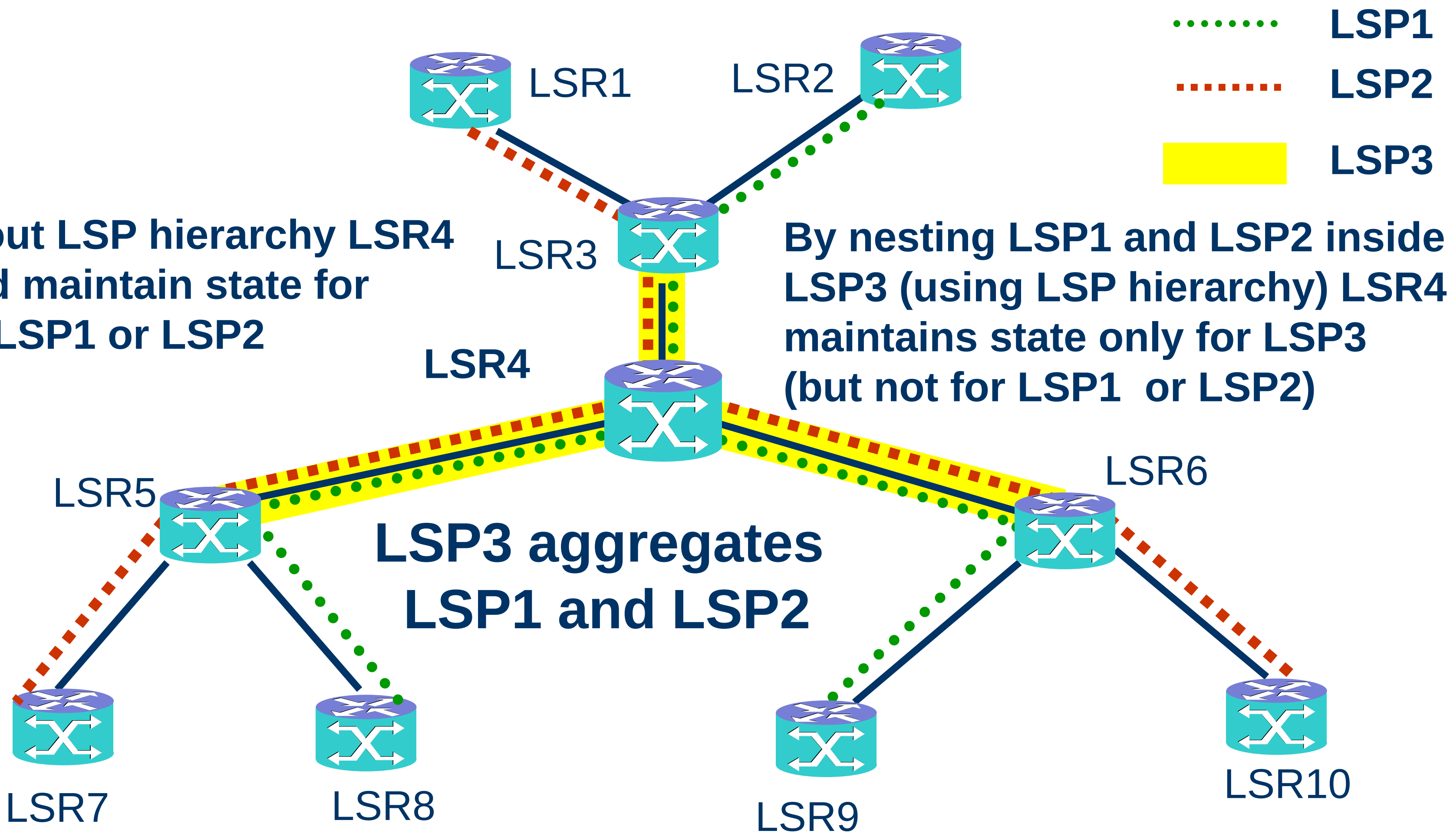
LDP:

- Initialization message extensions: Upstream Label Assignment Capability TLV indicates support for upstream labels
  - Implies support for context specific label spaces
- Label Request message extension: Upstream-Assigned Label Request TLV
  - From downstream LSR towards upstream LSR
- Label Mapping, Label Abort, and Label Withdraw messages extension: Upstream-Assigned Label TLV – carried the (upstream) label
  - From upstream LSR towards downstream LSR

P2MP LSP Hierarchy

- Why ?
  - To improve scalability of the data plane by reducing the MPLS forwarding state for P2MP LSPs
  - To improve scalability of the control plane by reducing the overhead associated with maintaining MPLS forwarding state for P2MP LSPs
    - By reducing MPLS forwarding state for P2MP LSPs
  - For the same reasons we use point-to-point/multipoint-to-point LSP hierarchy with unicast
- How ?
  - By nesting several P2MP inner LSPs inside a common outer P2MP LSP
  - By using the MPLS label stack construct
  - The same as for point-to-point LSP hierarchy with unicast
    - At least conceptually, but there are some differences as well
      - See the following slides for more on this…

P2MP LSP Hierarchy - example

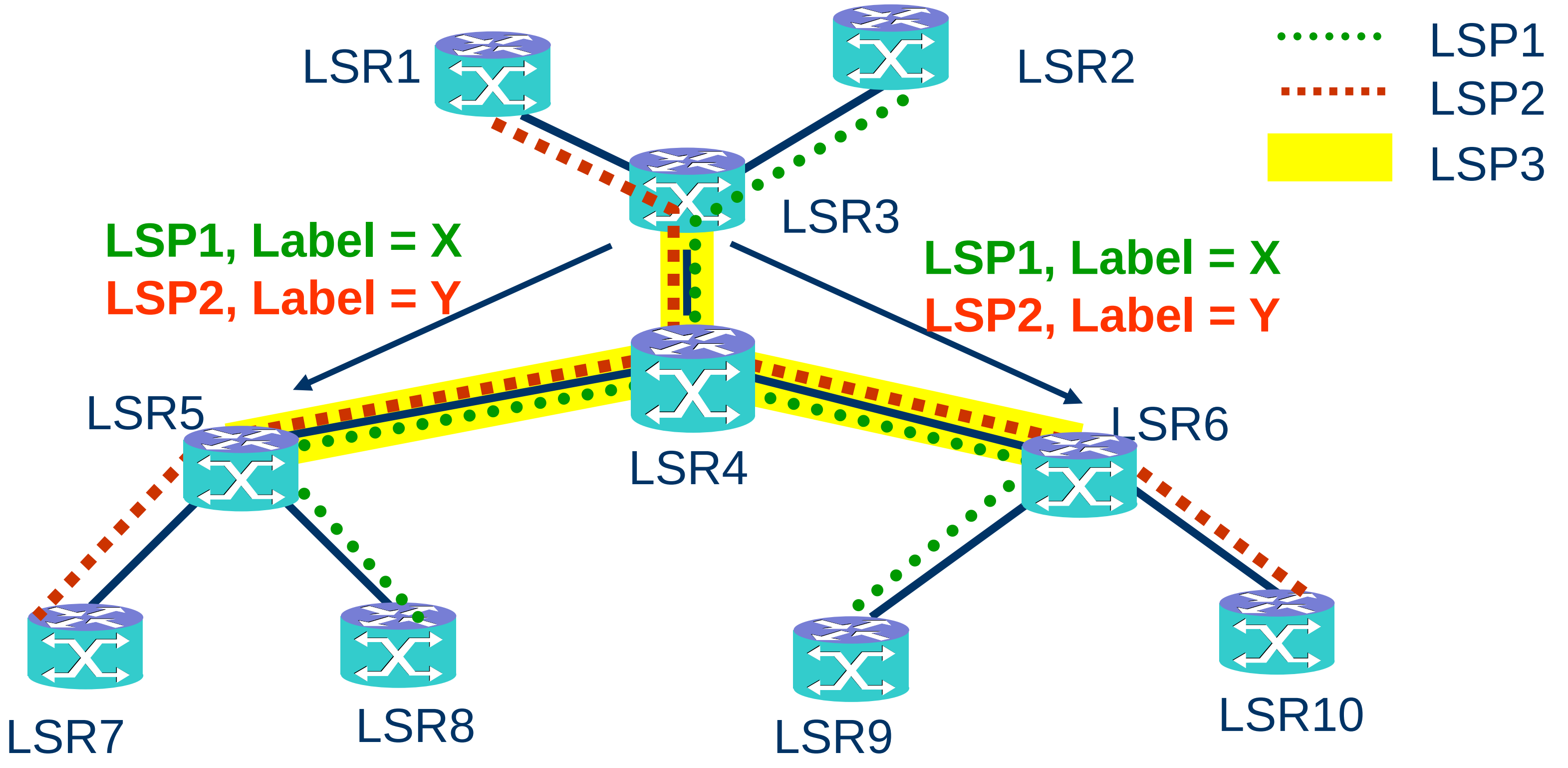


Without LSP hierarchy LSR4 would maintain state for both LSP1 or LSP2

By nesting LSP1 and LSP2 inside LSP3 (using LSP hierarchy) LSR4 maintains state only for LSP3 (but not for LSP1 or LSP2)

LSP3 aggregates LSP1 and LSP2

P2MP LSP Hierarchy: upstream labels



Q: How can LSR5/LSR6 distinguish between LSP1 and LSP2 ?

A: Use upstream labels (LSR3 sends upstream labels to LSR5, LSR6)

P2MP LSP Hierarchy

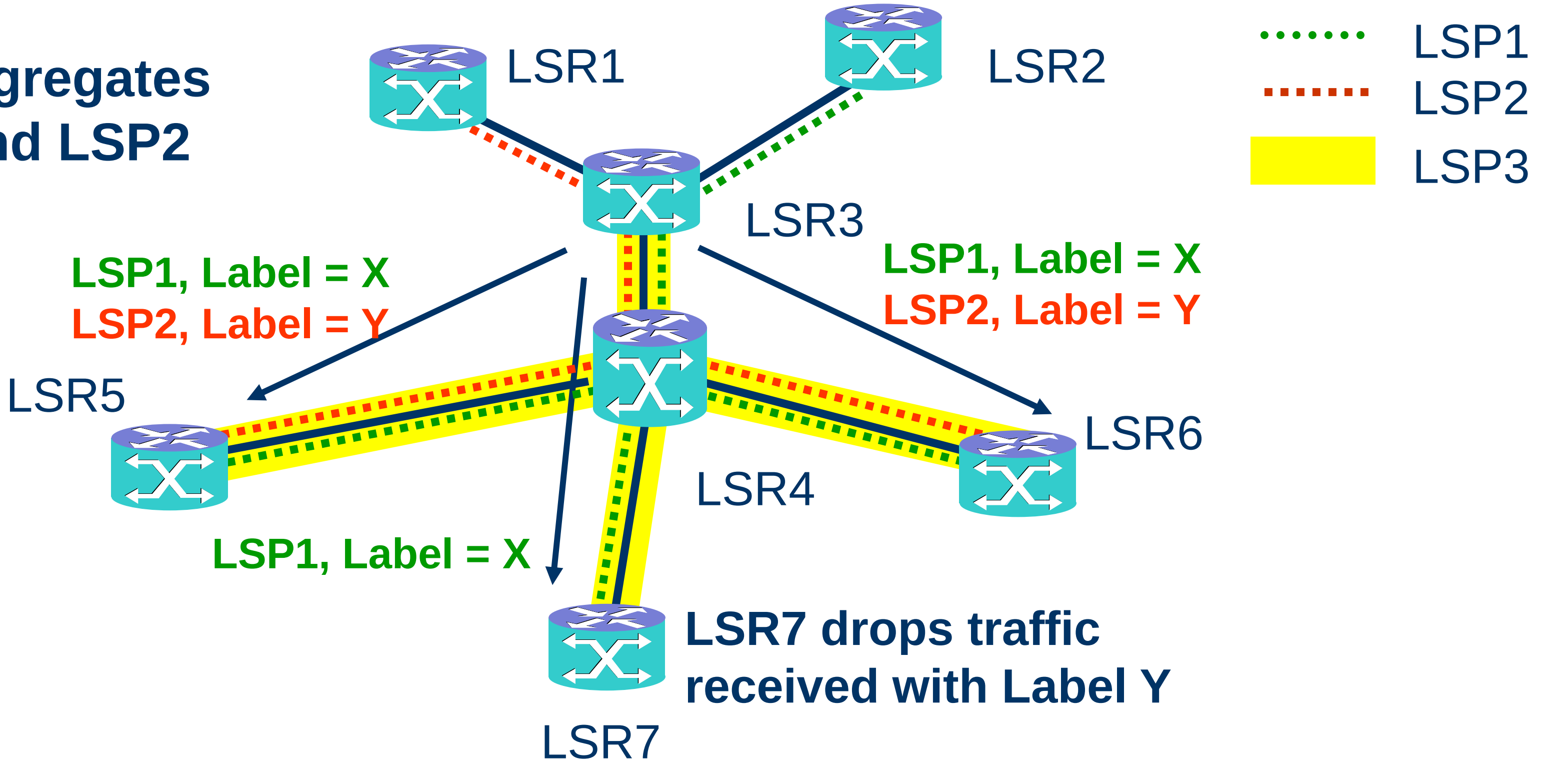
- Several (inner) P2MP LSPs may be nested inside one (outer) P2MP LSP
- Root of the outer P2MP LSP assigns a distinct (upstream) label to each inner P2MP LSP, and communicates this assignment/binding to all the leaves of the outer LSP
- Leaves of the outer P2MP LSP use the label binding information received from the root to distinguish among different inner P2MP LSPs
- Since a given LSR may be a leaf or more than one (outer) P2MP LSP, the (leaf) LSR has to interpret the inner (upstream) labels in the context of the root of the outer P2MP LSP
- LSRs that are leaves of the outer P2MP LSP use the upstream (indirect) neighbor label space
  - The (indirect) upstream neighbor is the root node of the outer P2MP LSP
  - Use of the upstream neighbor specific label space requires to disable penultimate hop popping on the outer LSP at the leaf LSRs

P2MP LSP Hierarchy: communicating upstream label binding

- Root of the outer P2MP LSP assigns a distinct (upstream) label to each inner P2MP LSP (see previous slide)
- Root of the outer LSP uses a label distribution protocol to communicate to the leaves of the outer LSP the label assignment/binding for each inner LSP, and the context
  - The context is the root of the outer LSP
- RSVP-TE as a label distribution protocol: for each inner P2MP LSP the Path message from the root to all the leaves of the outer P2MP includes
  - UPSTREAM_ASSIGNED_LABEL object - carries the upstream label (assigned by the root) for each inner P2MP LSP
  - IF_ID RSVP_HOP object - carries the context (the identity of the outer LSP)
- LDP as a label distribution protocol: for each inner P2MP LSP the Label Mapping message from the root to all the leaves of the outer P2MP LSP includes:
  - Upstream Assigned Label TLV - carries the upstream label (assigned by the root) for each inner P2MP LSP
  - Interface ID TLV - carries the context (the identity of the outer LSP)

P2MP LSP Hierarchy: aggregation with partial congruency

LSP3 aggregates LSP1 and LSP2



LSP1, Label = X
LSP2, Label = Y

LSP1, Label = X
LSP2, Label = Y

LSP1, Label = X

LSR7 drops traffic received with Label Y

Leaves are not exactly the same

LSP1 leaves: LSR5, LSR6, LSR7
LSP2 leaves: LSR5, LSR6
(LSP1 and LSP2 are partially congruent)

P2MP LSP Hierarchy: aggregation with partial congruency

- Several (inner) P2MP LSPs may be nested inside one (outer) P2MP LSP, even if the set of leaf nodes of each of these (inner) P2MP LSPs is not exactly the same
  - Provides more flexibility with respect to the set of P2MP LSPs that could be aggregated
    - By relaxing the leaf congruency constraint on the set of P2MP LSPs that could be aggregated
  - Results in less efficient use of bandwidth compared to the case where the set of leaf nodes of each of the (inner) P2MP LSP is exactly the same
    - Compared to the case where the leaf congruency constraint is enforced

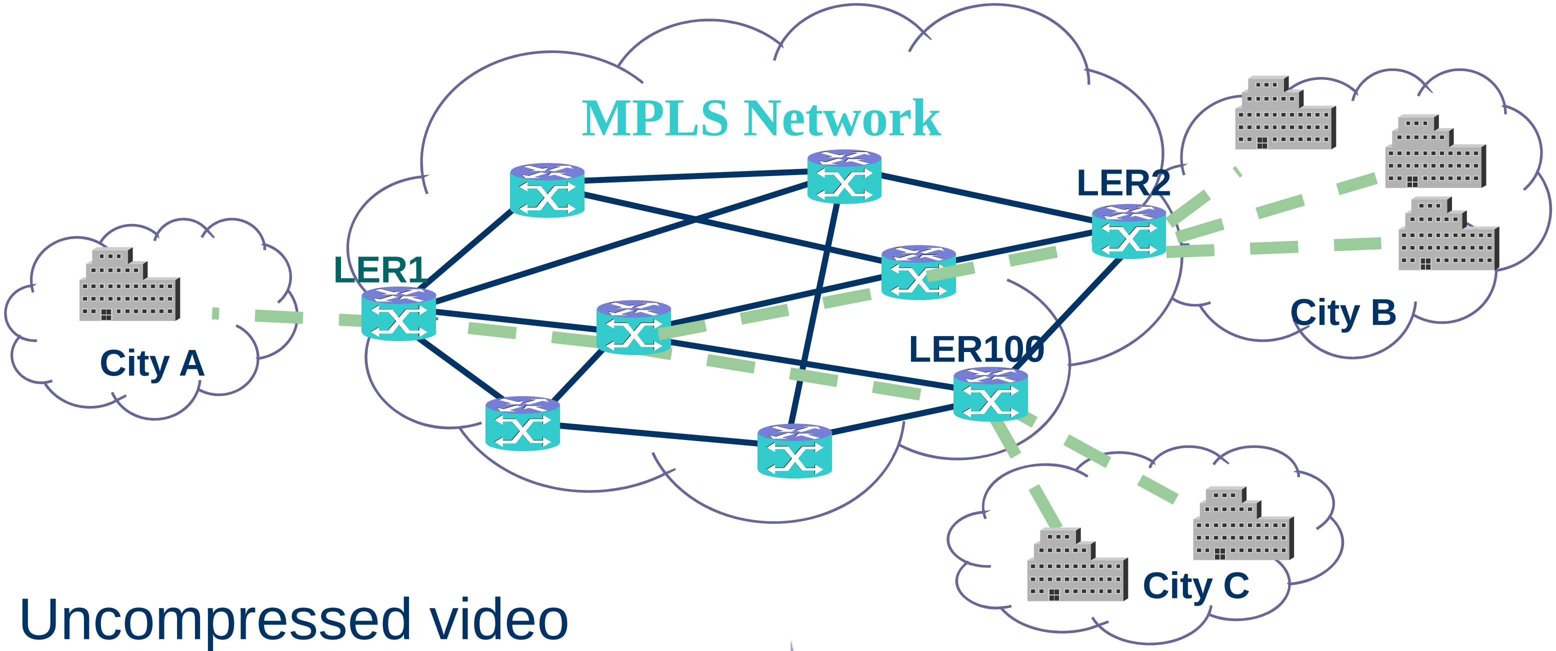
Agenda

- P2MP LSP
  - Setting up P2MP LSP with LDP
  - Setting up P2MP LSP with RSVP-TE
- Upstream labels with P2MP LSPs
  - P2MP LSP over LAN
  - Distribution of upstream labels
  - P2MP LSP Hierarchy
- Usage of P2MP LSPs

Applications: examples

- Video distribution
- Multicast in VPLS
- Multicast in 2547 VPNs

Broadcast TV Distribution between Production Studios



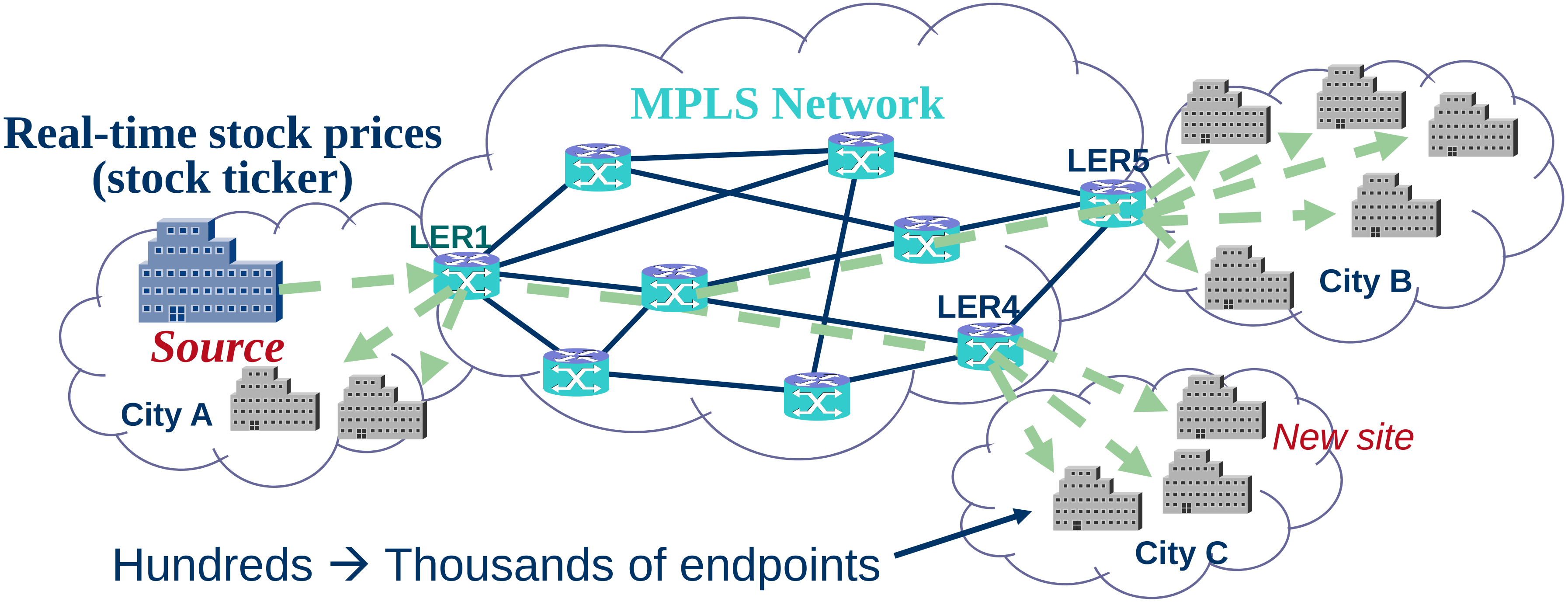
- Uncompressed video 1 Mbps →100 Mbps
- 1+1 protection for all streams
- Future streams as large as 1.5 Gbps

→

Bandwidth at each LSR can grow large very quickly

Multicast improves network performance/efficiency

Financial Information Distribution


MPLS Network
Real-time stock prices (stock ticker)
LER1
LER5
LER4
Source
City A
City B
City C
New site
Hundreds → Thousands of endpoints

Benefits

- Improved network performance/efficiency
- Rapid new site turn-up

Applications: providing discovery for P2MP LSPs

- For P2MP LSP with RSVP-TE: an application provides the root node with the identity of all the leaf nodes
- For P2MP LSP with LDP: an application provides all the leaf nodes with the address of the root node and the identity of the LSP
  - P2MP FEC

Applications: mapping multicast traffic into P2MP LSP

- Mapping multicast traffic into a P2MP LSP
  - At the root node – what traffic to put into a given P2MP LSP
  - At the leaf nodes – how to handle traffic received on a given P2MP LSP
- Mapping has to be consistent among the root and all the leaf nodes of the P2MP LSP
  - Option 1: via provisioning (configuration) at the root and all the leaves
  - Option 2: Root defines the mapping and distributes the mapping to all the leaves
    - Using application-specific procedures (e.g., BGP auto-discovery for multicast in 2547 VPNs)
  - Option 3: (unambiguous) algorithmic mapping
    - E.g., use IP multicast <S,G> as the FEC for a P2MP LSP

Suggested readings

- RFC4875
- draft-ietf-mpls-ldp-p2mp
- draft-ietf-mpls-multicast-encaps
- draft-ieft-mpls-upstream-label
- draft-ietf-mpls-rsvp-upstream
- draft-ietf-mpls-ldp-upstream

Section 2

Multicast in BGP/MPLS VPNs

Agenda

- ⇨ BGP/MPLS MVPN – what are the goals ?
- Supporting PIM-SM in SSM mode MVPNs
- Supporting PIM-SM in ASM mode MVPNs
- Summary

BGP/MPLS MVPN – what are the goals ?

- Extend 2547 VPN service offering to include support for IP multicast for 2547 VPN customers
- Follow the same architecture/model as 2547 VPN unicast
  - No need to have the Virtual Router model for multicast and the 2547 model for unicast
- Re-use 2547 VPN unicast mechanisms, with extensions, as necessary
  - No need to have PIM/GRE for multicast and BGP/MPLS for unicast
- Retain as much as possible the flexibility and scalability of 2547 VPN unicast

Agenda

- BGP/MPLS MVPN – what are the goals ?
- Supporting PIM-SM in SSM mode MVPNs
- Supporting PIM-SM in ASM mode MVPNs
- Summary

IP Multicast with PIM-SM in SSM mode and MVPN

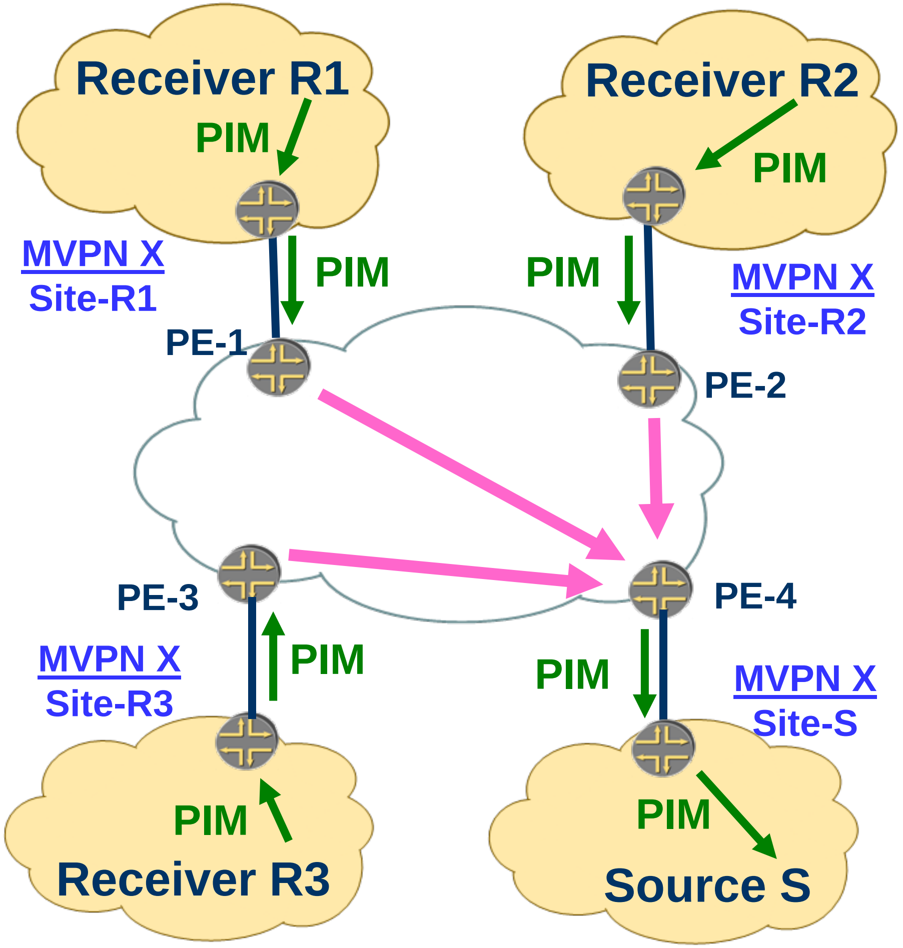
Plain IP Multicast:

- Multicast Sources need to know that there are Multicast Receivers – the Receivers have to inform the Sources that the Receivers want to receive traffic from the Sources
- With PIM-SM in SSM mode the Receivers discover the Sources by means outside of PIM
- There has to be multicast forwarding state from the Sources to the Receivers to carry multicast traffic from the Sources to the Receivers

In the context of MVPN:

- Carrying multicast routing information from the Receivers to the Sources may involve MVPN service providers
  - As multicast sources and multicast receivers may be in different sites
- Carrying multicast data traffic from the Sources to the Receivers may involve MVPN service providers
  - As multicast sources and multicast receivers may be in different sites
- Different MVPNs may use the same address space (e.g., RFC1918), including IP multicast addressing space

Carrying MVPN multicast routing information (from Receivers to Source)



Step 1: from the Receivers to the PEs of the sites that contain the Receivers (PEs of Site-R1, Site-R2, Site-R3)

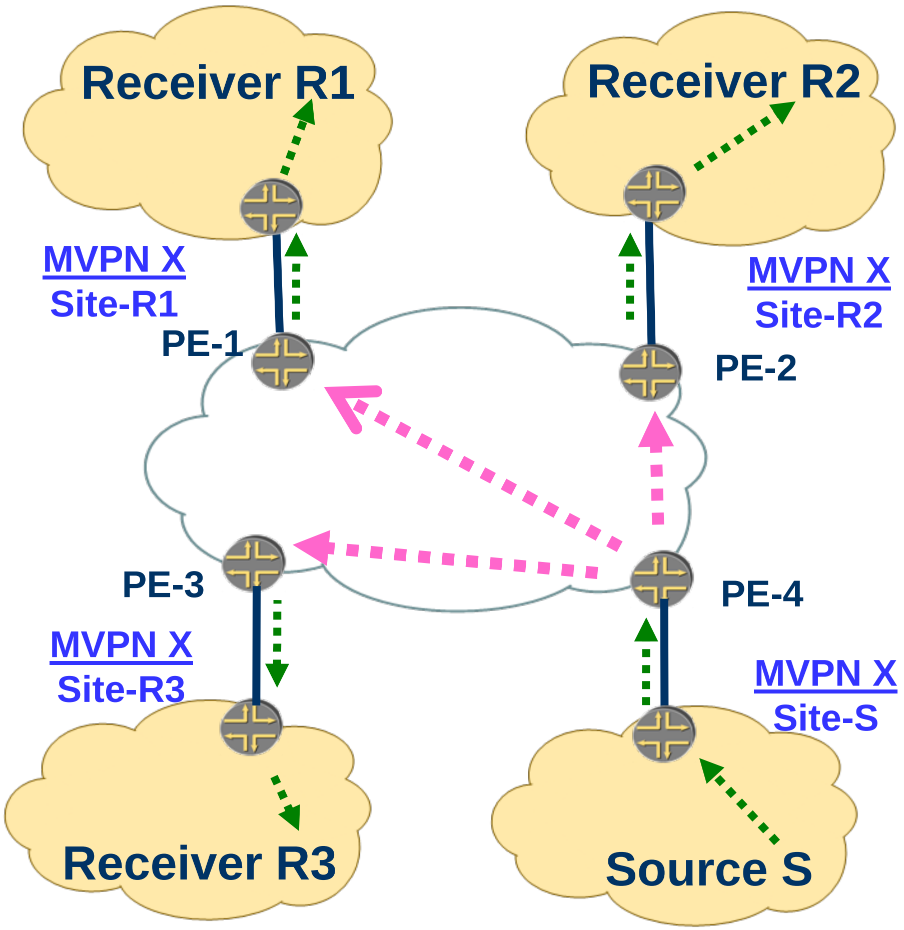
- Using plain PIM

Step 2: from the PEs of the sites that contain the Receivers (PEs of Site-R1, Site-R2, Site-R3) to the PE of the site that contains the Source (PE of Site-S)

Step 3: from the PE of the site that contains the Source (PE of Site-S) to the Source S

- Using plain PIM

Carrying MVPN multicast data traffic (from Source to Receivers)



Step 1: from the Source S to the PE of the site that contains the Source (PE of Site-S)

- Using forwarding state established by plain PIM

Step 2: from the PE of the site that contains the Source (PE of Site-S) to the PEs of the sites that contain the Receivers (PEs of Site-R1, Site-R2, Site-R3)

Step 3: from the PEs of the sites that contain the Receivers (PEs of Site-R1, Site-R2, Site-R3) to the Receivers

- Using forwarding state established by plain PIM

MVPN – what are the required service provider mechanisms ?

- A mechanism to carry MVPN multicast routing information from the PEs connected to the sites that contain the Receivers to the PE connected to the site that contains the Source
  - E.g., from the PEs connected to Site-R1, Site-R2, Site-R3 to the PE connected to Site-S
  - So that Receivers inform the source S that the Receivers want to receive traffic from S
- A mechanism to carry multicast traffic from the PE connected to the site that contains the Source to the PEs connected to the sites that contain the Receivers
  - E.g., from the PE connected to Site-S to the PEs connected to Site-R1, Site-R2, Site-R3
  - So that multicast traffic will flow from the source S to the Receivers

Agenda

- BGP/MPLS MVPN – what are the goals ?
- Supporting PIM-SM in SSM mode MVPNs
  - ➡ Carrying MVPN multicast routing information
  - Carrying MVPN multicast traffic
- Supporting PIM-SM in ASM mode MVPNs
- Summary

Carrying MVPN multicast routing information: BGP C-multicast routes

- ***BGP C-multicast (customer multicast) routes*** **carry MVPN customer multicast routing information from the PEs connected to the sites that contain the Receivers to the PE connected to the site that contains the Source**
- **C-multicast routes are carried in Multiprotocol BGP (RFC4760) using new NLRI – MCAST-VPN**
- **C-multicast route NLRI contains:**
  - **Multicast Source (S), Multicast Group (G)**
  - **Route Distinguisher (RD)**
    - **Needed to support MVPNs that may use the same address space (just like with unicast)**
- **Import of a C-multicast route into a VRF is controlled by the Route Target carried by the route**
  - **A C-multicast route is imported into a VRF if the VRF Route Import of this VRF matches the Route Target carried in the C-multicast route**
- **Re-use the existing BGP mechanisms (e.g., extended communities, Route Target constraint, Route Reflectors, etc…)**

Import of C-multicast routes

- C-multicast route carrying a given Multicast Source S should be imported ONLY into the VRF on the PE connected to the site that contains S
  - And not into any other VRF, even of the same MVPN (this is in contrast to VPN-IPv4 routes)
- To accomplish this each VRF has an additional import Route Target, called *C-multicast Import RT*:
  - In addition to the Route Targets used to control import of VPN-IPv4 routes and auto-discovery routes
  - C-multicast import RT controls import of C-multicast routes into VRF

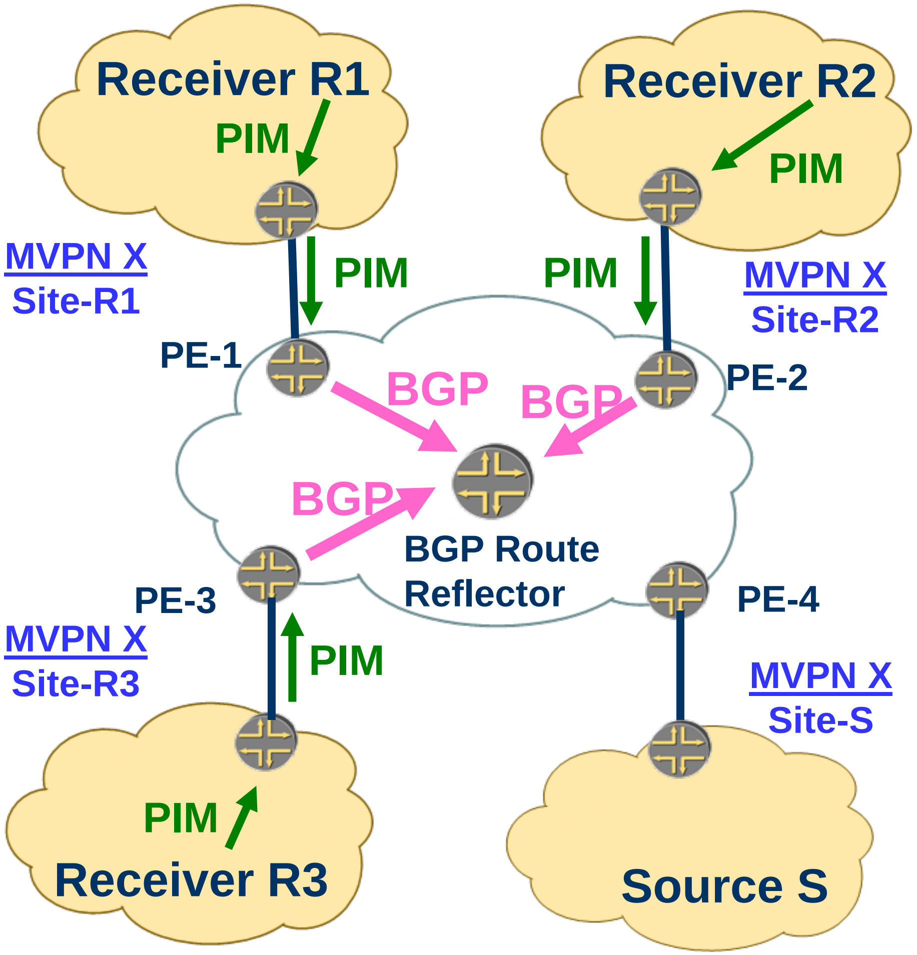
AND

- All the VRFs within a given MVPN have the information about VRF Route Imports of each of these VRFs
  - Accomplished by encoding the value of C-multicast Import RT in the VRF Route Import extended community, and carrying VRF Route Import with the (unicast) VPN-IPv4 routes
- To make a C-multicast route carrying Multicast Source S be imported only into the VRF on the PE connected to the site that contains S:
  - find the (unicast) VPN-IPv4 route to S
  - set the Route Target of the C-multicast route to the VRF Route Import carried by the found VPN-IPv4 route

More on C-multicast RT

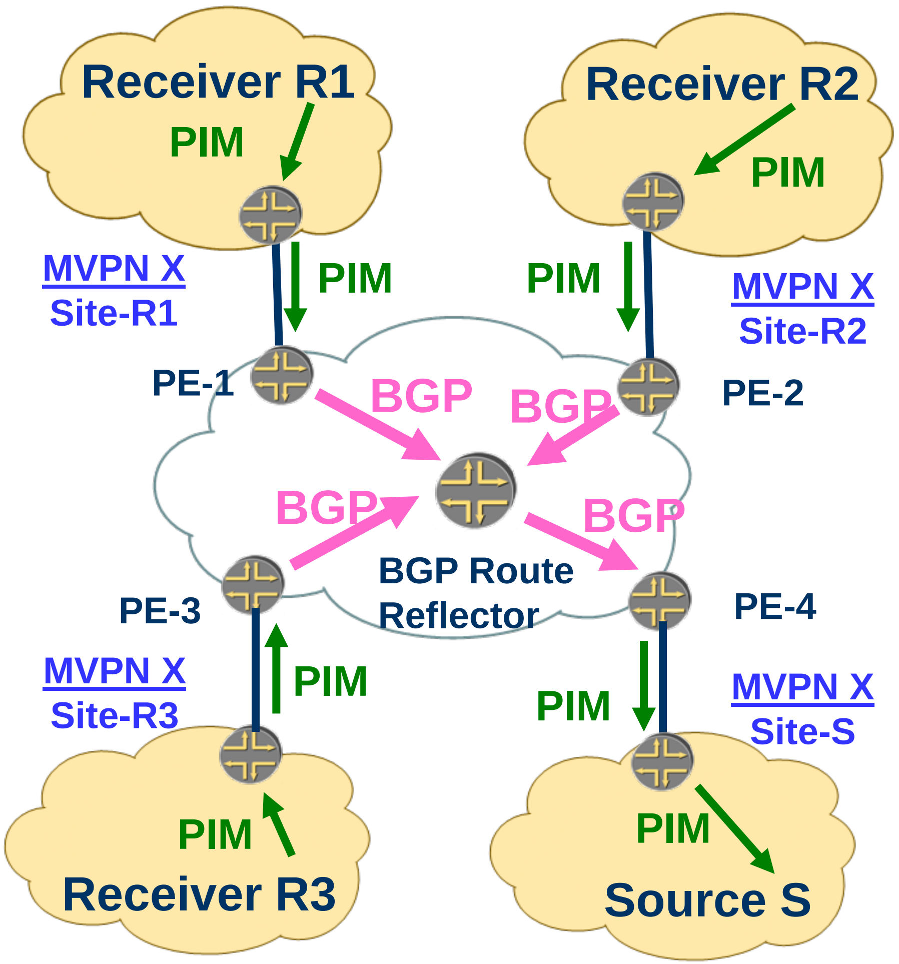
- C-multicast Import RT controls import of C-multicast routes into a VRF
- C-multicast Import RT needs to be distinct across all VRFs on all PEs
  - As a given C-multicast route needs to be imported only into the VRF of the PE connected to the site that contains S
  - Contains PE’s IP address + local (to the PE) number ->
    - Different MVPNs within a given PE have different C-multicast Import RTs
    - Within a given MVPN VRFs on different PEs have different C-multicast Import RTs
- C-multicast Import RT is auto-configured

Originating C-multicast route



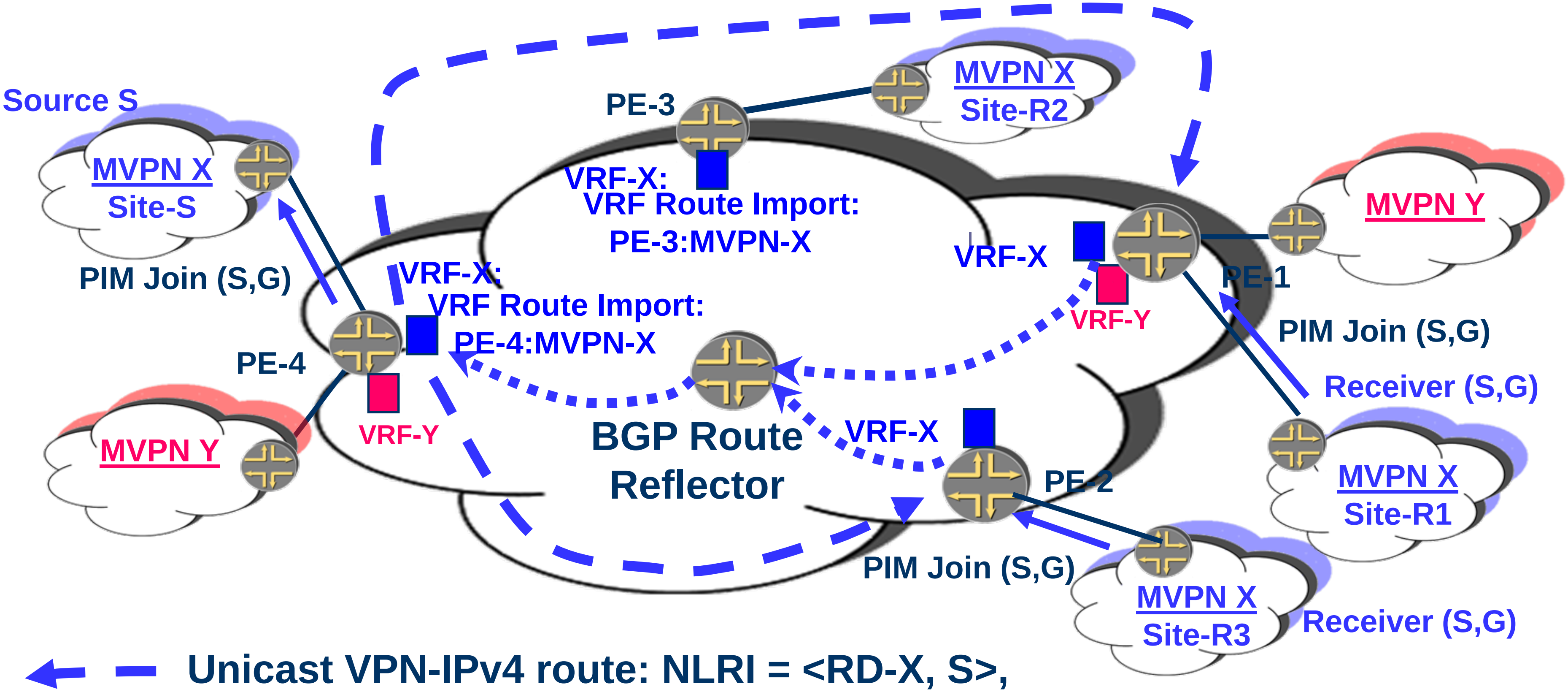
- PE-1 determines that Site-R1 has receiver(s) for (S,G)
  - PE-1 receives from Site-R1's CE PIM Join (S, G)
- PE-1 constructs and originates a C-multicast route as follows:
  - Finds (unicast) VPN-IPv4 route to S
  - Extracts from this route: RD and VRF Route Import
  - C-multicast route carries:
    - <Source, Group> - from PIM Join (S, G)
    - RD – from the VPN-IPv4 route
    - Route Target – constructed from VRF Route Import of the VPN-IPv4 route
- Same applies to PE-2 and PE-3

Receiving C-multicast route



- **PE-4 receives the C-multicast route originated by PE-1 (PE-2, PE-3)**
  - **These C-multicast routes are aggregated by BGP Route Reflector**
- **PE-4 accepts the C- multicast route into the VRF for MVPN X**
  - **Because the VRF Route Import on this VRF matches the Route Target carried in the C-multicast route**
- **PE-4 creates (S,G) state in the VRF, and propagates <S,G> information to the CE of Site-S (the site that contains S)**
  - **Using PIM as PE-CE multicast routing protocol**

Carrying MVPN multicast routing information – putting it together



◄ — Unicast VPN-IPv4 route: NLRI = <RD-X, S>,
VRF Route Import = <PE-4:MVPN-X>

◄ ▪▪▪ C-multicast route: NLRI = <RD-X, S, G>,
Route Target = <PE-4:MVPN-X>

Digression on Route Target Constraint (RT Constraint)

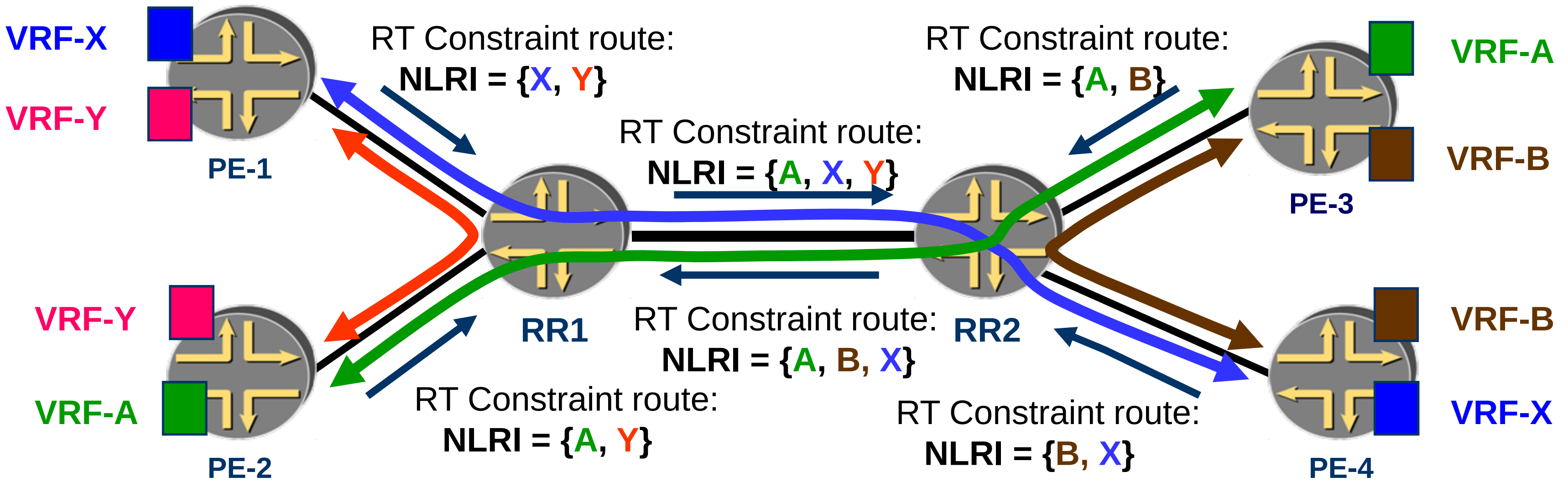
Route Target Constrain (RT Constraint) Primer

- **Distribution of BGP routes that carry the Route Target (RT) extended community could be constrained by filtering based on the Route Target**
  - **Outbound filtering: a BGP speaker could advertise to its peer only the routes that carry one or more Route Targets from a particular set of Route Targets**
  - **Inbound filtering: a BGP speaker accepts from its peer only the routes that carry one or more Route Targets from a particular set of Route Targets**
  - **Applies to any BGP routes that carry Route Target extended community: VPN-IPv4 routes, C-multicast routes, etc…**
- **RT Constraint mechanism (RFC4684) uses BGP to distribute such filters among PEs, Route Reflectors, and ASBRs**
  - **Both within a single Autonomous System (AS), as well as across multiple Autonomous Systems (ASes)**
  - **By using RT Constraint routes**

Carrying RT Constraint in BGP

- *RT Constraint routes* are carried in Multiprotocol BGP (RFC4760) using new NLRI – RT Membership
  - NLRI format: <AS:RT>
- A single RT Membership NLRI can carry a set of RTs
  - Expressed as an RT Membership prefix
- RT Constraint routes are handled and distributed similar to any other BGP routes
- RT Constraint routes advertised by a router to its peer provide the peer with the set of RTs; the peer should advertise to the router only the routes that carry one or more RTs from that set
- Distribution of a route that carry a given RT is constrained by sending the route only along the reverse path of the best learned <AS:RT> RT Constraint route
  - Applies to the distribution of any route (e.g., VPN-IPv4 route, C-multicast route, etc…), as long as this route carries the RT

RT Constraint use with VPN-IPv4 routes: example



VRF-X

VRF-Y

PE-1

RT Constraint route:
NLRI = {X, Y}

RT Constraint route:
NLRI = {A, B}

VRF-A

VRF-B

PE-3

RT Constraint route:
NLRI = {A, X, Y}

VRF-Y

VRF-A

PE-2

RR1

RT Constraint route:
NLRI = {A, B, X}

RR2

VRF-B

VRF-X

PE-4

RT Constraint route:
NLRI = {A, Y}

RT Constraint route:
NLRI = {B, X}

VRF-A: Export RT A, Import RT A
VRF-B: Export RT B, Import RT B
VRF-X: Export RT X, Import RT X
VRF-Y: Export RT Y, Import RT Y

⟷ VPN-IPv4 routes for VPN A
⟷ VPN-IPv4 routes for VPN B
⟷ VPN-IPv4 routes for VPN X
⟷ VPN-IPv4 routes for VPN Y

No VPN B routes on RR1 !
No VPN Y routes on RR2 !

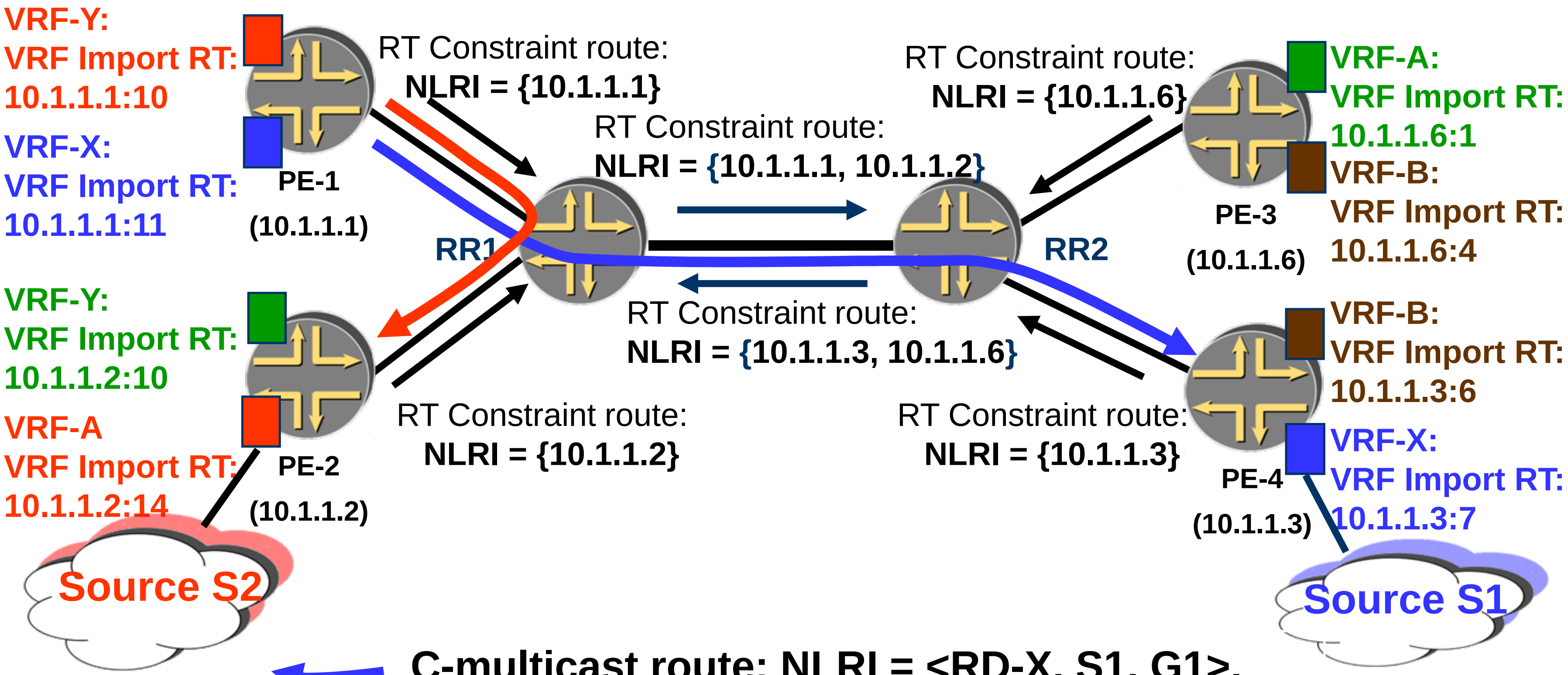
Note: RT Membership NLRI does not show AS number

RT Constraint use with C-multicast routes

- Each PE advertises a single RT Constraint route[1]
  - RT Membership NLRI of the route contains PE’s AS number + PE’s IP address
    - RT Membership prefix
- Single RT Constraint route advertised by a PE covers all the VRF Route Import RTs present on the PE
  - By virtue of assignment of VRF Route Import RTs
    - As each VRF Route Import RT consists of PE’s IP address + local (to the PE) number

[1] PE may also advertise other RT Constraint routes to constrain distribution of VPN-IPv4 routes

RT Constraint use with C-multicast routes: example



VRF-Y:
VRF Import RT:
10.1.1.1:10

VRF-X:
VRF Import RT:
10.1.1.1:11

PE-1
(10.1.1.1)

RT Constraint route:
NLRI = {10.1.1.1}

RT Constraint route:
NLRI = {10.1.1.6}

RT Constraint route:
NLRI = {10.1.1.1, 10.1.1.2}

VRF-A:
VRF Import RT:
10.1.1.6:1

VRF-B:
VRF Import RT:
10.1.1.6:4

PE-3
(10.1.1.6)

RR1

RR2

VRF-Y:
VRF Import RT:
10.1.1.2:10

VRF-A
VRF Import RT:
10.1.1.2:14

PE-2
(10.1.1.2)

RT Constraint route:
NLRI = {10.1.1.3, 10.1.1.6}

VRF-B:
VRF Import RT:
10.1.1.3:6

VRF-X:
VRF Import RT:
10.1.1.3:7

RT Constraint route:
NLRI = {10.1.1.2}

RT Constraint route:
NLRI = {10.1.1.3}

PE-4
(10.1.1.3)

Source S2

Source S1

C-multicast route: NLRI = <RD-X, S1, G1>,
Route Target = <10.1.1.3:7>

C-multicast route: NLRI = <RD-Y, S2, G2>,
Route Target = <10.1.1.2:14>

Note: RT Membership NLRI does not show AS number

End of Digression on Route Target Constraint (RT Constraint)

Agenda

- BGP/MPLS MVPN – what are the goals ?
- Supporting PIM-SM in SSM mode MVPNs
  - Carrying MVPN multicast routing information
  - ➡ Carrying MVPN multicast traffic
- Supporting PIM-SM in ASM mode MVPNs
- Summary

Carrying MVPN multicast traffic: inter-PE tunnels and BGP auto-discovery routes

- **Inter-PE tunnels are used to carry MVPN multicast data traffic from the PEs connected to the sites that contain the Sources to the PEs connected to the sites that contain the Receivers**
- ***BGP auto-discovery routes* perform two functions:**

  (1) **Enable establishment of inter-PE tunnels**

  - **Auto-discovery routes do NOT (directly) establish tunnels – tunnels are established by the appropriate signaling protocol associated with a particular type of a tunnel (e.g., RSVP-TE, LDP)**
  - **Signaling protocols use the information carried in the auto-discovery routes**

  (2) **Bind one or more MVPN multicast <Source (S), Group (G)> streams to a particular inter-PE tunnel**

  - **Many-to-one binding**

More on auto-discovery routes

- Auto-discovery routes are carried in Multiprotocol BGP (RFC4760) using MCAST-VPN NLRI
- Handled similar to VPN-IPv4 routes:
  - RDs to distinguish among different MVPNs
  - import/export based on Route Target extended communities
- Re-use the existing BGP mechanisms (e.g., extended communities, Route Target constraint, Route Reflectors, etc…)

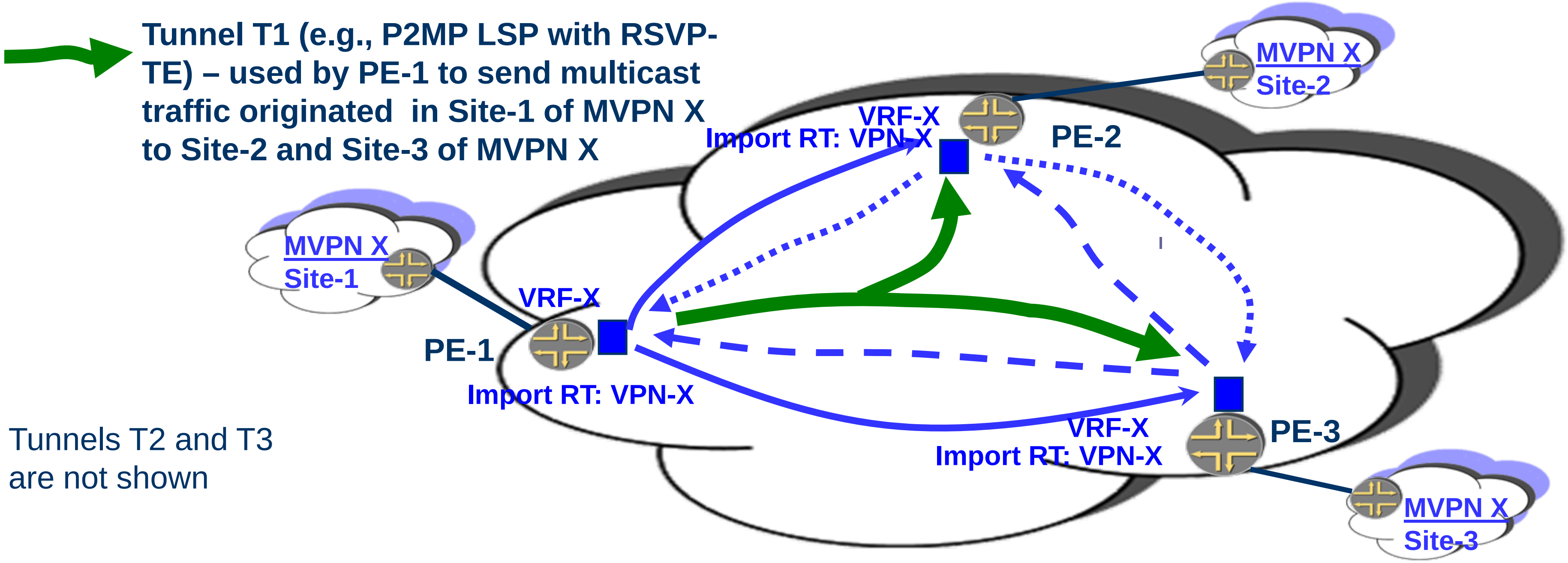
More on auto-discovery routes (cont.)

- PE connected to a MVPN site advertises an auto-discovery route that carries the export Route Targets for that MVPN configured on the PE
- PE connected to an MVPN site imports received auto-discovery route into the VRF of that MVPN only if at least one of the import Route Targets configured on the PE for that MVPN matches the Route Target carried in the auto-discovery route
- MVPN import/export Route Targets can be the same as Route Targets used by unicast 2547 VPNs

Auto-discovery routes and inter-PE tunnels

- **Auto-discovery routes received by a PE provides the PE connected to a MVPN site with:**
  - **(a) the information about the identity of other PEs connected to the sites of that MVPN**
    - **Auto-discovery route originated by a PE carries the identity of that PE (IP address of the PE)**
  - **(b) the identity of the tunnels used by other PEs for sending multicast traffic of that MVPN**
    - **Auto-discovery route originated by a PE for a particular MVPN carries the identity of the inter-PE tunnel that the PE will use to send to other PEs that have sites of that MVPN multicast traffic coming from the sources in the MVPN site(s) connected to that PE**
    - **Tunnel identity is carried in the BGP PMSI Tunnel attribute of the auto-discovery route**
    - **Tunnel identity includes the type of the Tunnel signaling protocol (e.g., RSVP-TE, LDP, etc…)**
- **Combination of (a) and (b) provides sufficient information for tunnel signaling**

Auto-discovery routes and inter-PE tunnels: example



Tunnel T1 (e.g., P2MP LSP with RSVP-TE) – used by PE-1 to send multicast traffic originated in Site-1 of MVPN X to Site-2 and Site-3 of MVPN X

Tunnels T2 and T3 are not shown

BGP Auto-Discovery Route originated by PE-1 for MVPN X:
<RD= RD-X, Origin PE = PE-1, RT = VPN-X, Tunnel-ID = (RSVP-TE, T1)>

BGP Auto-Discovery Route originated by PE-2 for MVPN X:
<RD= RD-X, Origin PE = PE-2, RT = VPN-X, Tunnel-ID = (RSVP-TE, T2)>

BGP Auto-Discovery Route originated by PE-3 for MVPN X:
<RD= RD-X, Origin PE = PE-3, RT = VPN-X, Tunnel-ID = (RSVP-TE, T3)>

Binding multicast streams to inter-PE tunnels: Inclusive vs Selective tunnels

- Inclusive tunnel advertised by a given PE carries MVPN multicast streams:
  - from all the sources in the MVPN site(s) connected to the PE
  - all the multicast streams originated by the sources
  - to all the PEs connected to all other sites of that MVPN
    - Even if some of these sites have no actual receivers for the traffic, HOWEVER
    - CEs in the sites with no actual receivers do not receive the traffic

- Selective tunnel advertised by a given PE carries MVPN multicast streams:
  - only from a particular source(s) in the MVPN site(s) connected to the PE
  - only a subset of multicast streams originated by the sources
  - to only the PEs connected to the other sites of that MVPN that have actual receivers for the traffic

Inclusive tunnels require less forwarding state than Selective tunnels.

Selective tunnels are more bandwidth efficient than Inclusive tunnels.

More on Inclusive and Selective tunnels

- Both Inclusive and Selective tunnels are established using the information carried in BGP auto-discovery routes
- BGP auto-discovery routes for Selective tunnels carry additional information about specific MVPN multicast sources (S) and groups (G) that would be carried over the Selective tunnels
  - Binds a particular <S,G> of a particular MVPN to a particular Selective tunnel
- Creating selective tunnel for (S,G) of a given MVPN is controlled solely by the PE connected to the site that contains S

More on inter-PE tunnels: tunneling technologies

- Available tunneling technologies:
  - MPLS-based: P2MP LSP with RSVP-TE, P2MP LSP with LDP, Ingress replication
  - GRE-based: PIM-SSM with GRE encapsulation, PIM-SM with GRE encapsulation, PIM-Bidir with GRE encapsulation
- Different MVPNs within the same Service Provider may use different tunneling technologies
- For a given inter-AS/inter-provider MVPN (an MVPN that spans multiple ASes/providers) each provider may use different tunneling technology
  - Segmented inter-AS tunnels

More on inter-PE tunnels: aggregation

- **Aggregate multiple P2MP LSPs using P2MP LSP hierarchy**
  - **Multiple P2MP LSPs (rooted at the same PE) may be “nested” inside a single (outer) P2MP LSP**
  - **Applicable to P2MP LSPs used for both Inclusive and Selective tunnels**
  - **Not constrained by MVPN boundaries**
    - **P2MP LSPs of multiple MVPNs could be aggregated into a single P2MP LSP**
  - **Could use aggregation with partial congruency**
    - **If the aggregated P2MP LSPs are only partially congruent with each other**
- **Aggregate multiple (S,G) of a given MVPN into a single Selective tunnel**
  - **As long as all the sources are in the sites connected to the PE that creates the Selective tunnel**
- **Results in improved scalability by reducing both forwarding plane and control plane overhead**

Agenda

- **BGP/MPLS MVPN – what are the goals ?**
- **Supporting PIM-SM in SSM mode MVPNs**
- **Supporting PIM-SM in ASM mode MVPNs**
  - ➡ Carrying MVPN multicast routing information
  - Carrying MVPN multicast traffic
- **Summary**

Supporting PIM-SM in the ASM mode: few observations (1)

In the context of plain IP multicast:

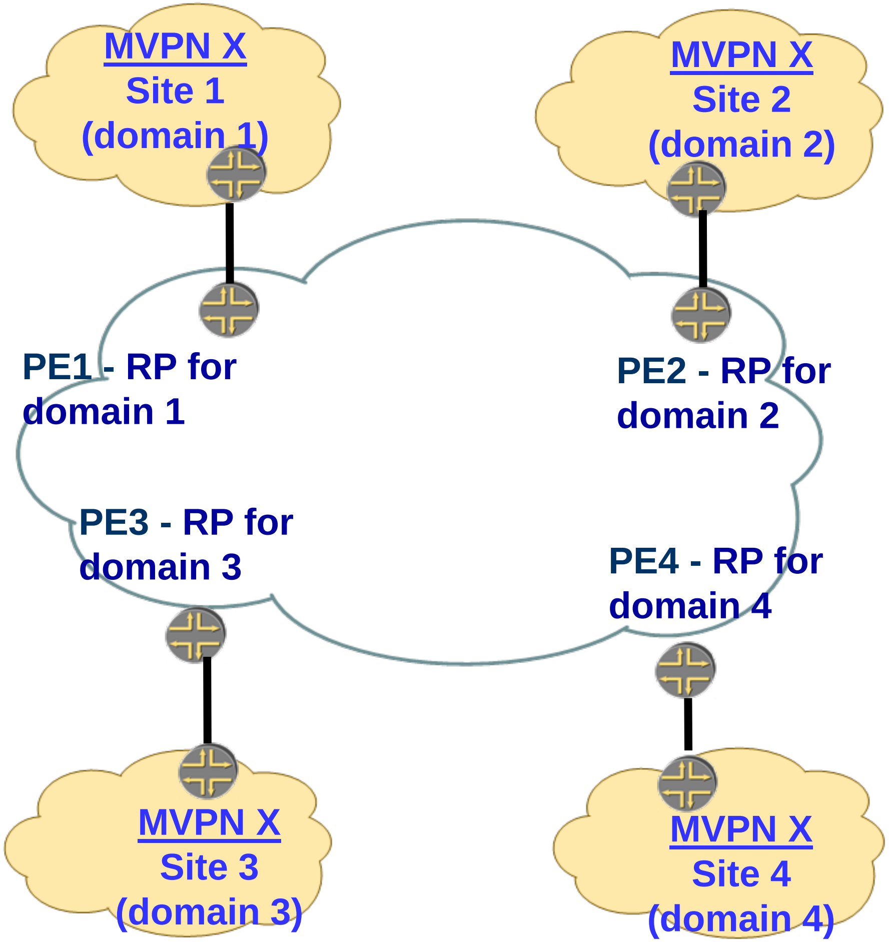
- The sole purpose of joining RP Tree (RPT) is for the receivers to discover the sources
  - As RP knows about all active sources
    - Designated Routers connected to active sources register the active sources with RP (using PIM Register)
- Switching from RP Tree (RPT) to Shortest Path Tree (SPT) usually occurs as soon as a receiver discovers a source
  - Usually on the first packet received from the source
- RPT/SPT interaction introduces a fair amount of additional complexity
  - PIM-SM in the SSM mode is (significantly) simpler than PIM-SM in the ASM mode

Supporting PIM-SM in the ASM mode: few observations (2)

<u>In the context of plain IP multicast:</u>

- Interconnecting multiple multicast domains with MSDP:
  - Within each domain use PIM-SM in the ASM mode
  - Exchange information about active (multicast) sources among all the participating domains
    - By using MSDP among RPs
  - Exchange only (S, G) state among domains
    - No (*, G) state exchange among domains
- Results in <u>PIM-SM in the SSM</u> mode behavior at the <u>inter-domain</u> level, while preserving <u>PIM-SM in the ASM</u> mode behavior at the <u>intra-domain</u> level
  - No PIM-SM in the ASM mode behavior at the inter-domain level

Co-located RP/PE – MVPN as a collection of (interconnected) multicast domains (1)



- Treat all MVPN sites connected to a given PE as a single (multicast) domain
- A given PE acts as an (MVPN) RP for all the sites of a given MVPN connected to that PE – *colocated RP/PE*
  - Distinct RP instance per directly connected MVPN on the PE
- MVPN = set of multicast domains interconnected by the MVPN Service Provider(s) infrastructure
- Plain PIM-SM in the ASM mode procedures for all the sites of a given MVPN connected to a given PE, including CE-PE interaction
  - Plain PIM-SM in the ASM mode within each (multicast) domain
- But what about inter-site (inter-domain) procedures among the MVPN sites connected to different PEs ?
  - See the following slide…

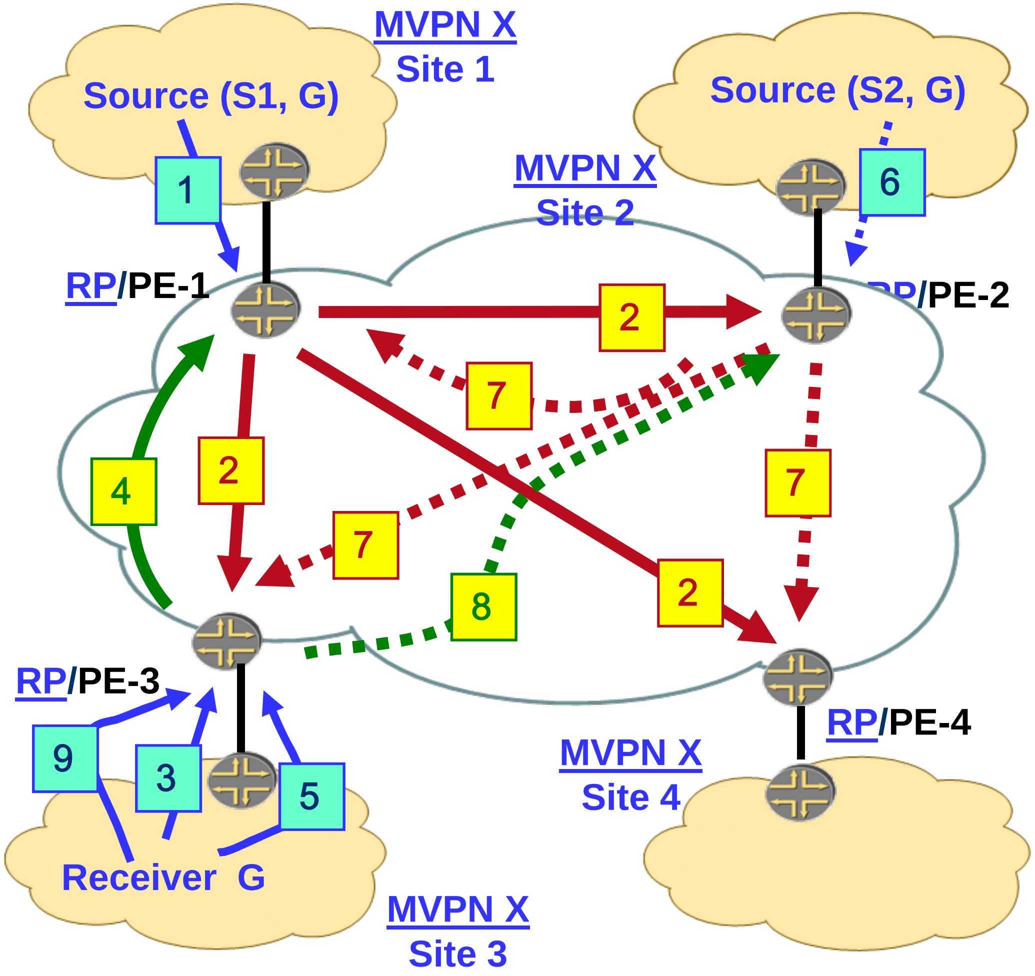
Co-located RP/PE – MVPN as a collection of (interconnected) multicast domains (2)

- **Use BGP to exchange information about active (multicast) sources among all the (multicast) domains that form a given MVPN**
  - **Among all the sites of that MVPN**
  - **Exchange of information involves only the PEs connected to the sites of that MVPN**
    - **PEs act as RPs of the domains that form that MVPN**
  - **By using** BGP Source Active auto-discovery routes
- **Use BGP procedures for supporting PIM-SM in SSM mode MVPNs to exchange (S, G) state among the domains that form a given MVPN**
  - **See previous slides for more on this…**

BGP Source Active auto-discovery routes

- Source Active auto-discovery routes are carried in Multiprotocol BGP (RFC4760) using MCAST-VPN NLRI
  - The same NLRI as used by C-multicast routes and auto-discovery routes
- Source Active auto-discovery route NLRI contains:
  - Multicast Source (S), Multicast Group (G)
  - Route Distinguisher (RD)
    - Needed to support MVPNs that may use the same address space (just like with unicast)
- Route Targets used to constrain distribution of Source Active auto-discovery routes of a given MVPN are the same as the MVPN import/export Route Target
  - MVPN import/export Route Targets can be the same as Route Targets used by unicast 2547 VPNs
- Re-use the existing BGP mechanisms (e.g., extended communities, Route Target constraint, Route Reflectors, etc…)

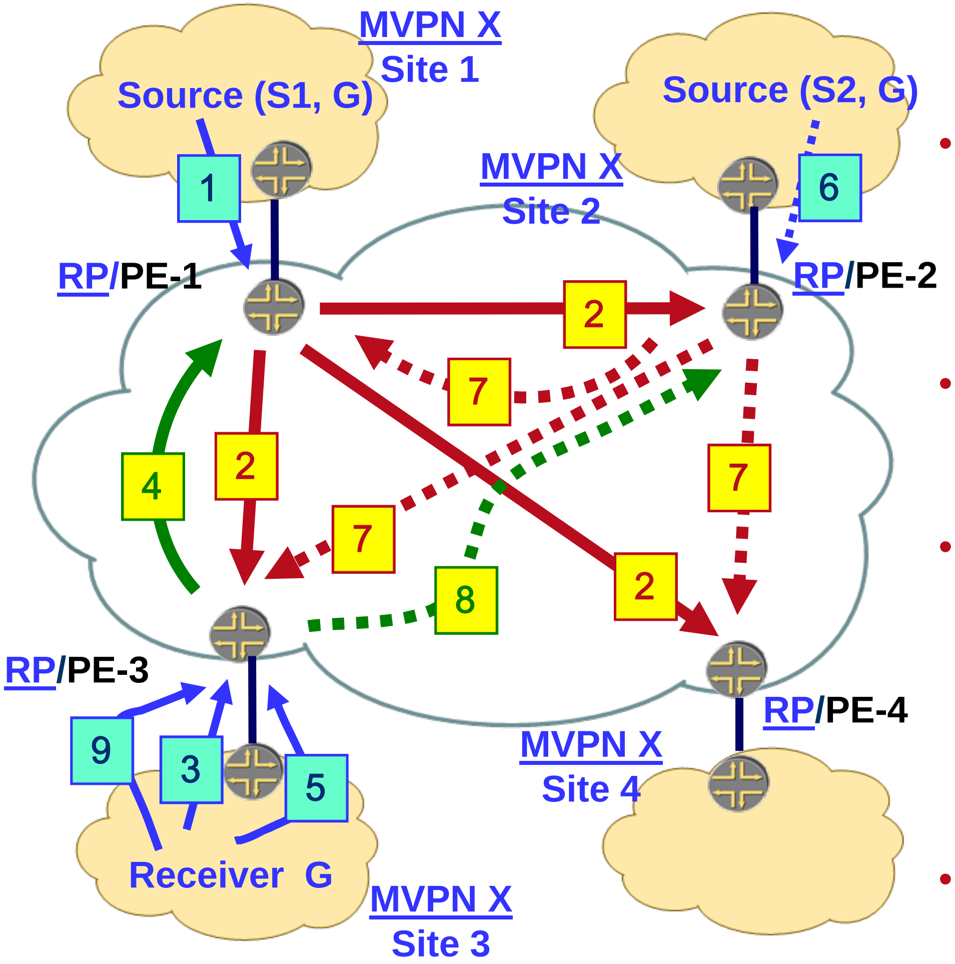
Co-located RP/PE: example



1. PIM Register (S1, G)
2. Source Active auto-discovery (S1, G)
3. PIM Join (*, G)
4. C-Multicast (S1, G)
5. PIM Join (S1, G)

6. PIM Register (S2, G)
7. Source Active auto-discovery (S2, G)
8. C-Multicast (S2, G)
9. PIM Join (S2, G)

Co-located RP/PE: Summary (1)



- Each RP/PE of a given MVPN discovers active sources (S,G) within the MVPN sites (directly) connected to the RP/PE by using plain PIM
  - (1) PIM Register (S1, G), (6) PIM Register (S2, G)
- RP/PE connected to the site that contains an active source (S,G) informs all other PEs that have sites of that MVPN about the active source (S,G) by originating Source Active auto-discovery route for the active source
  - (2) Source Active (S1, G), (7) Source Active (S2, G)
- Receivers within each MVPN site use plain PIM to inform the PE connected to the site that they want to receive traffic for a particular group G
  - (3) PIM Join (*, G)
- When a PE receives Join(*,G) from one of its directly connected CEs, the PE converts it into C-multicast routes, one per each received Source Active auto-discovery route that has G
  - (4) C-Multicast route (S1, G), (8) C-Multicast route (S2, G)
  - Informs the PE connected to the active source (S,G) that there are receivers for (S,G) connected to some other PEs
- When a receiver switches from Shared Tree (RPT) to Source Tree (SPT), this switch is localized to the site that contains the receiver and the PE connected to the site
  - (5) PIM Join (S1, G), (9) PIM Join (S2, G)

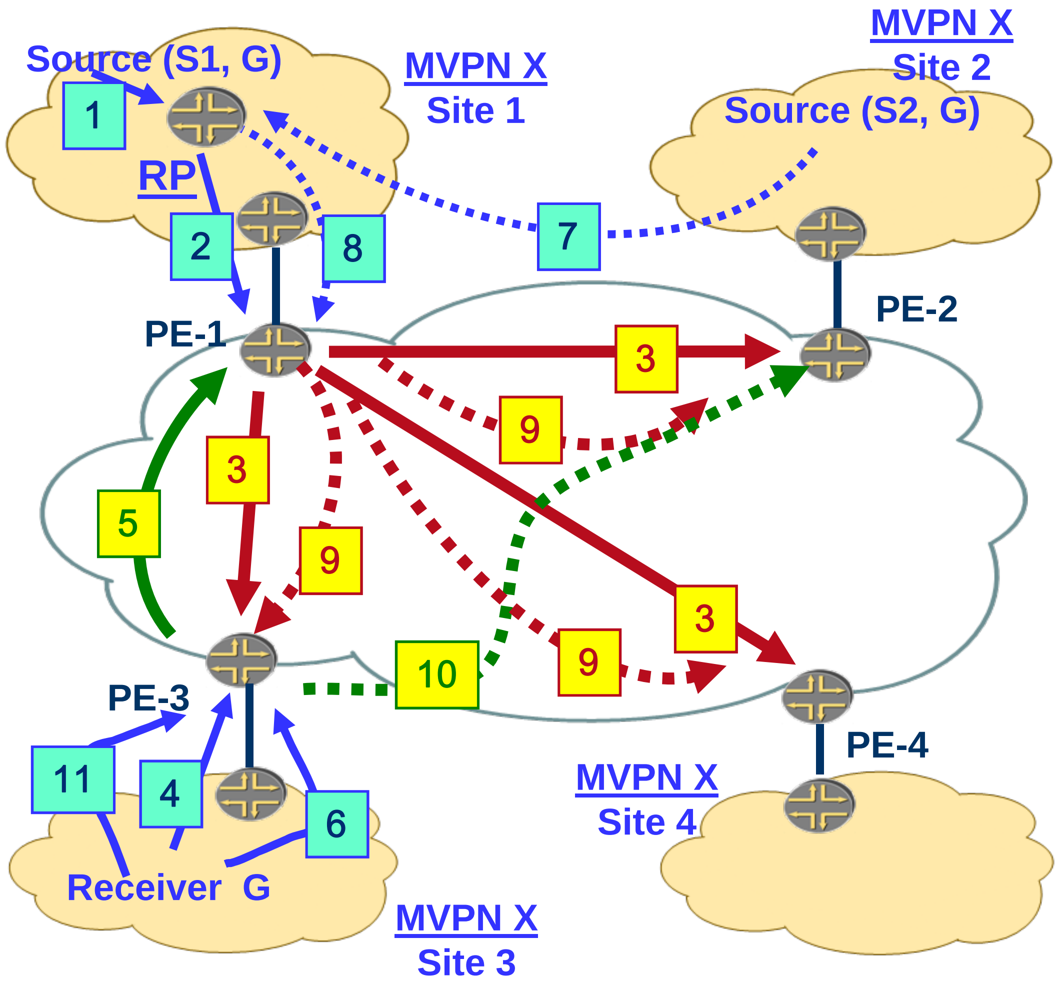
Co-located RP/PE: Summary (2)

- Requires every PE that has one or more sites of a given MVPN connected to it to act as an RP for that MVPN
- Suitable if an MVPN customer wants to outsource its RP infrastructure to the service provider
- Problematic if an MVPN customer wants to retain its own RP infrastructure
- See next slides for other options…

MSDP/PIM Register options

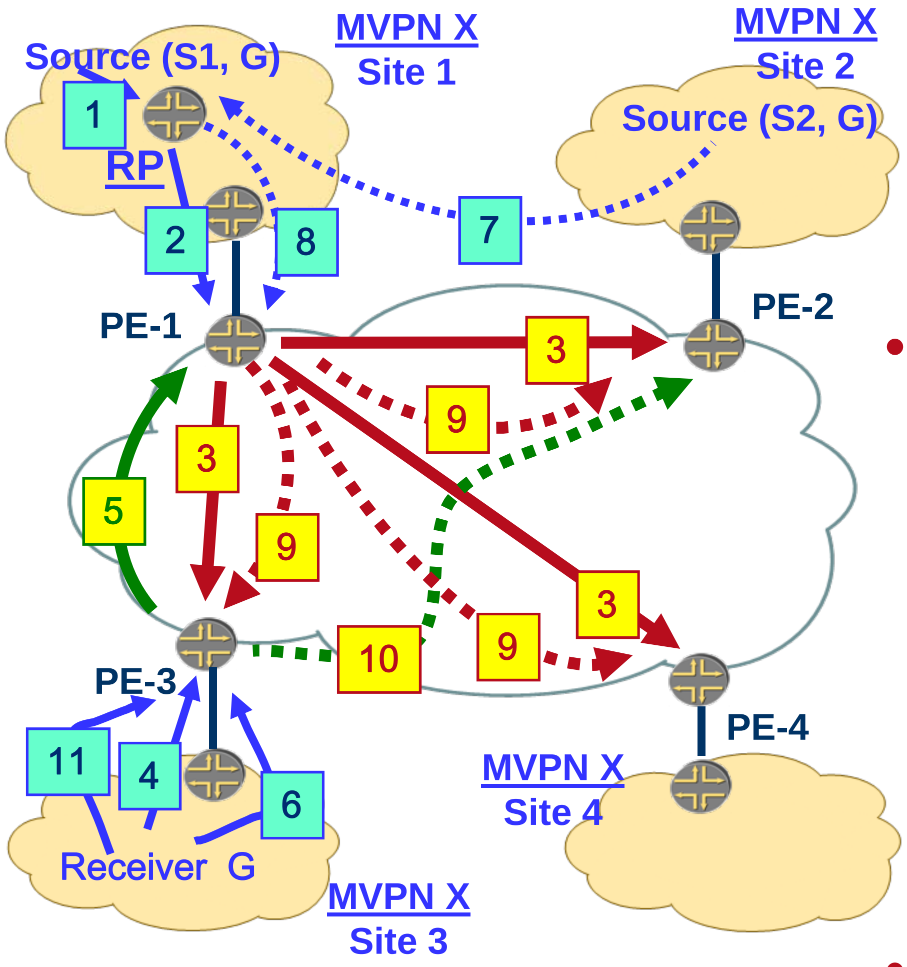
- MVPN customer maintains its own RP infrastructure
- Use the existing IP multicast mechanisms to communicate information about active sources (S,G) from MVPN RP(s) to one or more PEs:
  - MSDP Option: use MSDP between MVPN RPs and PEs, or
  - PIM Register Option: use PIM Register between MVPN RPs and PEs
  - These PEs maintain information about active sources (S,G) for a given MVPN
    - Just like with co-located RP/PE
  - These PEs do NOT act as MVPN RPs
    - PEs do NOT receive PIM Register from any of the MVPN Designated Routers (DRs)
    - With MSDP option MSDP Source Active (SA) advertisements flow only from RP to PE, but NOT from PE to RP
    - With PIM Register option PIM Registers flow only from RP to PE, but NOT from PE to RP
    - This is in contrast to co-located RP/PE
- The rest is the same as with co-located RP/PE

MSDP Option: example



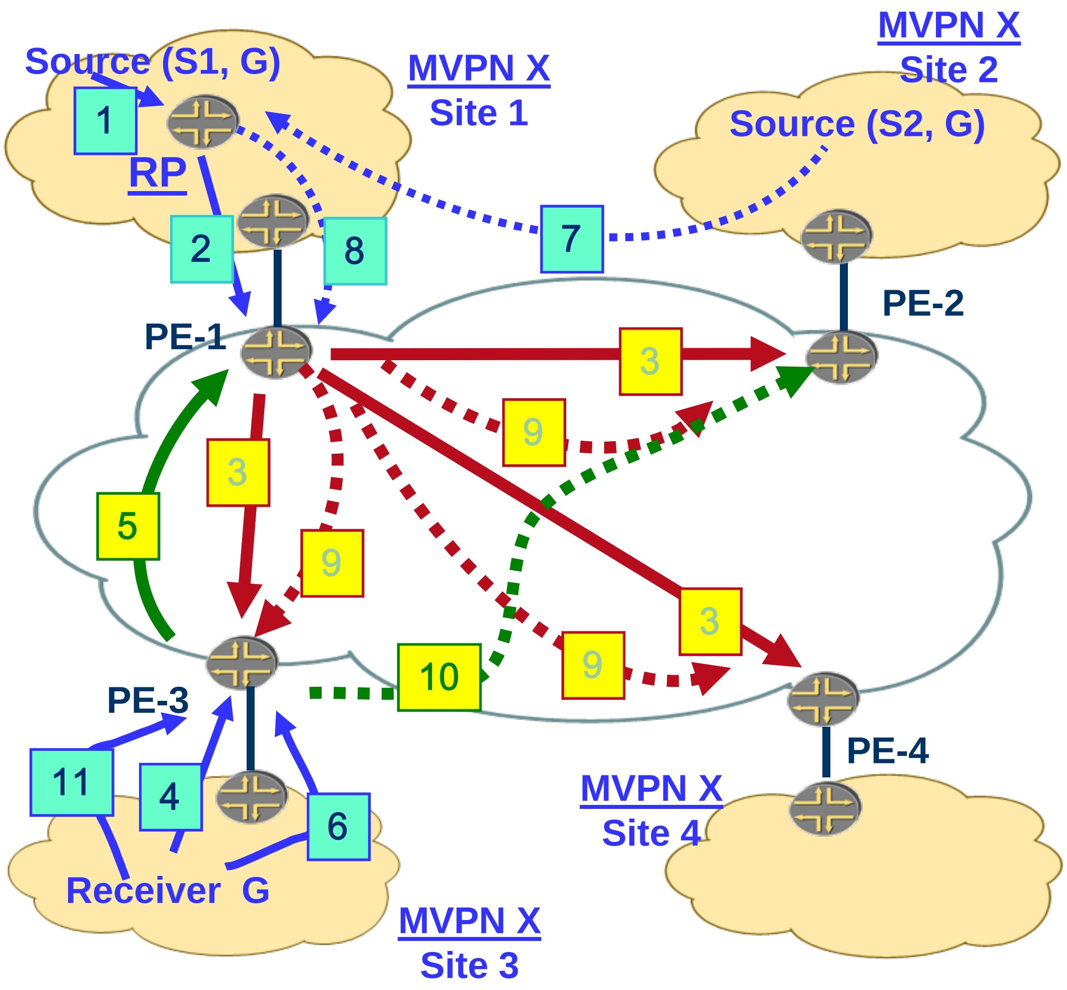
1. PIM Register (S1, G)
2. MSDP SA (S1, G)
3. Source Active auto-discovery (S1, G)
4. PIM Join (*, G)
5. C-Multicast (S1, G)
6. PIM Join (S1, G)
7. PIM Register (S2, G)
8. MSDP SA (S2, G)
9. Source Active auto-discovery (S2, G)
10. C-Multicast (S2, G)
11. PIM Join (S2, G)

MSDP Option: Summary



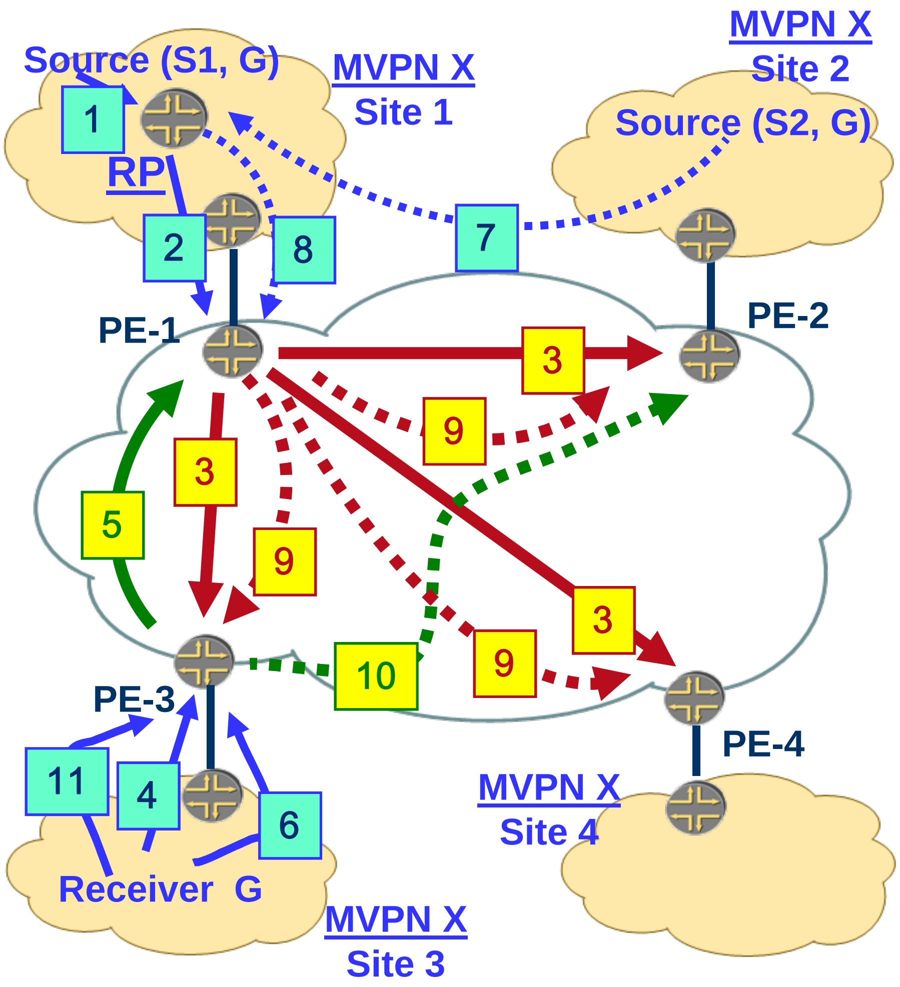
- RPs within each MVPN discover information about MVPN active sources (S,G) by using plain IP multicast mechanism (PIM Register messages)
  - Does not require any multicast within the service provider
  - (1) PIM Register (S1, G), (7) PIM Register (S2, G)
- PE obtains from an RP of a given MVPN the information about active sources (S,G) within that MVPN using plain IP multicast mechanisms - MSDP between PE and RP
  - Distinct MSDP instance on the PE per each distinct MVPN
  - MSDP Source Active (SA) advertisements flow from RP to PE, but NOT from PE to RP
  - (2) MSDP SA (S1, G), (8) MSDP SA (S2, G)
  - PEs do NOT maintain MSDP peering with each other
- Rest is the same as with Option 1

PIM Register Option: example



1. PIM Register (S1, G)
2. PIM Register (S1, G)
3. Source Active auto-discovery (S1, G)
4. PIM Join (*, G)
5. C-Multicast (S1, G)
6. PIM Join (S1, G)
7. PIM Register (S2, G)
8. PIM Register (S2, G)
9. Source Active auto-discovery (S2, G)
10. C-Multicast (S2, G)
11. PIM Join (S2, G)

PIM Register Option: Summary



- RPs within each MVPN discover information about MVPN active sources (S,G) by using plain IP multicast mechanisms (PIM Register messages)
  - Does not require any multicast within the service provider
  - (1) PIM Register (S1, G), (7) PIM Register (S2, G)
- PE obtains from an RP of a given MVPN the information about active sources (S,G) within that MVPN using plain IP multicast mechanisms – PIM Register between PE and RP
  - PIM Register messages flow from RP to PE, but NOT from PE to RP
  - (2) PIM Register (S1, G), (8) PIM Register (S2, G)
- The rest is the same as with Option 1

MSDP/PIM Register: Summary

- Works well if an MVPN customer wants to have full control over its own RP infrastructure
  - Supports Anycast RP in customer’s RP infrastructure
  - Supports BSR, Auto-RP in customer’s RP infrastructure
- Does this option make sense if a customer wants to completely outsource the RP infrastructure ?
  - NO, as this option assumes that none of the PEs act as an RP

Co-located RP, MSDP, PIM Register comparison

- Among MVPN sites the (multicast) traffic is ALWAYS carried over Shortest Path (SPT) trees
  - Inter-site (multicast) traffic never flows through customer’s RP
- All these options have exactly the same procedures for:
  - Originating and receiving BGP Source Active auto-discovery routes for active sources (S,G)
  - Originating and receiving BGP C-multicast routes that carry (S,G)
  - Handling PIM messages received by PEs from the directly connected CEs
- The main difference is whether a PE acts as a fully functional MVPN customer RP
  - Yes with co-located RP
  - No with MSDP/PIM Register

Agenda

- BGP/MPLS MVPN – what are the goals ?
- Supporting PIM-SM in SSM mode MVPNs
- Supporting PIM-SM in ASM mode MVPNs
  - Carrying MVPN multicast routing information
  - ⇨ Carrying MVPN multicast traffic
- Summary

Carrying MVPN multicast traffic

- Exactly the same as for PIM-SM in SSM mode MVPNs
  - No new procedures

Agenda

- BGP/MPLS MVPN – what are the goals ?
- Supporting PIM-SM in SSM mode MVPNs
- Supporting PIM-SM in ASM mode MVPNs
- ➡ Summary

BGP/MPLS MVPN – Summary

- Extends 2547 VPN service offering to include support for IP multicast for 2547 VPN customers
- Follows the same architecture/model as 2547 VPN unicast
  - Uniform control plane to support both unicast and multicast
  - Eliminates the need to have the Virtual Router (VR) model for multicast and the 2547 model for unicast
- Re-uses 2547 VPN unicast mechanisms: BGP, MPLS
  - With extensions, as necessary
  - Common set of mechanisms to support both unicast and multicast
- Retains as much as possible the flexibility and scalability of 2547 VPN unicast

Suggested reading

- RFC4834
- draft-ietf-l3vpn-ppvpn-mcast-reqts
- draft-ietf-l3vpn-2547bis-mcast
- draft-ietf-l3vpn-2547bis-mcast-bgp

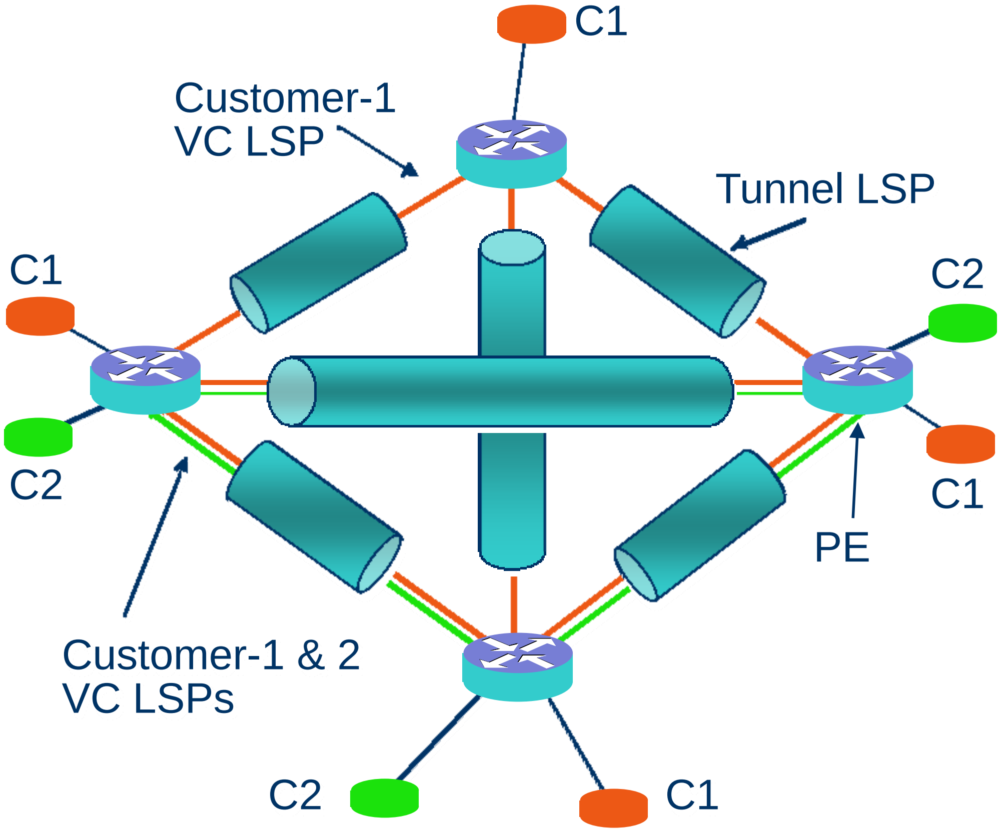
Section 3

Multicast in VPLS Networks

Agenda

- Introduction to VPLS and H-VPLS
- Requirements and solution aspects
  - Optimizing multicast transport
  - Resiliency
  - Assured QoE
- VPLS and P2MP MPLS
- Conclusions

Introduction to VPLS

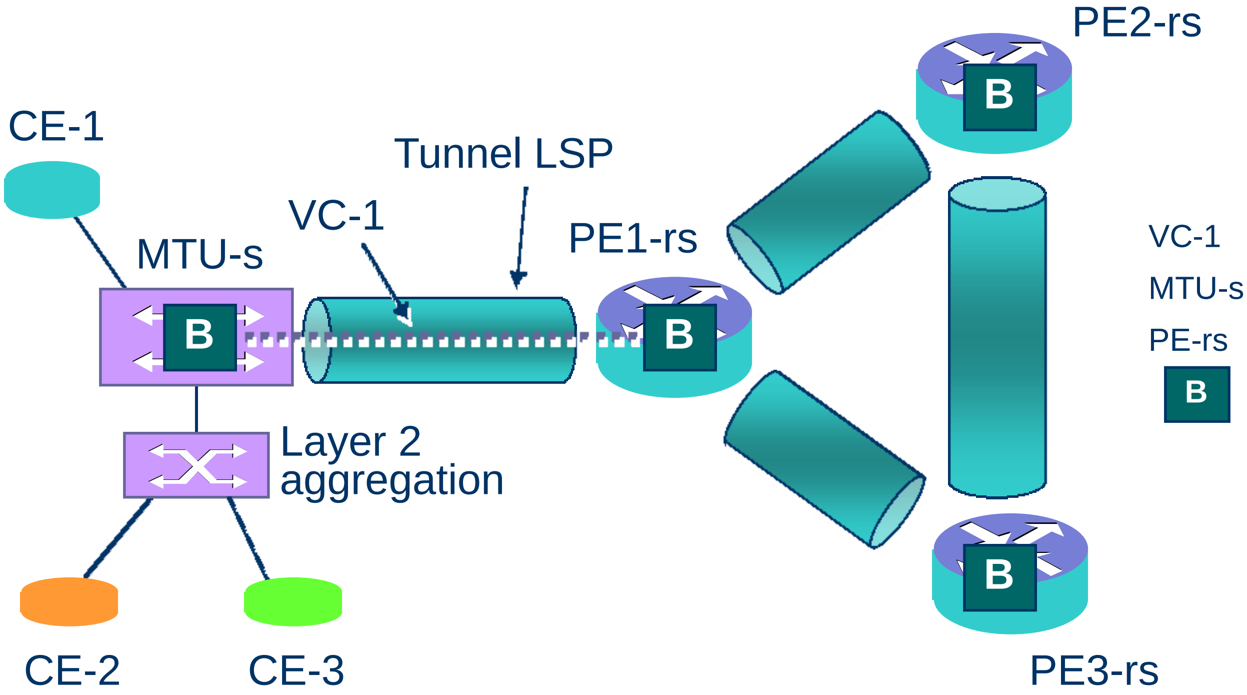


Transparent L2 VPN for Ethernet

- **Learns MAC addresses per PW**
- **Forwarding based on MAC addresses**
- **Replicates multicast & broadcast frames**
- **Floods unknown frames**
- **Split-horizon for loop prevention**

H-VPLS Architecture

- Uses PWs / LSPs between edge MTU and VPLS aware PE devices
- Reduces signaling and packet replication to allow large scale deployment



VC-1	= Single PW
MTU-s	= Bridging Capable Multi-Tenant Unit
PE-rs	= VPLS Capable PE
B	= Virtual VPLS (Bridge) Instance

VPLS and Metro Ethernet

- Metro Ethernet Network plays pivotal role in next generation services
  - Converged traffic: all residential & growing enterprise traffic
  - Subscriber aware: interface with multiple access networks
  - Service aware: connect to multiple service back-ends
- Ethernet famous for flexible, cost-effective bandwidth … but:
  - Not optimized for reliable, efficient IP multicast (IPTV)
  - Inherent gaps in security, OAM&P and QoS
- VPLS leverages:
  - MPLS
  - Multicast for triple play and enterprise applications

Requirements for Multicast in VPLS
draft-ietf-l2vpn-vpls-mcast-reqts-03.txt

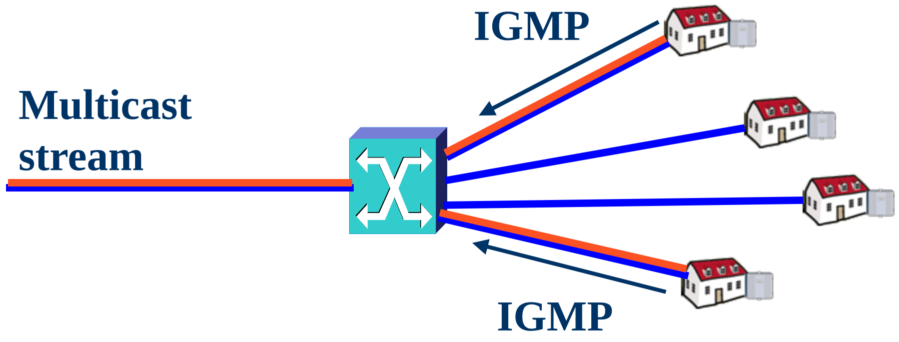
- By default, VPLS will replicate mcast traffic on ingress
  - To non-member site
  - Duplication on PWs sharing a path
- Multicast VPLS resolves both these issues
  - Reduce wasted bandwidth
  - Keep core P- routers stateless
  - Trim multicast tree to only include group members
    - Only replicate to sites with mcast receivers

VPLS Multicast Solution Aspects

- **IGMP-Snooping**
  - This tutorial shows how this applies for triple-play
  - draft-ietf-magma-snoop-12.txt
- **PIM-Snooping in VPLS**
  - This tutorial shows how this applies for enterprise applications
  - draft-ietf-l2vpn-vpls-pim-snooping-00.txt
- **PE-to-PE Mcast State Distribution**
  - draft-qiu-serbest-l2vpn-vpls-mcast-ldp-00.txt
- **Mcast Trees in Provider Core**
  - draft-raggarwa-l2vpn-vpls-mcast

What is snooping?

- Snooping switches use information in upper protocol layers to determine processing at lower layers
- IGMP/PIM snooping switch:
  - Inspects IGMP/PIM messages
  - Changes layer 2 forwarding behaviour accordingly



Conserve bandwidth on network links

Avoid sending multiple copies to nodes

Avoid forwarding to nodes not interested in receiving mcast packets from group

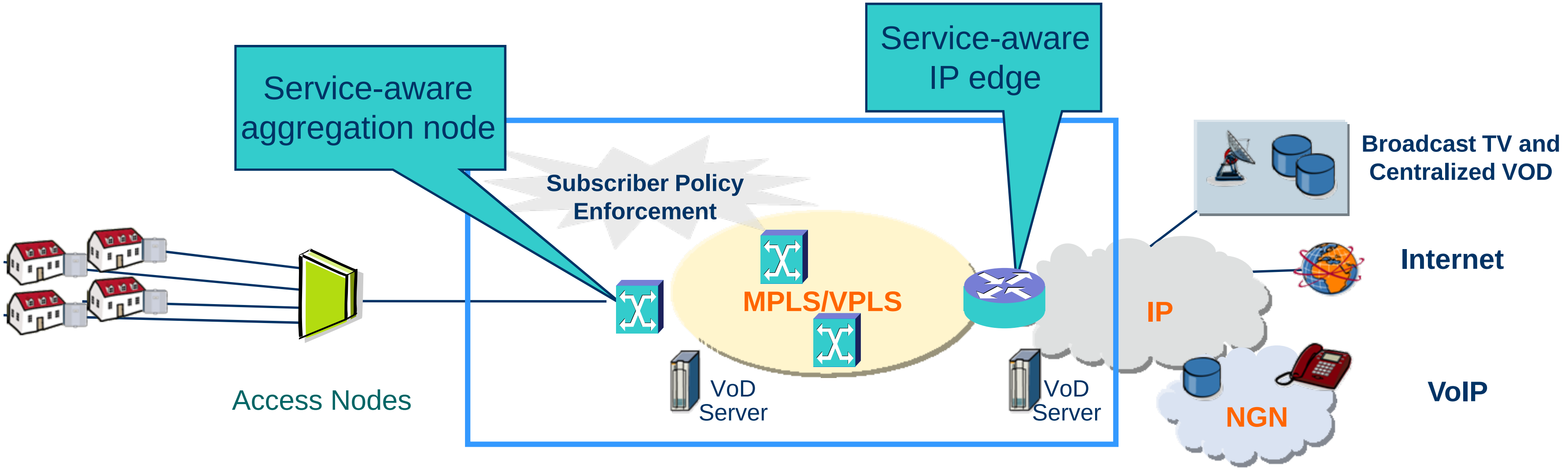
Multicast in VPLS for IPTV

- **Used when more than one person requests the** SAME PROGRAM **at the** SAME TIME
  - **Broadcast TV**
  - **Near VoD (same program starts at regular intervals)**

- Goal:
  - **Reduce bandwidth usage by** sharing a single copy **of a flow** as far as possible
  - **Optimize for different physical topologies**

- Why use layer 2 aggregation?
  - **Channels directly available at layer 2 (IGMP snooping)**
  - **No need to signal "join" and "leave" up to the source**
  - **Share resources to handle multicast**
  - **Decrease zapping time**

Optimizing VPLS for IPTV/Video
Key Requirements


Service-aware aggregation node
Service-aware IP edge
Subscriber Policy Enforcement
MPLS/VPLS
IP
Broadcast TV and Centralized VOD
Internet
VoIP
NGN
Access Nodes
VoD Server
VoD Server

Traffic engineering, QoS, security, carrier OAM

- **VPLS: combines strengths of Ethernet and MPLS**
- **Service separation: unicast/mcast VPLS instances**
- **QoS: per-subscriber and per-service flow control**
- **Security: residential split horizon, anti-spoofing, …**

Multicast optimization and flexibility

- **VPLS multicast registration (IGMP proxy/snooping)**
- **H-VPLS: optimize rings and/or mesh topologies**
- **Distributed multicast and content insertion**

Assuring the user experience

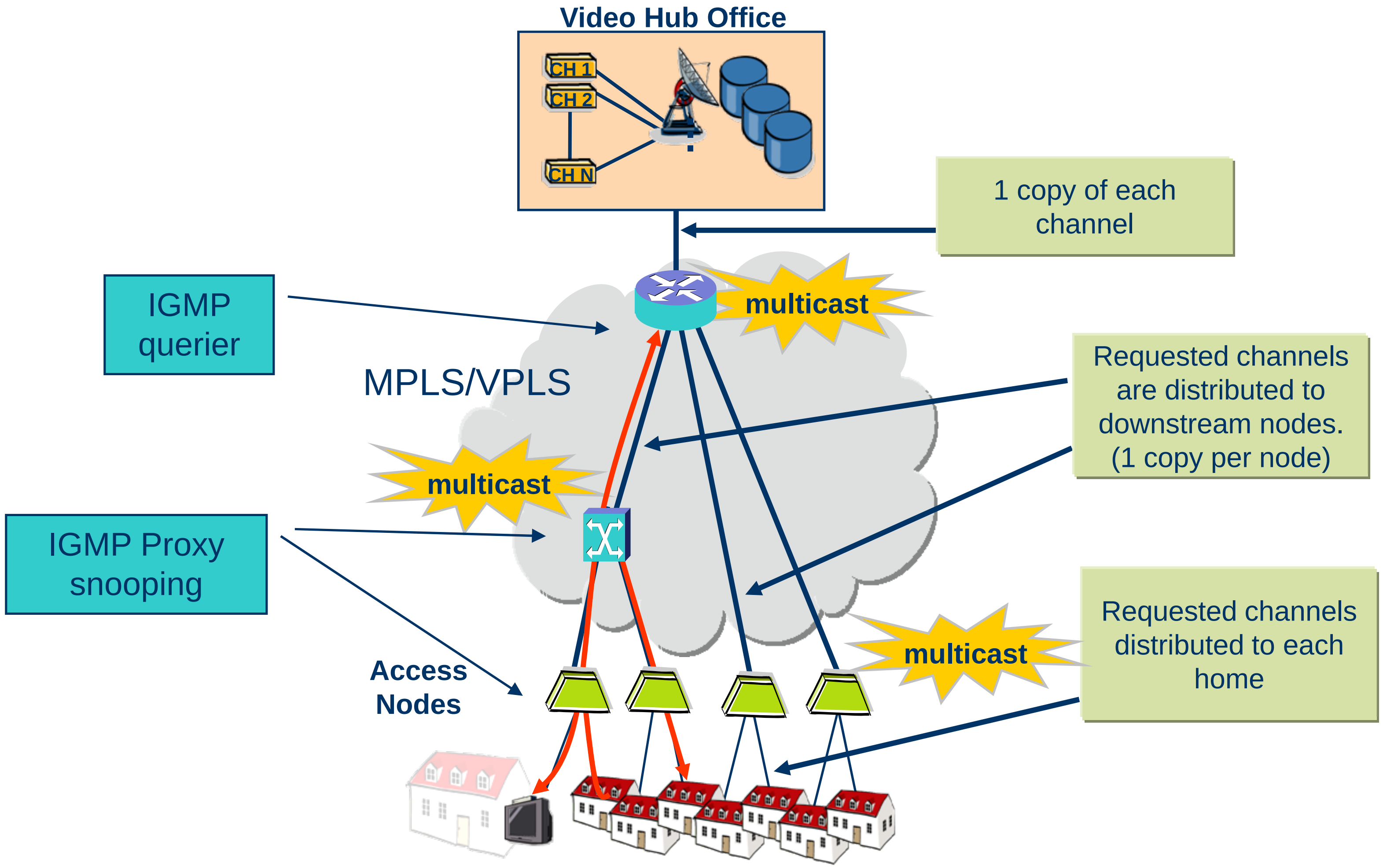
- **IGMP performance, HQoS, ICC***
- **Service admission control**
- **QoE and performance monitoring**
- **End-to-end management**

High-availability, non stop services

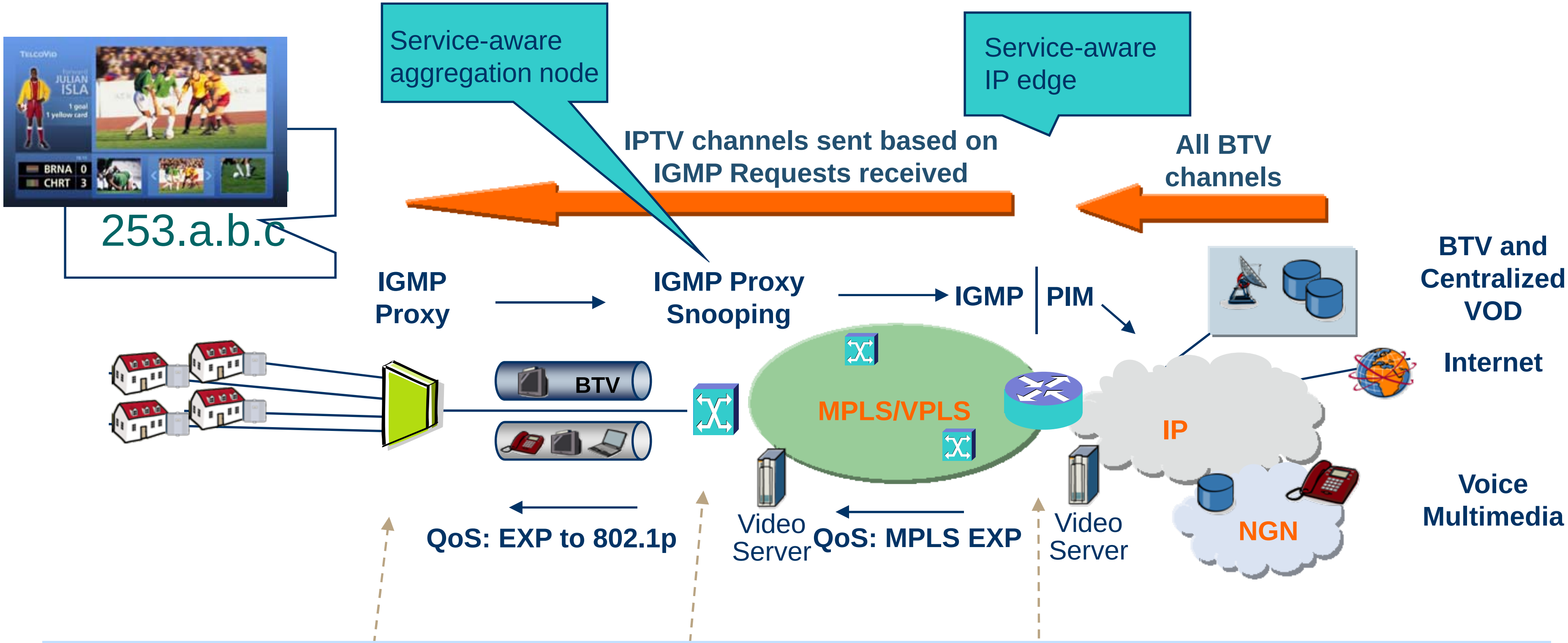
- **Rapid restoration (MPLS FRR)**
- **Source redundancy (PIM-BFD)**
- **Reliable nodes (NSR, ISSU)***

* ICC: Instant Channel Change
NSR: Non-stop Routing
ISSU: In service software upgrade

Scaling Multicast in VPLS (IGMP Snooping)

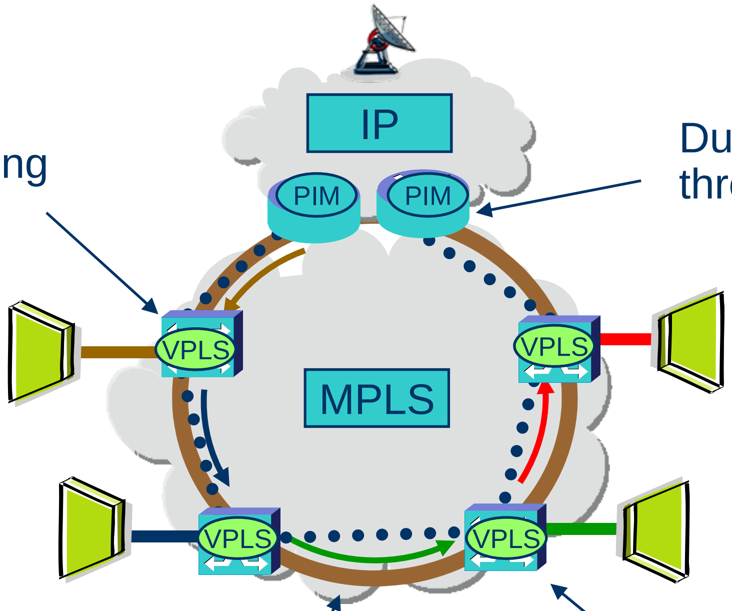


Multicast VPLS Registration (IGMP Snooping)



BTV: End-to-end multicast support in VPLS – adapts traffic replication and dynamically optimizes multicast mechanisms for actual viewing patterns

Multicast using H-VPLS rings



H-VPLS ring

Dual source redundancy through sources or PIM

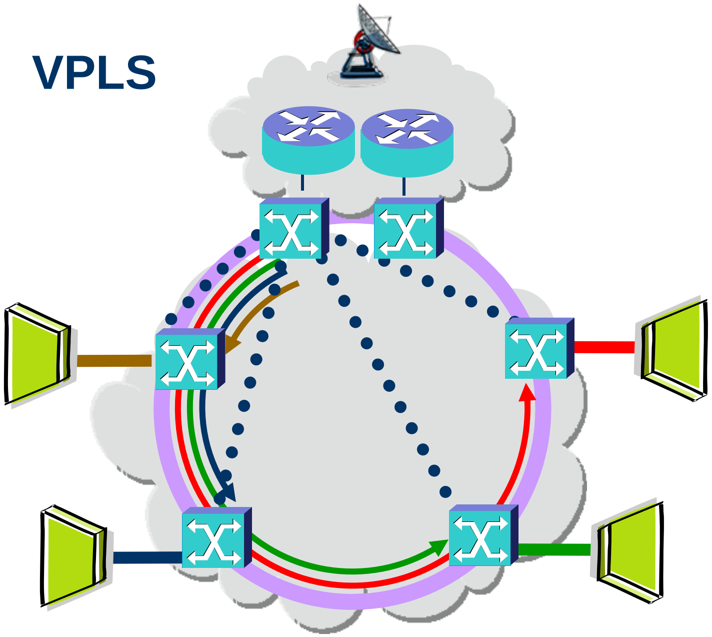
Efficient distributed replication in VPLS aggregation nodes

MPLS tunnels provide fast convergence and high availability

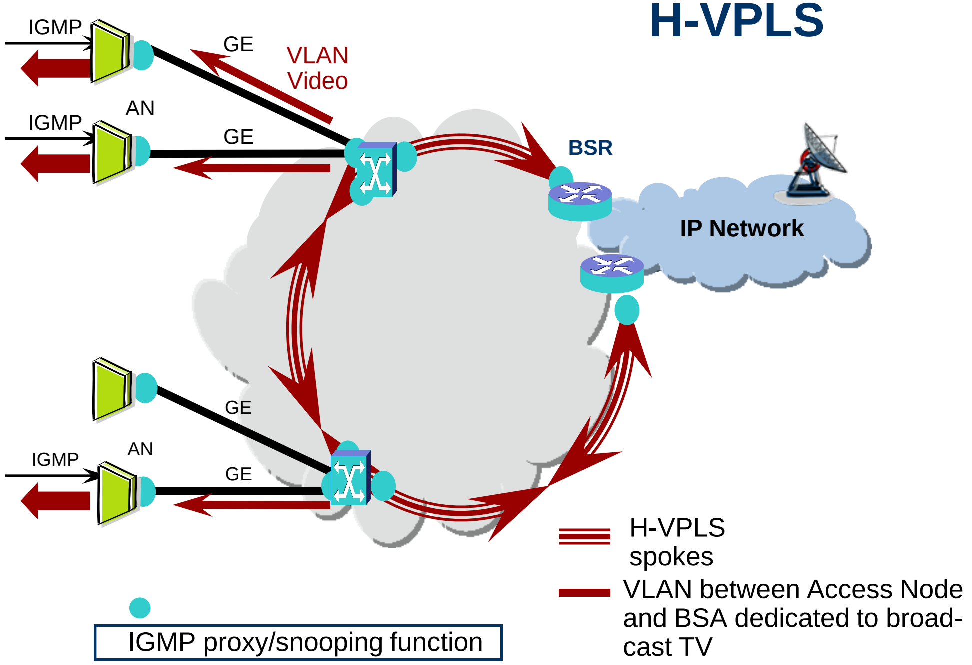
Comparing VPLS-based Multicast in Ring Topologies

- **By default, VPLS builds full mesh of connections**
- **Multiple copies of multicast flows sent on initial segments of the ring**

- **In H-VPLS, only a single copy travels the ring**
- **IGMP snooping used to replicate joined streams to the drop sites**



VPLS



H-VPLS

IGMP

GE

VLAN
Video

AN

IGMP

GE

BSR

IP Network

GE

AN

IGMP

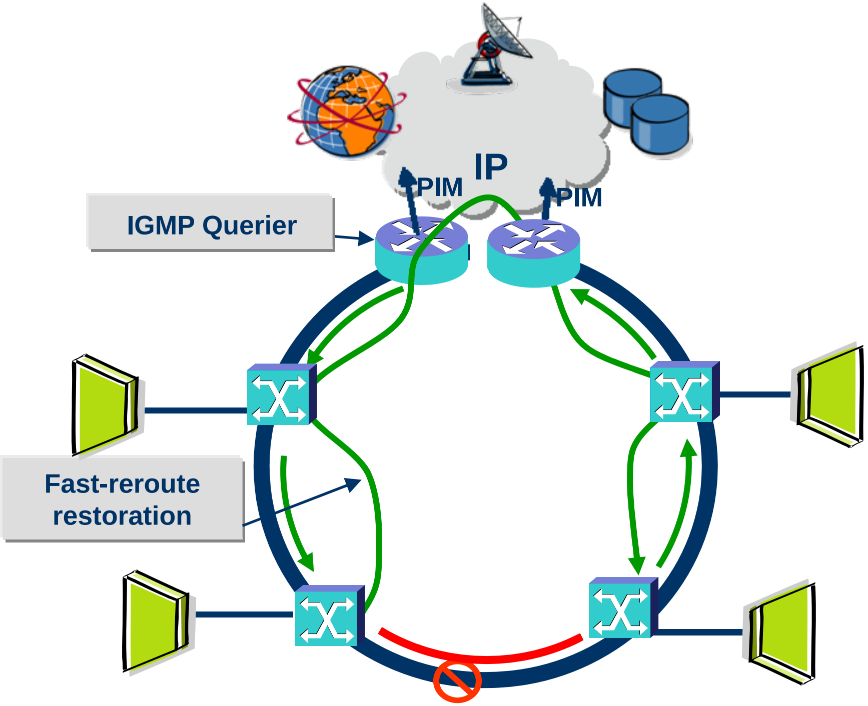
GE

H-VPLS
spokes

VLAN between Access Node and BSA dedicated to broad-cast TV

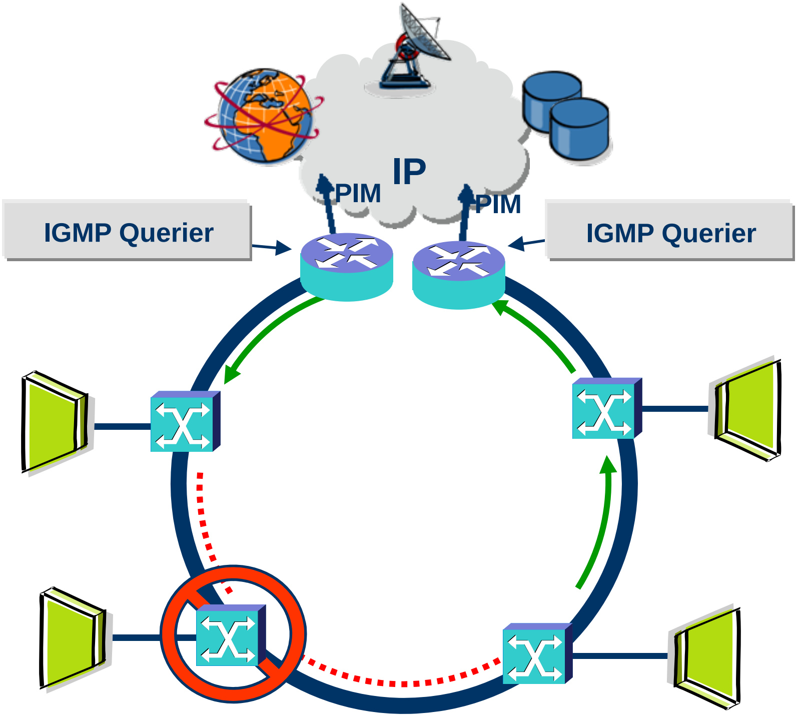
IGMP proxy/snooping function

Multicast Service Availability
Restoration from Link or Node Failures



Inter-aggregation node link failure:

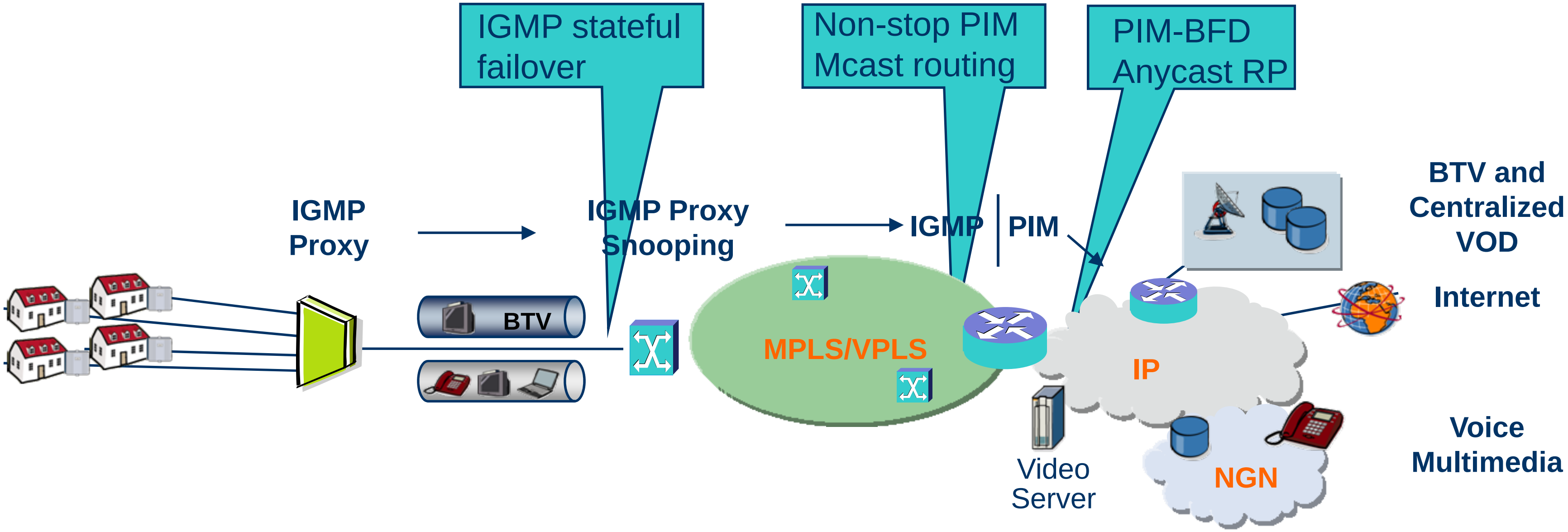
- MPLS FRR



Aggregation node failure:

- MPLS FRR available
- Both multicast routers become active
- Recovery via other half of ring

Multicast Service Availability
Protocol Resilience/Recovery



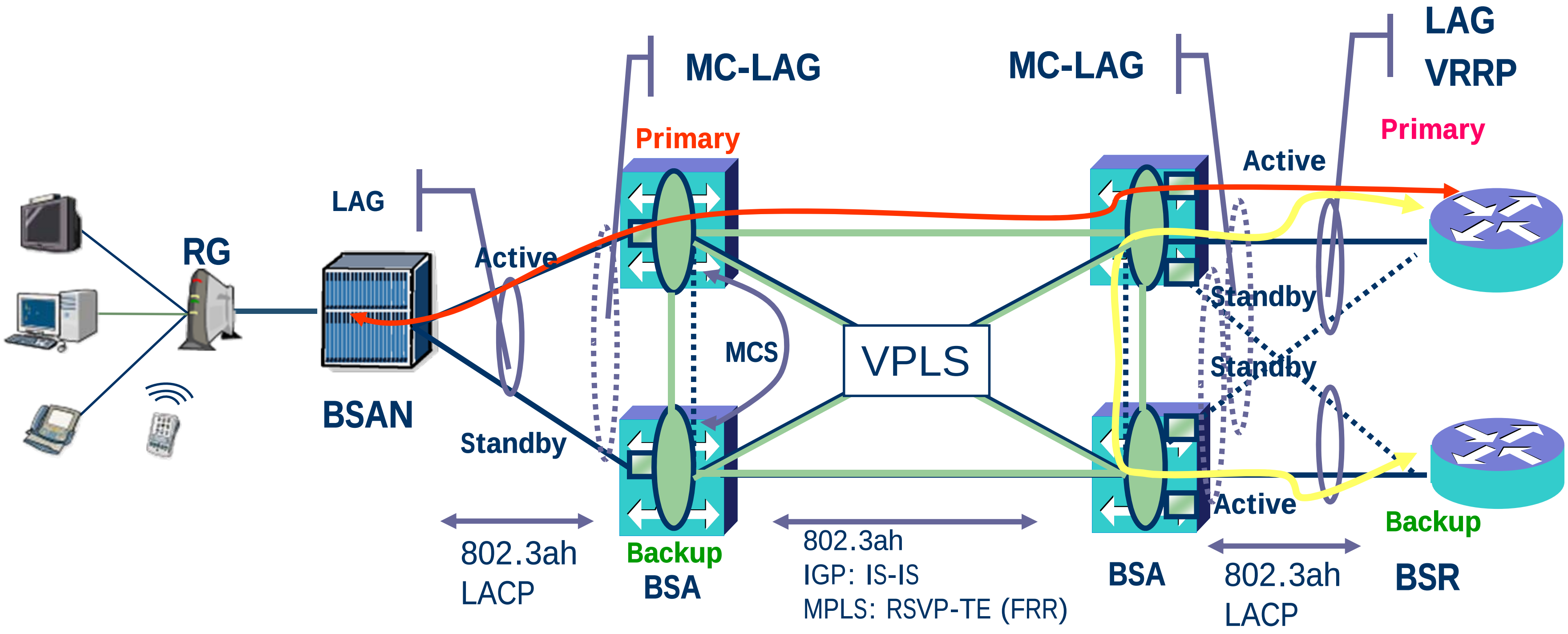
Enabling uninterrupted viewing:

- **IGMP stateful switchover:** preserves BTV channel forwarding state if a CPM fails
- **Non-stop multicast routing:** preserves PIM routing tables if a CPM fails
- **PIM-BFD:** fast detection of upstream PIM router failure
- **Anycast RP (RFC 3446):** fast convergence when PIM/MSDP rendezvous Point (RP) fails by allowing receivers and sources to Rendezvous at the closest RP

Multicast Service Availability
Node/Link Redundancy

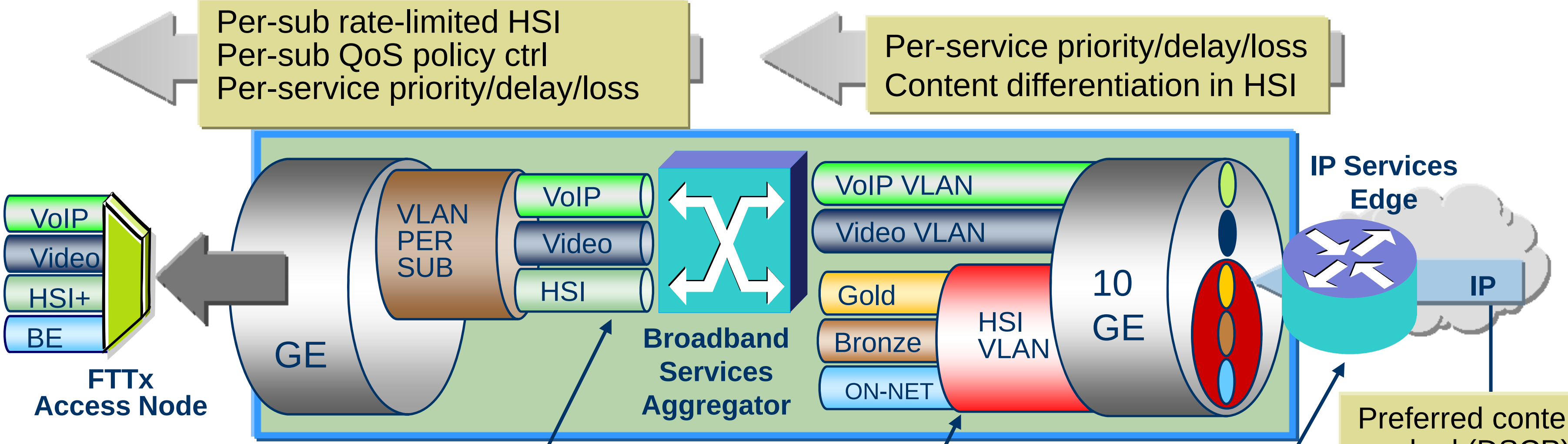
Provide resilience against complete node or link failure

Automatic failover with preservation of subscriber and service state (non-stop services)



MC-LAG: Multi-chassis Link Aggregation Group

Delivering Quality of Service for mcast: Hierarchical QoS



Per-sub rate-limited HSI
Per-sub QoS policy ctrl
Per-service priority/delay/loss

Per-service priority/delay/loss
Content differentiation in HSI

Per-sub queuing & PIR/CIR policing/shaping for HSI. HSI service classified on Source IP range

Per-service prioritization for VoIP and video. VoIP prioritized over video. Destination IP and/or DSCP classification

802.1p marking for prioritization in the access and home

VoIP and video is queued and prioritized as per VLAN QoS policy

HSI content differentiation based on DSCP. Each queue may have individual CIR/PIR and shaping

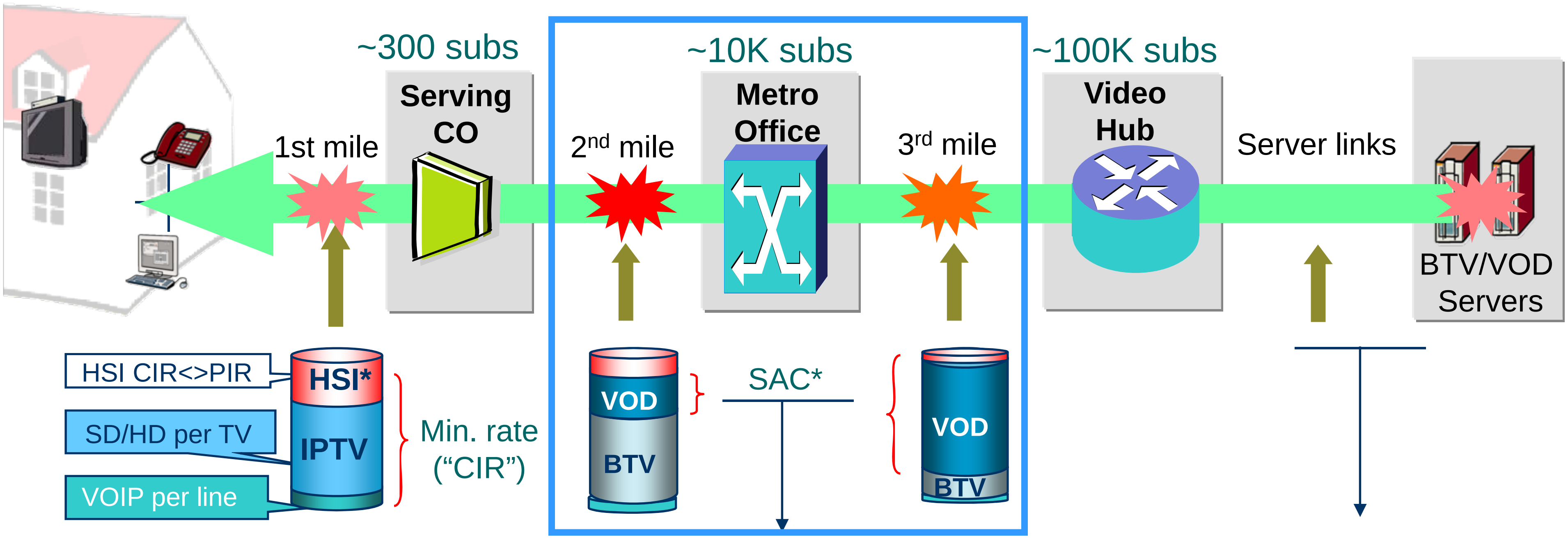
Optional overall subscriber rate limiting on VLAN (H-QoS)

Preferred content marked (DSCP) at trusted ingress points of IP network

For HSI content differentiation there is queuing for Gold, Silver, Bronze based on DSCP classification

Optional overall subscriber rate limiting on VLAN

Bandwidth Management / Admission Control - Potential resource bottlenecks



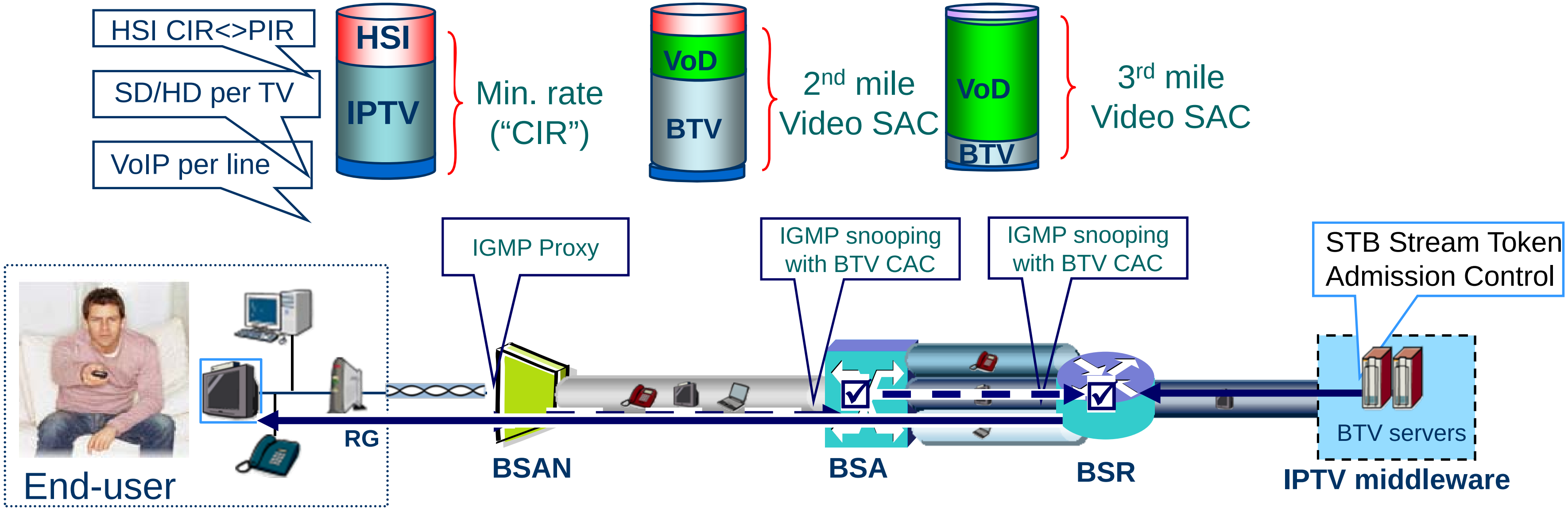
"Aggregate CIR bandwidth of subscribed services shall not exceed access capacity"

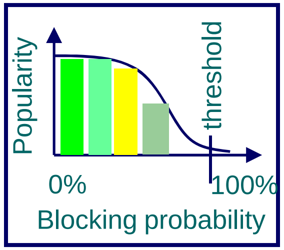
"Concurrent video bandwidth in use shall not exceed 2nd or 3rd mile bandwidth capacity budget"

Server link capacity must match server streaming capacity

* HSI: High Speed Internet
SAC: Session Admission Control

Multicast Admission Control





BTV Channel Admission Ctrl

- Triggered by STB, IGMP join
- High scalability, low latency
- On-path CAC enforcement
- Prioritize mandatory/popular channels and bundles

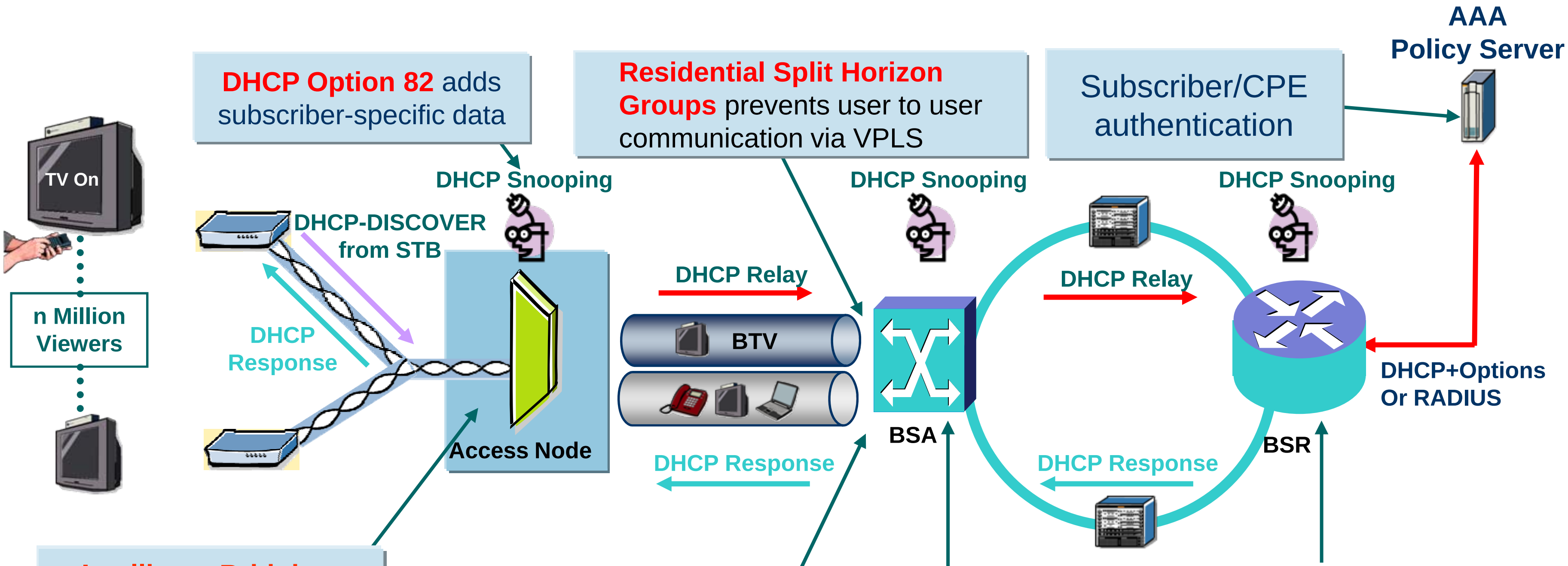
Multicast Grp	Channel b/w	Channel type	Bundle
224.1.1.1	2.5 Mbps	Mandatory	1
224.1.1.2	2.0 Mbps	Mandatory	1
224.1.1.3	2.5 Mbps	Optional	1
224.1.1.4	2.5 Mbps	Optional	1

Network Security Aspects
In addition to Content DRM

DHCP Option 82 adds subscriber-specific data

Residential Split Horizon Groups prevents user to user communication via VPLS

Subscriber/CPE authentication



Intelligent Bridging prevents user to user communication

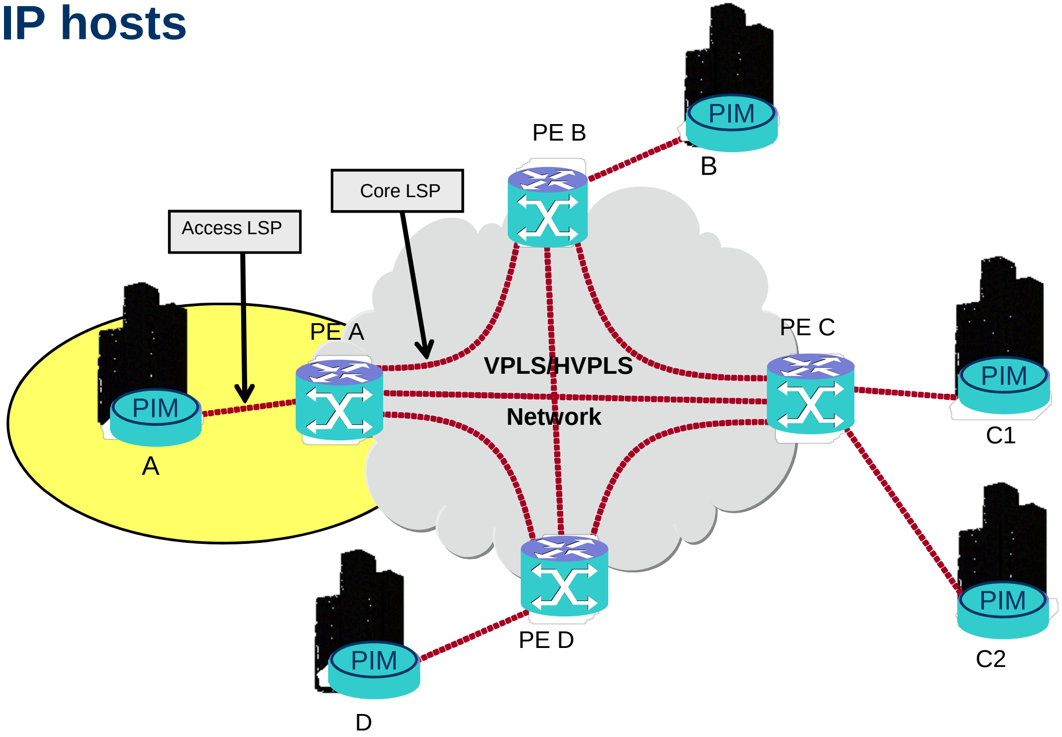
Proxy ARP enables controlled customer to customer interaction (e.g., for Internet only)

Periodic subscriber host connectivity verification

Auto-filter	Automatic anti-spoofing security filters to prevent service theft and DoS
attacks	Block user-to-user bridging
	Block ARP broadcasts & ICMP redirects
Rate-limit	User initiated broadcasts
Antispoofing	DHCP-configured ARP table

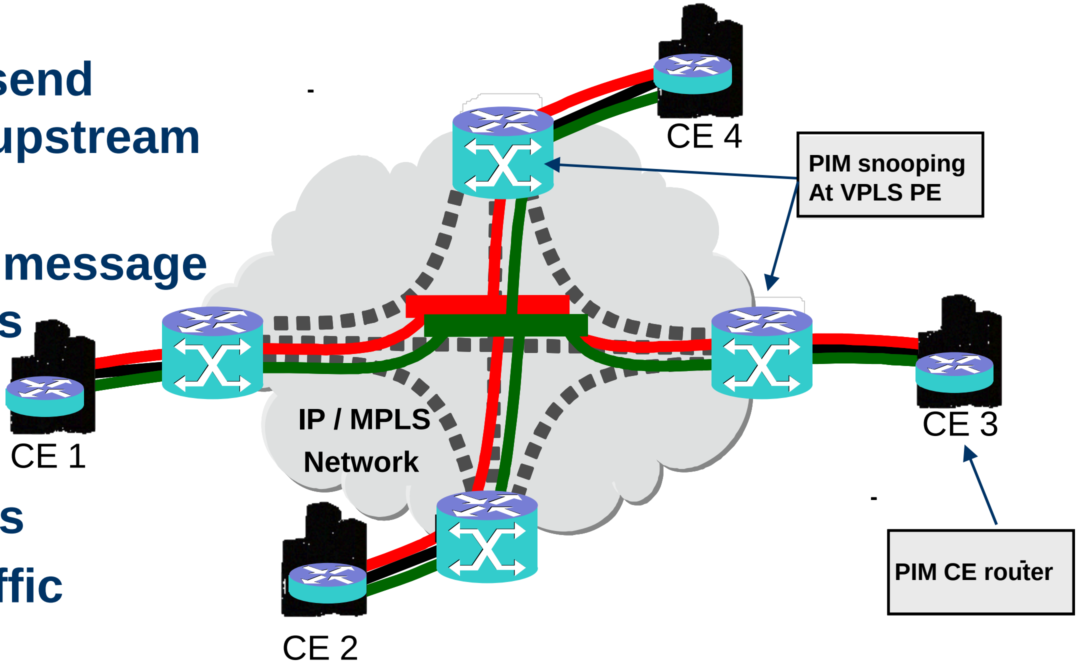
Multicast Issues for Enterprise VPNs

- IGMP snooping optimizes for a Layer 2 aggregation network between IGMP host and PIM-capable multicast PE router
- PIM snooping applies when using VPLS as a transparent VPN service to interconnect PIM-capable enterprise CE-router peers
- But, if a CE wants to multicast IP packets to subset of VPLS sites, all IP hosts will receive it

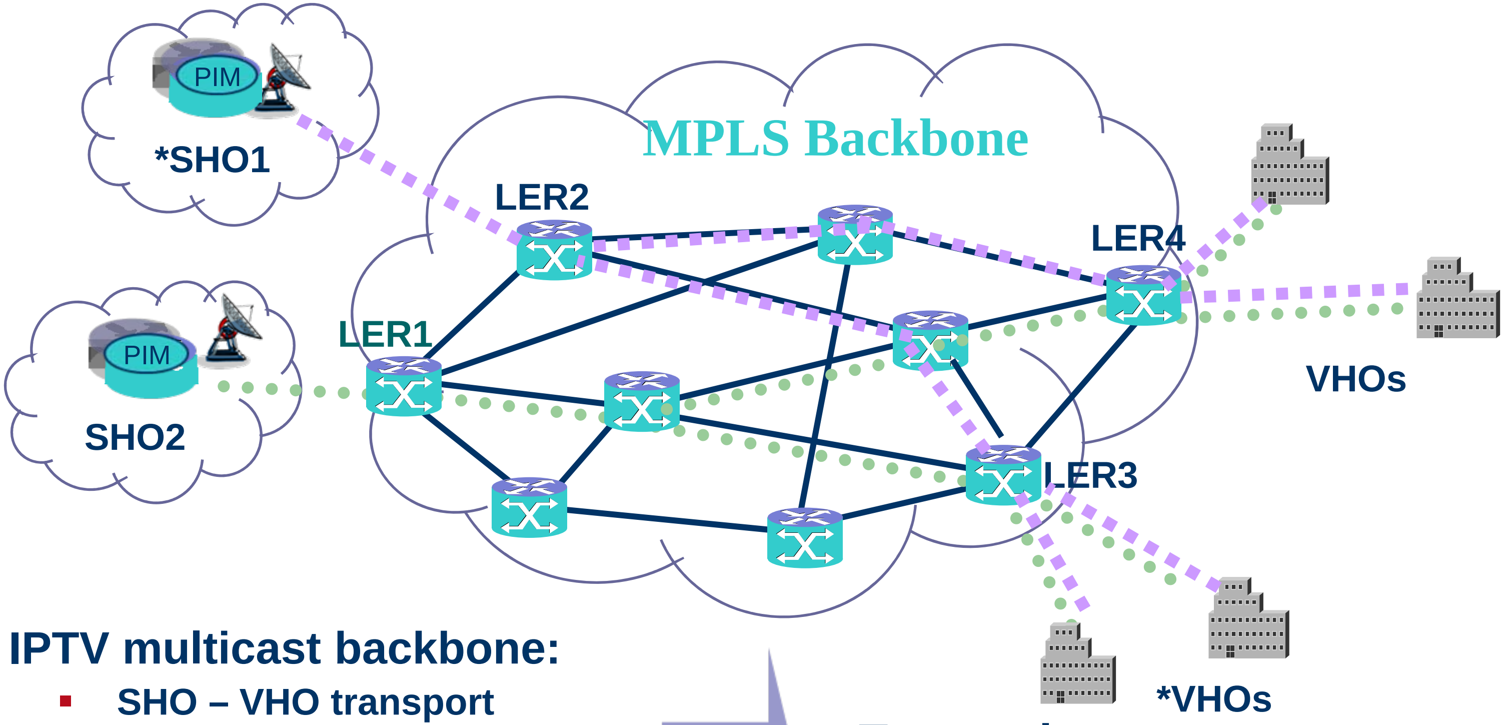


PIM Snooping in VPLS

- PIM snooping defined for PIM-SM/SSM and bidirectional PIM (BIDIR-PIM)
  - draft-ietf-l2vpn-vpls-pim-snooping
- PIM routers periodically exchange hello messages to discover neighbours & maintain session state
- PIM routers can then signal their intention to join/ prune specific multicast groups
  - Downstream routers send explicit join/prune to upstream routers
- VPLS PE snoops the PIM message exchange between routers
  - PIM join/prunes flooded in VPLS
  - Builds multicast states
- Forwards IP multicast traffic accordingly



Mcast trees in the provider core



- **IPTV multicast backbone:**
  - **SHO – VHO transport**
  - **Static multicast topology**
  - **100s of MPEG2/4 channels**
- **1+1 protection for all streams**
- **QoS and TE**

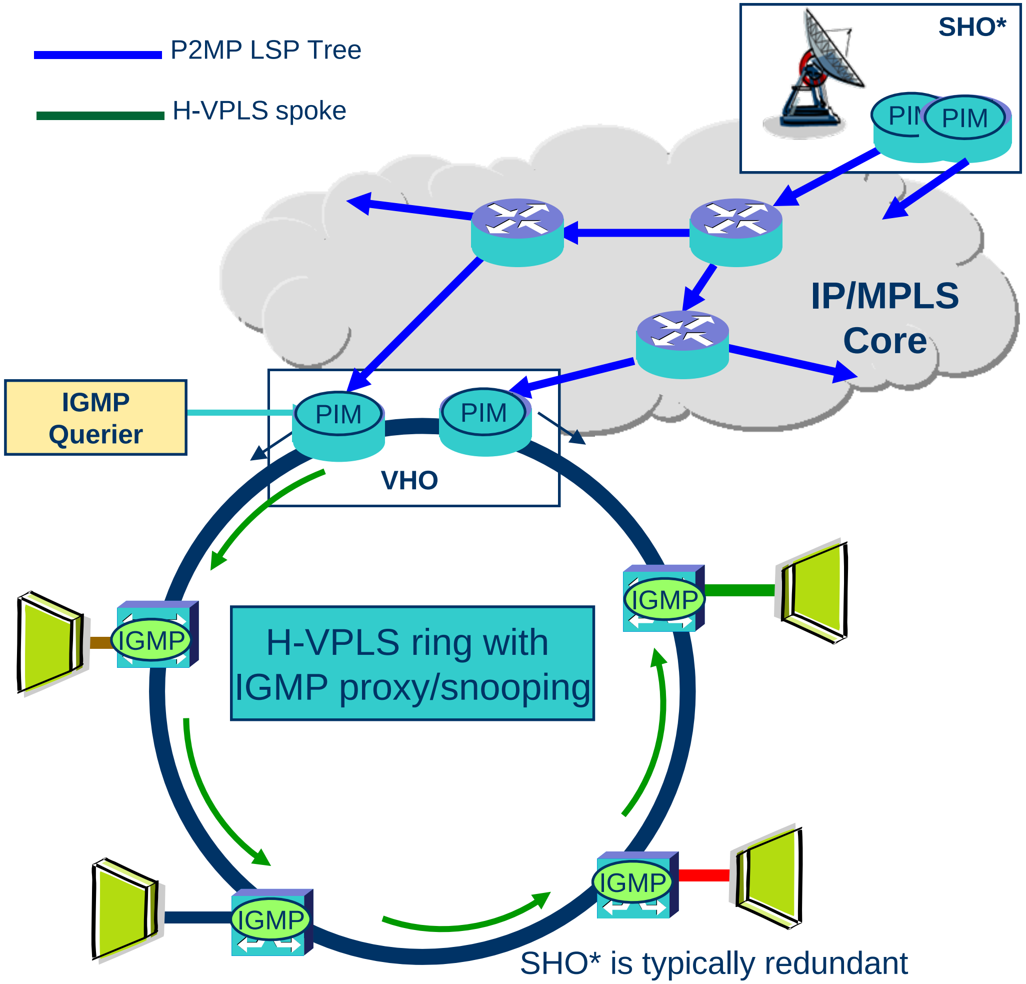
Two options:

- **PIM multicast routed backbone**
- **P2MP MPLS (“PIM-less” IP core)**

SHO: Super Hub Office
VHO: Video Hub Office

How mcast trees and VPLS mcast work together

- P2MP MPLS backbone**:**
  - **Static L2/L3 mcast topology**
  - **No PIM-routing in core**
  - **Sub 50 msec recovery**
- IP Edge (VHO locations)
  - **National channels from SHO**
  - **Local/regional channel insertion**
  - **PIM routing at edge**
- Mcast distribution network**:**
  - **Static L2, semi-static L3 mcast topology (channel requests)**
  - **H-VPLS ring w/IGMP proxy**
  - **Sub-50 msec recovery time**
  - **Mcast replication efficiency**



SHO* is typically redundant

Conclusions

- VPLS combines strengths of MPLS and Ethernet for multicast services
- Need to create and maintain a bandwidth efficient topology
  - **Flexible to support new modes of operations**
  - **Scalable and adaptable to handle evolving traffic patterns**
- Quality of Experience must be high, measurable and controlled
  - **QoS, security and accounting must be adopted to the reality of triple play services – protect premium content and services**
  - **Underpinned by non-stop service delivery and operation capabilities**
- As topology and services become more advanced, centralized control and administration tools must simplify operation
  - **OAM tools are critical to troubleshoot advanced services**
  - **Service-aware OAM tools help to quickly resolve issues**

Further Reading

- Requirements for Multicast Support in VPLS
  - draft-ietf-l2vpn-vpls-mcast-reqts-03.txt
- IGMP-Snooping
  - RFC 4541
- PIM-Snooping in VPLS
  - draft-ietf-l2vpn-vpls-pim-snooping-00.txt
- PE-to-PE Mcast State Distribution
  - draft-qiu-serbest-l2vpn-vpls-mcast-ldp-00.txt
- Mcast Trees in Provider Core
  - draft-raggarwa-l2vpn-vpls-mcast

For More Information. . .

- http://www.ipmplsforum.org
- http://www.ietf.org
- http://www.itu.int
- http://www.mplsrc.com

For questions, utilize the IP-MPLS Forum Message Board

Website: http://www.ipmlpsforum.org/board/

Thank you **for attending the**

Multicast in MPLS/VPLS Networks Tutorial

Please visit the IP/MPLS Forum Booth in the Exhibit Area