BUILDING MPLS-BASED MULTICAST VPN SOLUTION

DENOG3 Meeting, 20.10.2011/Frankfurt

Carsten Michel

Agenda

- Multicast VPN (mVPN) Overview
- L3VPN Multicast Solution using PIM/GRE (Draft-Rosen)
- MPLS Multicast Building Block: Point-to-Multipoint LSPs
- L3VPN Multicast Solution using BGP/MPLS (Next Generation mVPN)

What Is Multicast And Why Use It?

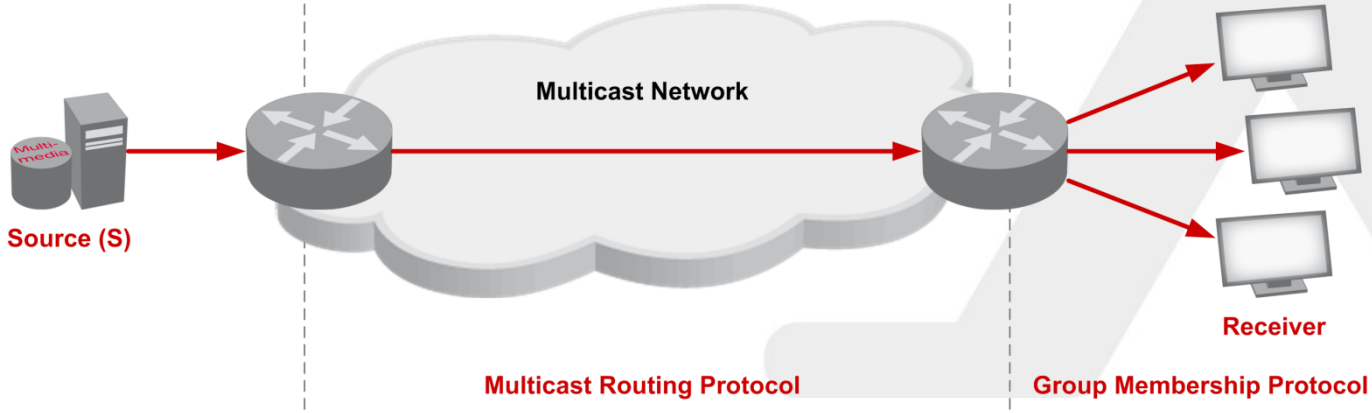
- Many applications transmit same data at the same time to multiple users

	Real Time	Non-real Time
Multimedia	• IPTV • Video Conferencing • Live Internet Audio	• Content delivery
Data	• Financial Market data • News feed • Interactive gaming	• Database replication • Server backup • Software distribution

- Network must have efficient mechanism to support these applications
- Multicast defines one-to-many or many-to-many communication across network infrastructure
  - Set of users seeking for the same data is called a (multicast) group
  - Will focus on IPv4 multicast in this talk

IP Multicast Terminology

- Source (S): originator of IP multicast packets
- Receiver: Host that is interested in receiving a particular multicast streams
- Group (G) address: IP address representing a multicast group (e.g. 224/8)
- Multicast routing protocol: protocol used between routers distribute multicast routing information (most likely PIM sparse mode)
  - Any source multicast (ASM) supports one-to-many and many-to-many applications
  - Source specific multicast (SSM) supports one-to-many applications
- Internet Group Membership Protocol (IGMP): protocol used to indicate host interest in receiving multicast traffic



Multicast VPN (mVPN) Overview

- BGP/MPLS IP Layer-3 VPNs as described in RFC 4364 (formerly RFC 2547) provide a way to transport customer IP unicast packets from one site to another
  - CE routers form unicast routing adjacencies with connected PE routers, but not with CE routers at other sites
- VPN services supporting native multicast transport upon MPLS VPN infrastructure are known as multicast VPN (mVPN)
  - Extend existing L3VPN framework to support IP multicast
  - Follow the same architecture as unicast L3VPN model
  - Re-use existing mechanism with extensions as needed
  - mVPN consists of two sets of sites, sender sites set and receiver sites set

mVPN Requirements

- Optimal multicast routing for a particular multicast flow (S,G) requires P routers to have state information about that specific flow
  - Remember, that P routers do not have any VPN-specific state in case of L3VPN unicast routing
  - Scalability is poor if number of P state is proportional to number of VPN multicast flows
- Optimal bandwidth usage requires network to avoid needless duplication and reception of data traffic
  - Same multicast data should not travel twice or more on the same link
  - PE routers should not receive multicast traffic if there are no downstream receivers connected
- Resource optimization requires a trade-off between state optimization and bandwidth optimization

PE-CE Multicast Routing

- As with unicast routing, CE routers maintain multicast adjacency with PE routers
  - No CE-CE overlay tunnels, i.e. no multicast routing adjacencies between CE routers at different sites
  - Multicast routing protocol on PE-CE link is assumed to be PIM
    - Support of PIM-SM in ASM and SSM mode as well as PIM-BIDIR required
    - IGMPv3 and MLDv2 required
  - No multicast operation/configuration changes requires within customer network, i.e. customer multicast deployment not affected at all (e.g. PIM modes, RPs, …)
- PE routers will have independent instances of multicast routing protocol per mVPN (multicast C-instances)

P-Multicast Service Interface (PMSI) and P-Tunnels

- P-Multicast Service Interface (PMSI) is a conceptual overlay that connects PE routers across the P network
  - PE within particular VPN sends packet to PMSI and packets will be delivered to some or all other PE routers allowing receiving PE to determine corresponding VPN, i.e. PMSI carries C-multicast data traffic
  - PMSI may be instantiated by different transport mechanisms called P-tunnel
- P-tunnels can be inclusive or selective.
  - Inclusive P-tunnel (I-PMSI P-tunnel) enables a PE router that is in the sender site set of an MVPN to transmit multicast data to all PE routers that are members of that MVPN.
  - Selective P-tunnel (S-PMSI P-tunnel) enables a PE router that is in the sender site set of an MVPN to transmit multicast data to a subset of the PEs.
- Different tunneling technologies can be used to create P-tunnels including
  - PIM-SM (ASM) signaled IP/GRE P-tunnels
  - mLDP or RSVP-TE signaled MPLS P-tunnels
  - MPLS p2p LSP using ingress replication

mVPN Control Plane

- The MVPN control plane according to draft-ietf-l3vpn-2547bis-mcast must support the following tasks
  - mVPN autodiscovery, i.e. a PE router discovers the identity of the other PE routers that participate in the same MVPN
  - Distribution of P-tunnel information, i.e. a sender PE router advertises the type and identifier of the P-tunnel that it will be using for transmitting VPN multicast packets
  - PE-PE C-multicast route exchange, i.e. a receiver PE router propagates C-multicast join messages (C-joins) received over its VPN interface towards the VPN multicast sources.

Agenda

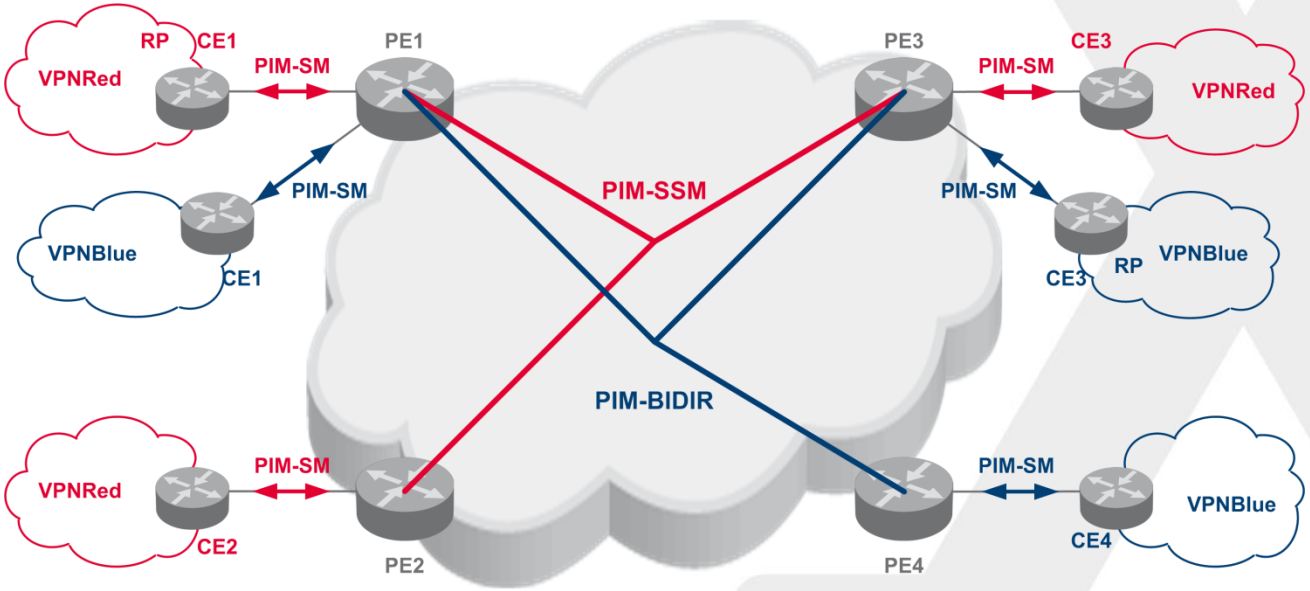
- Multicast VPN (mVPN) Overview
- L3VPN Multicast Solution using PIM/GRE (Draft-Rosen)
- MPLS Multicast Building Block: Point-to-Multipoint LSPs
- L3VPN Multicast Solution using BGP/MPLS (Next Generation mVPN)

Draft-Rosen mVPN Solution

- Based on draft-rosen-vpn-mcast-12.txt
- Many similarities to unicast MPLS VPNs (RFC 4364)
- PIM-SM used to build P-Trees connecting PEs
- Optionally auto-discovery of PE routers using BGP
- PE exchange customer routing with PIM
  - mVPN uses single inclusive P-tunnel (aka default MDT)
  - mVPN may use one or more selective P-tunnels (aka data MDT)
  - Encapsulation is GRE (outer header contains group address of MDT)

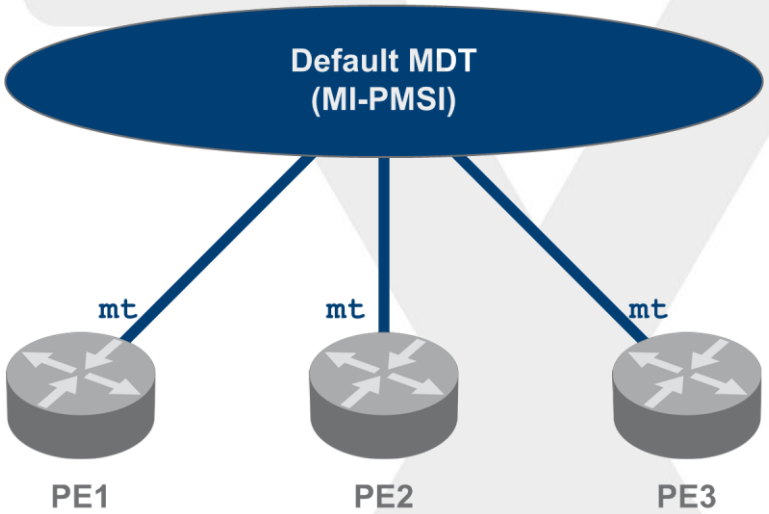
Draft-Rosen mVPN Overview

- Each mVPN is connected via an individual core multicast tree (P-Tree)
  - Provider network uses its own PIM instance (P-PIM)
  - Customer multicast deployment (C-PIM) independent of core multicast implementation; CE routers maintain PIM adjacency with PE routers only



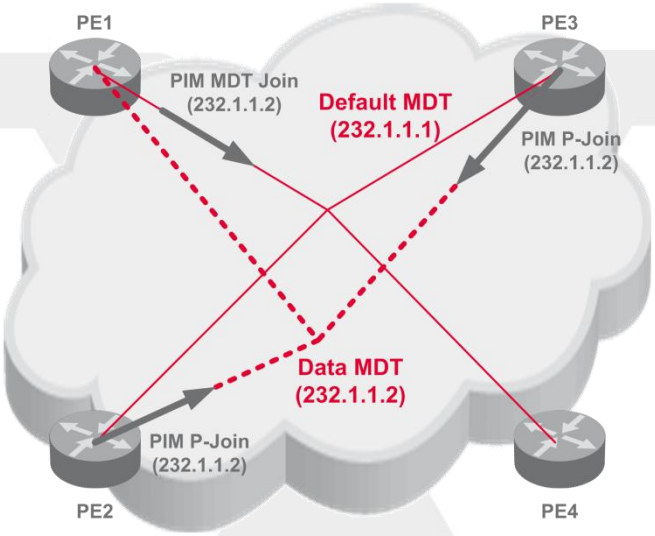
Default MDT and Multicast Tunnel Interface

- Default MDT is used as a permanent channel carrying PIM control messages and low bandwidth streams
- PE establishes an PIM adjacency with all other PEs in mVPN
- Access to default MDT from mVPN via Multicast Tunnel interface
  - Default MDT is treated as LAN
  - RPF is executed against multicast tunnel interface
- PE is always a root (source) of the MDT and also a leaf (receiver) to the MDT rooted on remote PEs



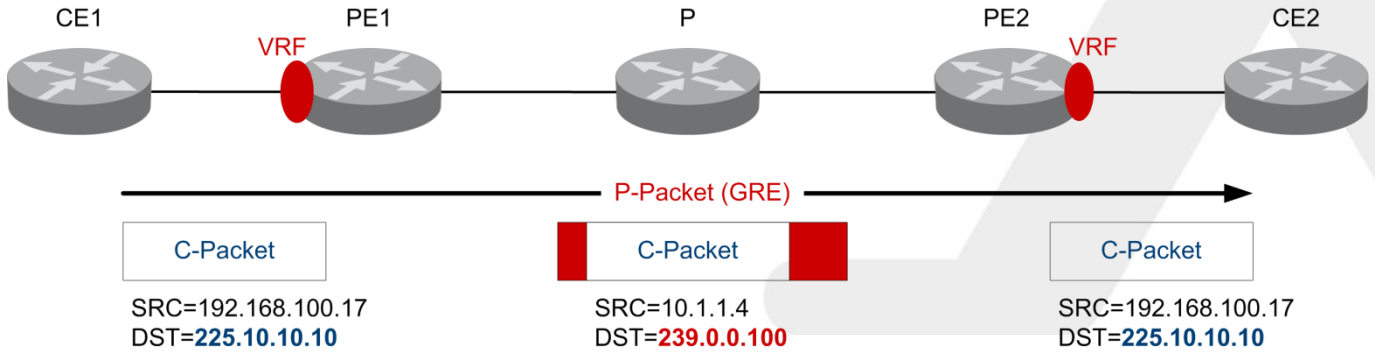
Data MDT

- Data MDTs created on demand between PE routers connecting to source and active receivers to optimize multicast forwarding trees for customer (S, G) states only
- New PIM MDT Join message used to announce when traffic exceeds pre-configured threshold
  - Sent via default MDT group every 60s as long as the traffic rate remains over threshold
- All PEs on the default MDT will receive message and interested PEs can join new data MDT



Draft-Rosen C-Multicast Packet Forwarding

- PE maps all (C-S,C-G) and (*,C-G) that can exist in a particular mVPN to a single (P-S,P-G) or (*,P-G) MDT group in the P-network
  - Mapping is achieved by encapsulating C-packets into P-packets using multicast GRE; no MPLS labels used for any multicast packet
    - Source address of GRE packet is ingress PE router address
    - Destination address of GRE packet is the MDT group address
    - C-packet IP ToS will be copied to P-packet
  - Number of P states is a function of number of mVPNs rather than customer multicast groups



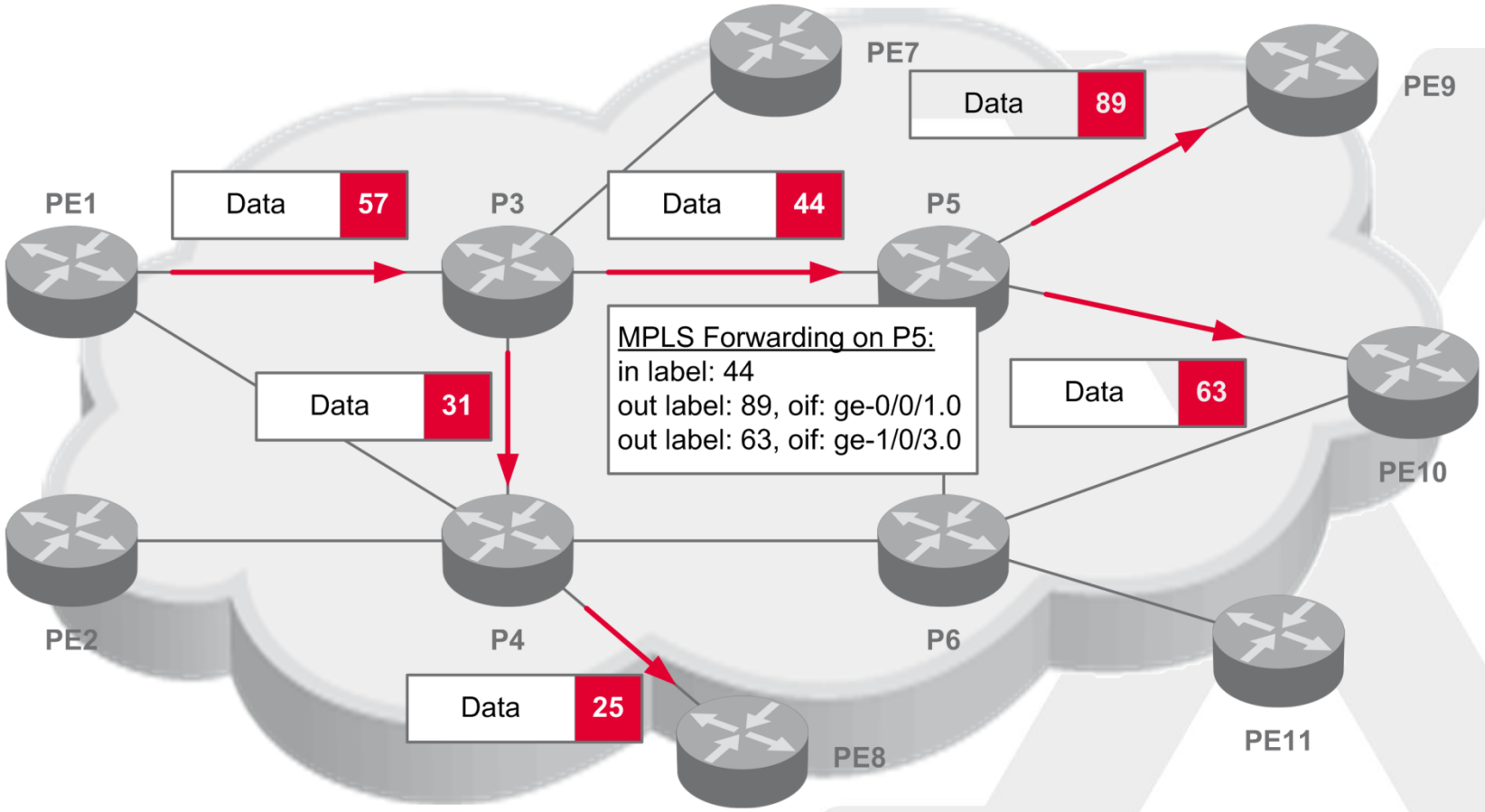
Agenda

- Multicast VPN (mVPN) Overview
- L3VPN Multicast Solution using PIM/GRE (Draft-Rosen)
- MPLS Multicast Building Block: Point-to-Multipoint LSPs
- L3VPN Multicast Solution using BGP/MPLS (Next Generation mVPN)

Point-to-Multipoint (P2MP) LSPs

- Point-to-multipoint LSP is the fundamental construct to support multicast over MPLS networks
  - Each p2mp LSP has a single root node, but multiple leave nodes
  - All leave nodes have to know the identity of the root node
  - Root node may or may not know the identity of the leave nodes
  - For a given p2mp LSP, the root node and all leave nodes must agree on the FEC bound the LSP
- Next Hop Label Forwarding Entry (NHLFE) maintained by each node along a p2mp LSP
  - Defines a set of next hops
  - Replicates the packets, i.e. a copy of each packet is sent to each of the specified next hops

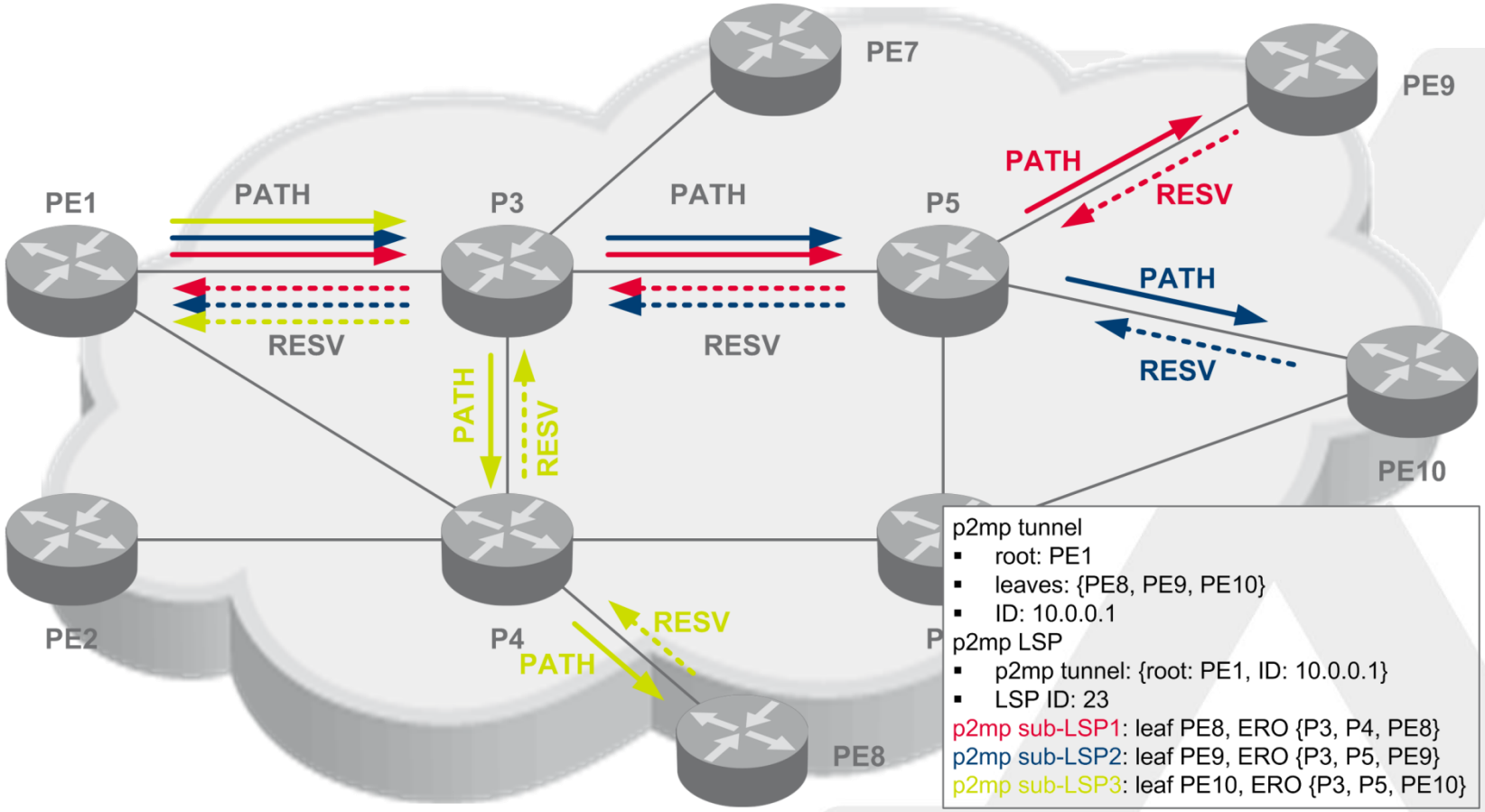
P2MP LSP Data Plane Example



P2MP LSP Setup using RSVP-TE

- Root node detects IP addresses of all the leave node
  - By means outside of RSVP-TE
- Root node computes path from itself to all leave 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
  - May also be pre-computed by off-line tools
- Root node uses RSVP-TE extensions to set up p2mp LSP
  - Establishes label forwarding state and may involve resource reservations
  - New building blocks include
    - p2mp tunnel is identified by p2mp session object and consists of multiple p2mp LSPs
    - p2mp LSP is a specific instance of a p2mp tunnel and comprises of one or more S2L sub-LSPs
    - S2L sub-LSP is an LSP from the root node to a particular leaf node

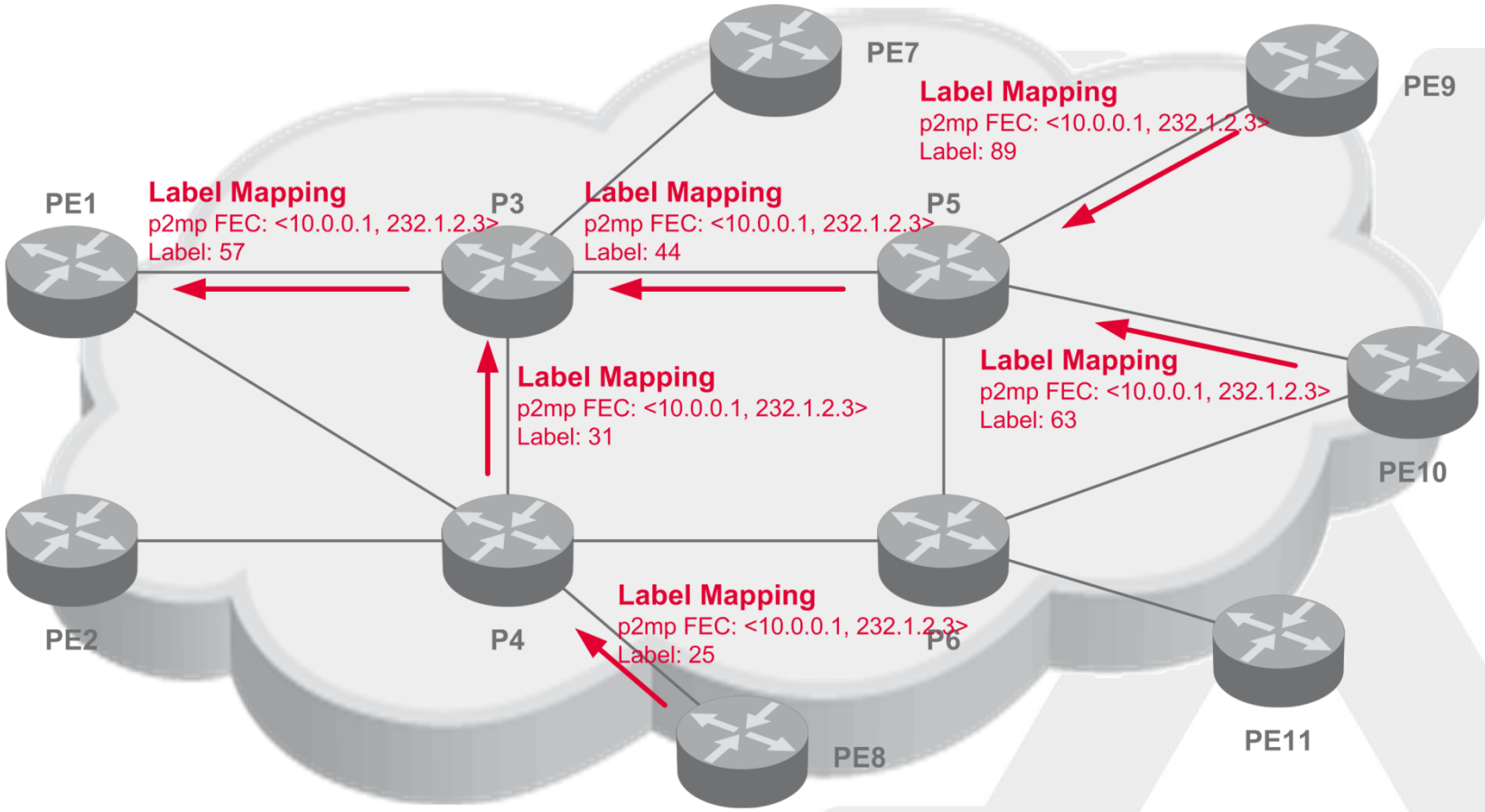
RSVP-TE P2MP Signaling Mechanism



P2MP LSP Setup using mLDP

- All leaf nodes find the identity of the LSP which includes the IP address of the root node
  - By means outside of LDP
- 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
- 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

mLDP P2MP Signaling Mechanism



Agenda

- Multicast VPN (mVPN) Overview
- L3VPN Multicast Solution using PIM/GRE (Draft-Rosen)
- MPLS Multicast Building Block: Point-to-Multipoint LSPs
- L3VPN Multicast Solution using BGP/MPLS (Next Generation mVPN)

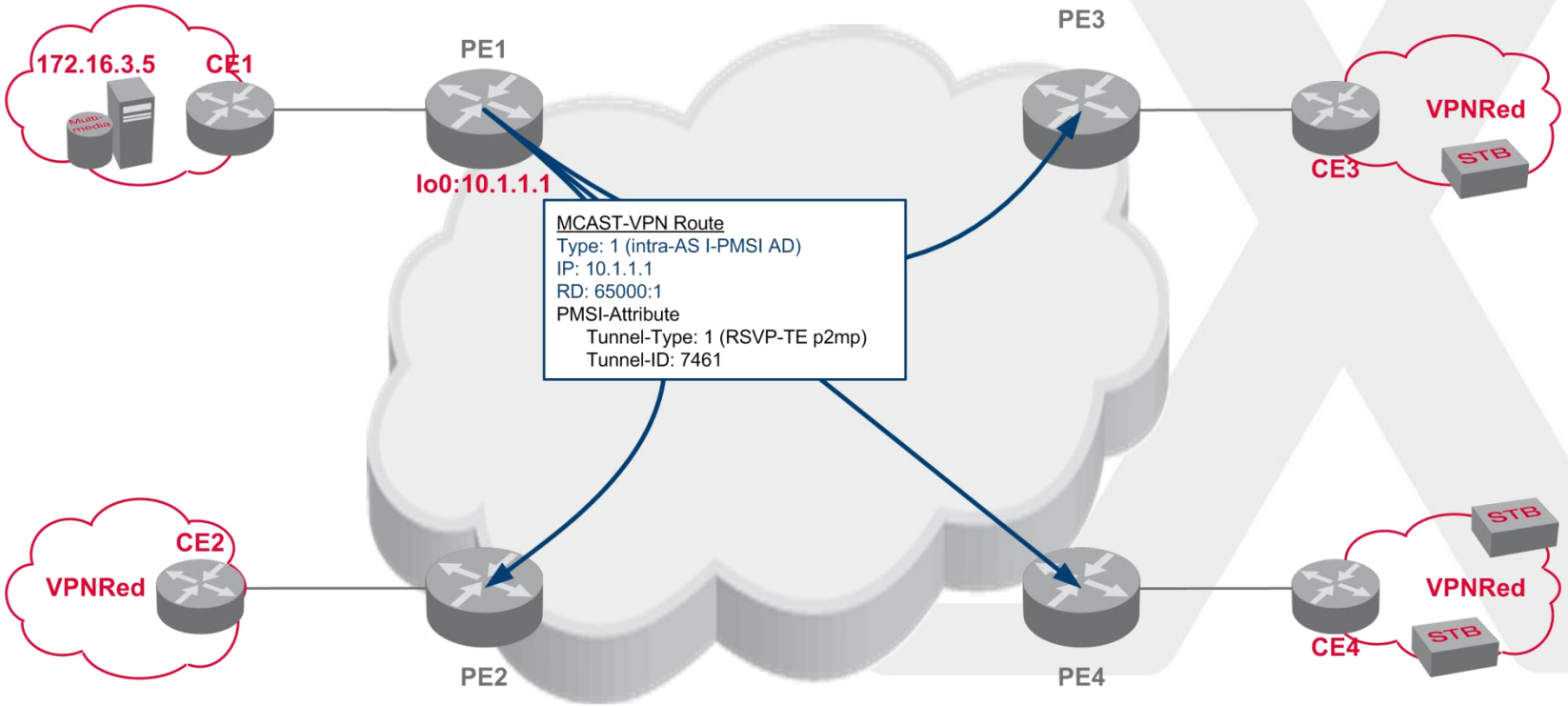
BGP MCAST-VPN Address Family and Route Types

- New BGP address family called MCAST-VPN for supporting NG MVPN control plane operations
  - Subsequent address family identifier (SAFI) of 5 by IANA.
  - MCAST-VPN NLRI contains route type, length, and variable fields depending on the route type

Route Type	Name	Used for …
1	Intra-AS I-PMSI AD route	Autodiscovery for inclusive P-tunnels
2	Inter-AS I-PMSI AD route	
3	S-PMSI AD route	Autodiscovery for selective P-tunnels
4	Leaf AD route	
5	Source active AD route	VPN multicast source discovery
6	Shared tree join route	C-multicast routes
7	Source tree join route	

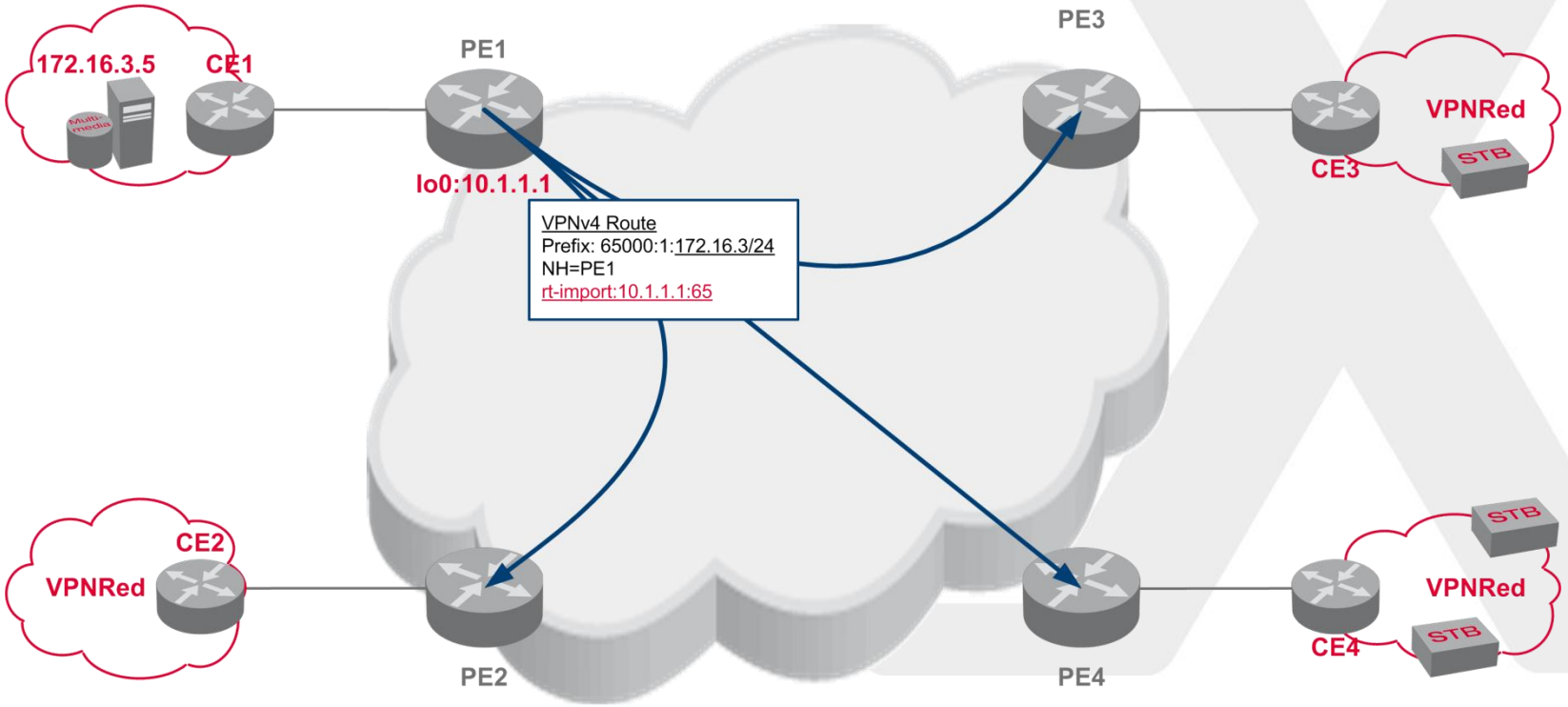
BGP Auto-Discovery

- BGP used for auto-discovery of which PEs are members of each mVPN
  - AD Route includes information which tunnel type will be used (PMSI tunnel attribute) by ingress PE to send traffic



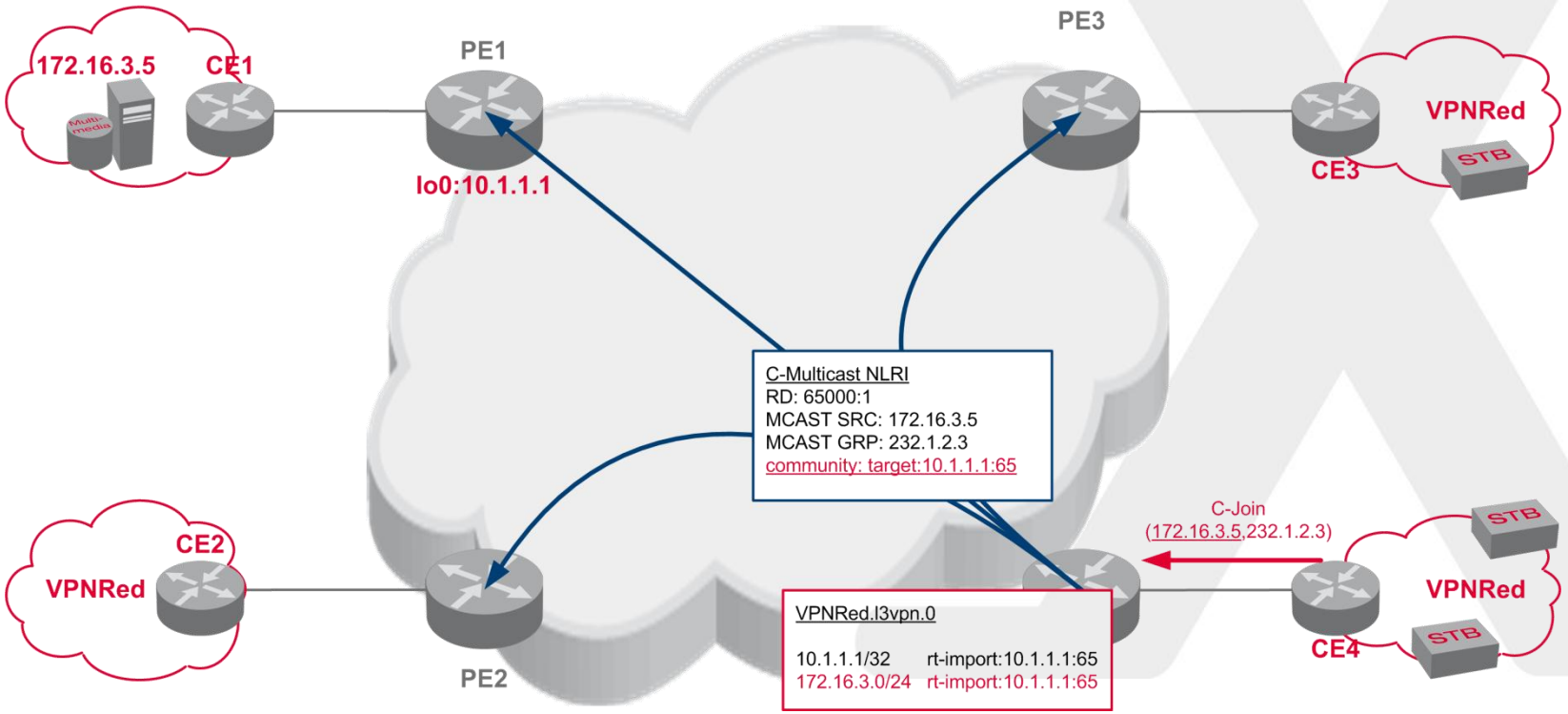
NG mVPN Control Plane (1/3)

- VPN-IPv4 routes carry two new extended community attributes
  - route-import is a auto-generated extended community attribute that controls import of C-multicast routes into a particular VRF



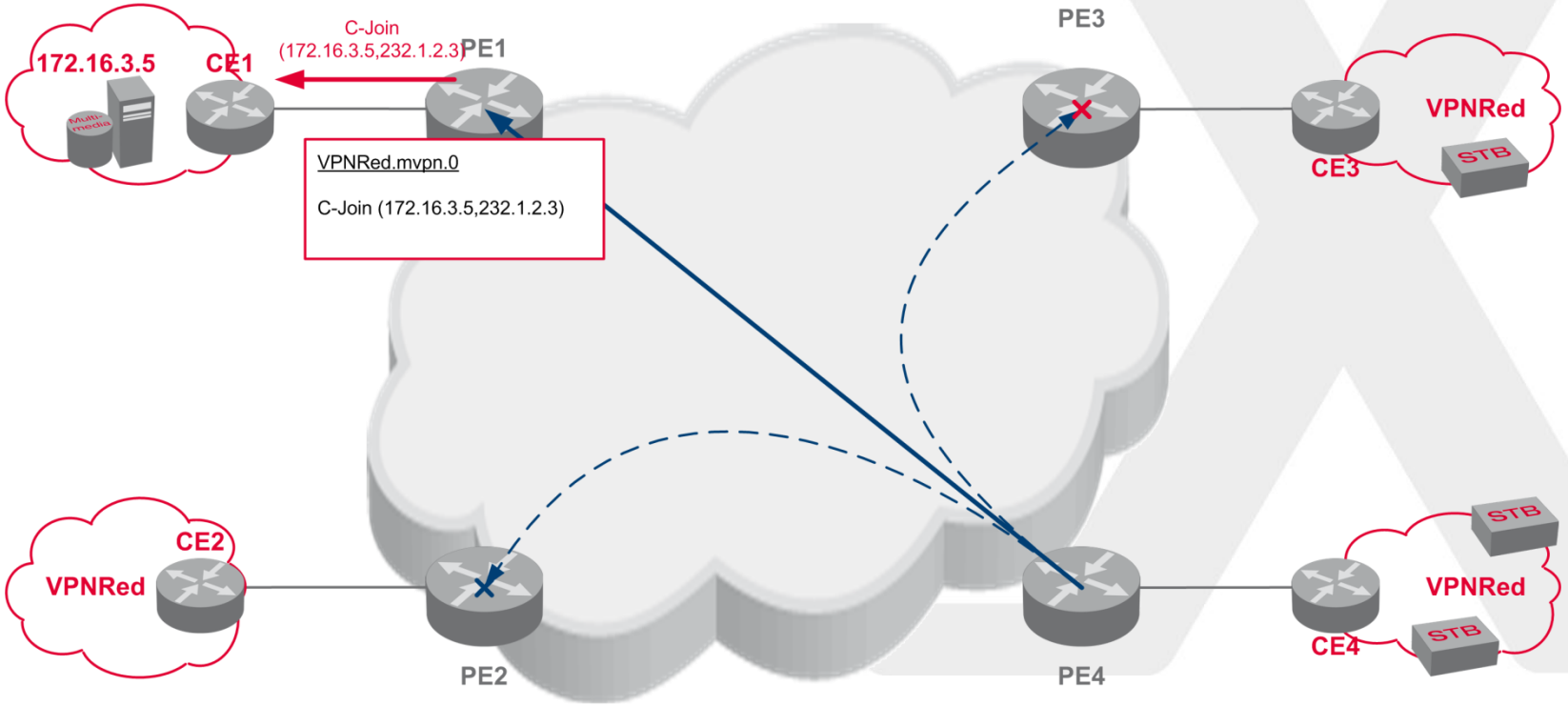
NG mVPN Control Plane (2/3)

- When PE router receives PIM Join messages from CE router for (S,G), it constructs a C-multicast route by looking up VPN-IPv4 route to source
  - Uses route-import attribute attached to the source VPN-IPv4 route for export



NG mVPN Control Plane (3/3)

- PE router accepts C-multicast route if attached route target matches the route-import for the specified VRF
  - Create (S,G) state in VRF and propagate information via PIM to CE router



mVPN Any Source Multicast (ASM)

- ASM by nature is fundamentally more complex that SSM due to rendezvous points (RP), RP tree, shortest path tree (SPT), etc.
  - NG mVPN implementation uses traffic optimization to allow multicast packet to flow immediately along shortest-path tree
- Within each MVPN use the existing IP multicast mechanisms to enable C-RPs to discover active sources (S,G) of that mVPN
- Use the existing IP multicast mechanisms to communicate this information from C-RPs to one or more PEs
  - Option 1: At least one PE is C-RP for the mVPN
  - Option 2: PE router learns active sources via MSDP from C-RP
- The PEs that learn active sources from customer sites advertise these to other PEs using BGP SA routes

Selective Trees

- Assume (S,G) is a high bandwidth stream, so ingress PE only wants to send it to those egress PE routers with interested receivers
  - Move particular multicast streams to selective trees (S-PMSI) for bandwidth optimization
- Static selective tree
  - Egress nodes of selective p2mp LSP must be configured on ingress PE
  - Can have multiple (S,G) mapping onto the same selective tree
  - Ingress PE advertises existence of selective tree using a BGP AD route
- Dynamic selective tree
  - One-to-one mapping between (S,G) and selective tree
  - Ingress PE advertises existence of the selective tree using a BGP AD route
  - Interested PEs need to send Leaf AD route to ingress PE

Single Forwarder Election

- Suppose multicast source is multihomed to two or more ingress PE routers for redundancy
  - Need to ensure that only one PE acts as ingress router forwarding traffic for a given (S,G) into the mVPN
  - Otherwise an egress PE may receive duplicate packets!
- Hence have single forwarder election process
  - PEs select the ingress PE with the highest IP address (default method)
  - Other methods may be implemented (e.g. local preference)
  - When egress PE sends a C-multicast route, it attaches the unique route -import associated with the selected ingress PE, so that only that ingress PE installs the route and forwards the multicast traffic

NG-mVPN Extranet

- Industry interest in NG-MVPN extranet functionality
  - IPTV wholesale
  - Financial real-time data feeds
- Definition of Multicast Extranet (c.f. http://tools.ietf.org/html/draft-raggarwa-l3vpn-bgp-mvpn-extranet-04)

  "…the ability for multicast sources in one MVPN to send multicast traffic to multicast receivers in other MVPN(s), and likewise, the ability for multicast receivers in one MVPN to receive multicast traffic from multicast sources in other MVPN(s)"
- Extranets have been available for unicast L3VPN for many years
  - Via manipulation of route targets
  - Can use very similar approach in NG-MVPN case, as NG-MVPN uses very similar mechanisms to unicast L3VPN
- For more information see presentation "Next-Generation Multicast VPN Extranets" at MPLS World Congress 2011, Paris, by Julian Lucek co-authored by Xantaro

Key Take Away

- Next Generation Multicast VPN solution deployed in more than 25 networks worldwide including
  - Professional Broadcast TV
  - Financial data feeds/trades
  - IPTV/CATV distribution
  - MVPN service offered by SPs to Enterprises/MNOs
  - CCTV monitoring
- Framework provides flexibility in data plane to adjust customer needs

" What did you take away from the meeting ? "