Multicast Support in 2547 VPNs and VPLS – is it Good Enough ?

Yakov Rekhter
Juniper Networks
yakov@juniper.net

MPLS 2004

Juniper NETWORKS

- **Assumption:** a large percentage of 2547 VPN and VPLS customers would have not only unicast, but multicast traffic as well
- **Assumption:** volume of the multicast traffic may be non-negligible relative to the volume of the unicast traffic
- **Disclaimer:** This talk is NOT about multicast in the Internet – this is about multicast support for 2547 VPNs and VPLS services

Agenda

- What are the (desirable) goals
- Support for multicast in 2547 VPNs - current proposals and their shortcomings
- Support for multicast in VPLS - current proposals and their shortcomings
- Final remarks

Design Objectives for multicast support in 2547 VPN/VPLS service (not a complete list)

Optimize Bandwidth:

- A given customer (multicast) packet should traverse a given service provider link at most once
- Deliver customer multicast traffic to only PEs that have (customer) receivers for that traffic
- Deliver customer multicast traffic along the "optimal" paths within the service provider (from the ingress PE to the egress PEs)
  - Shortest Path tree (optimizes delay) or Minimum Cost tree (optimizes total bandwidth) ?

Optimize State:

- The amount of state within the service provider network required to support Multicast in 2547 VPN and VPLS service should be no greater than what is required to support Unicast in 2547 VPN and VPLS service
- The overhead of maintaining the state to support Multicast in 2547 VPN and VPLS service should be no greater than what is required to support Unicast in 2547 VPN and VPLS service
  - including the overhead due to the protocol(s) that maintain the state

Agenda

- What are the (desirable) goals
- **Support for multicast in 2547 VPNs - current proposals and their shortcomings**
- Support for multicast in VPLS - current proposals and their shortcomings
- Final remarks

Current proposals:

- "Multicast in MPLS/BGP IP VPNs" (draft-rosen-vpn-mcast-07.txt)
- "Base Specification for Multicast in BGP/MPLS VPNs"(draft-raggarwa-l3vpn-2547-mvpn-00.txt)

Multicast Support for 2547 VPN: components

- **Control plane:** exchanging VPN multicast routing information:
  - Between CE and PE
  - Among PEs
- **Data plane:** forwarding VPN multicast traffic within the service provider

From the decomposition point of view is very similar to the unicast support for 2547 VPNs

Exchanging VPN multicast routing information

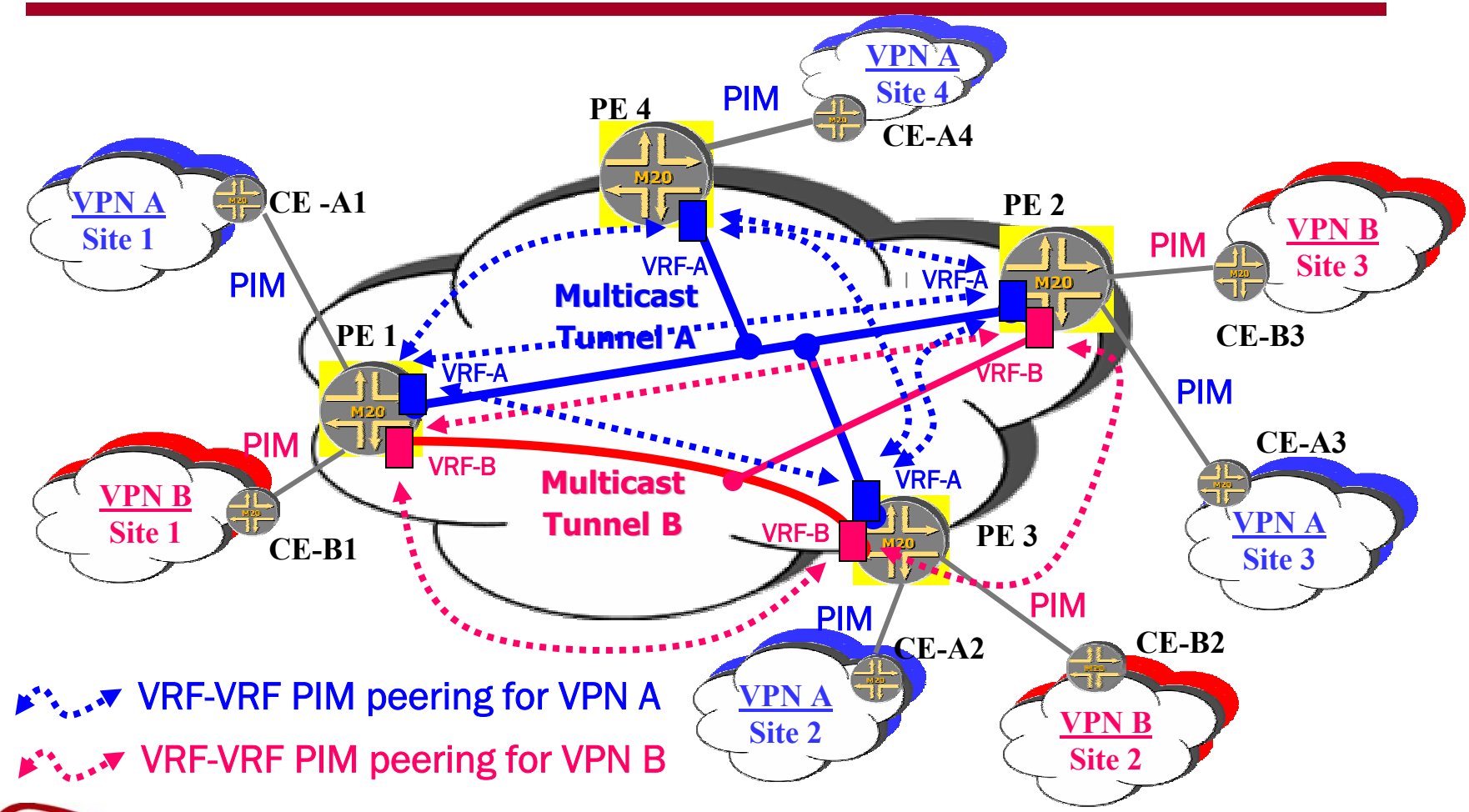
Between CE and PE:

- PIM
  - No changes to the protocol
- CE maintains PIM peering with its directly connected PE
  - CE does NOT maintain peering with CEs in other sites of that VPN
- PIM implementation on PE has to be VPN-aware:
  - PE processes PIM messages exchanged between PE and CE in the context of a particular VPN

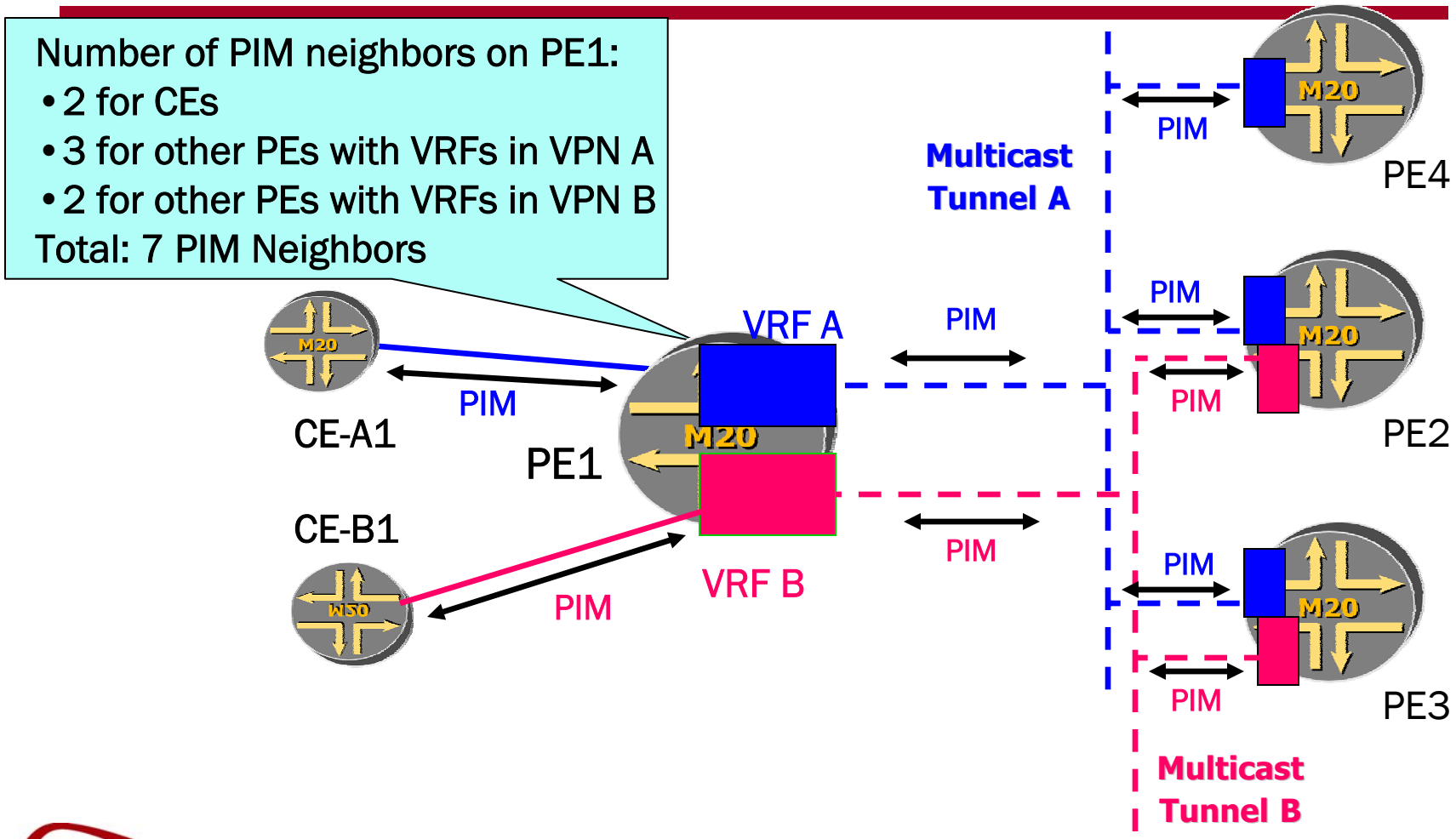
Among PEs:

- PE maintains a distinct PIM instance for each VPN
  - Need it anyway for exchanging multicast routing information between PE and CE
- PE maintains PIM peering with **all the other PEs** that have VPN in common
- PIM peering is on a **per VPN** (NOT per PE) **basis**
- Use Multicast Tunnel to exchange PIM messages
- From the VRF's point of view Multicast Tunnel looks like a LAN

Exchanging VPN multicast routing information: example



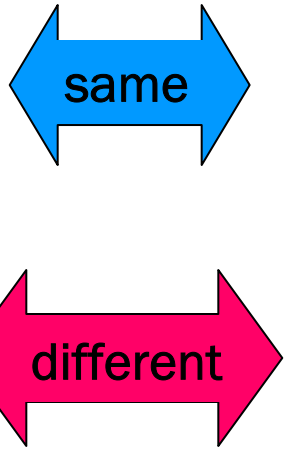
Exchanging VPN multicast routing information: PE1 perspective



Routing peering on PE routers: 2547 VPN Unicast vs 2547 VPN Multicast

2547 VPN Unicast:

- PE router maintains routing peering (IGP or BGP) with its directly connected CEs
- PE router maintains routing peering (BGP) with a small number of BGP Route Reflectors (to exchange VPN routing information)



2547 VPN Multicast:

- PE router maintains routing peering (PIM) with its directly connected CEs
- PE router maintains routing peering (PIM) with all other PE routers for which it has at least one VPN in common
  - Distinct instance of PIM peering per each VPN !

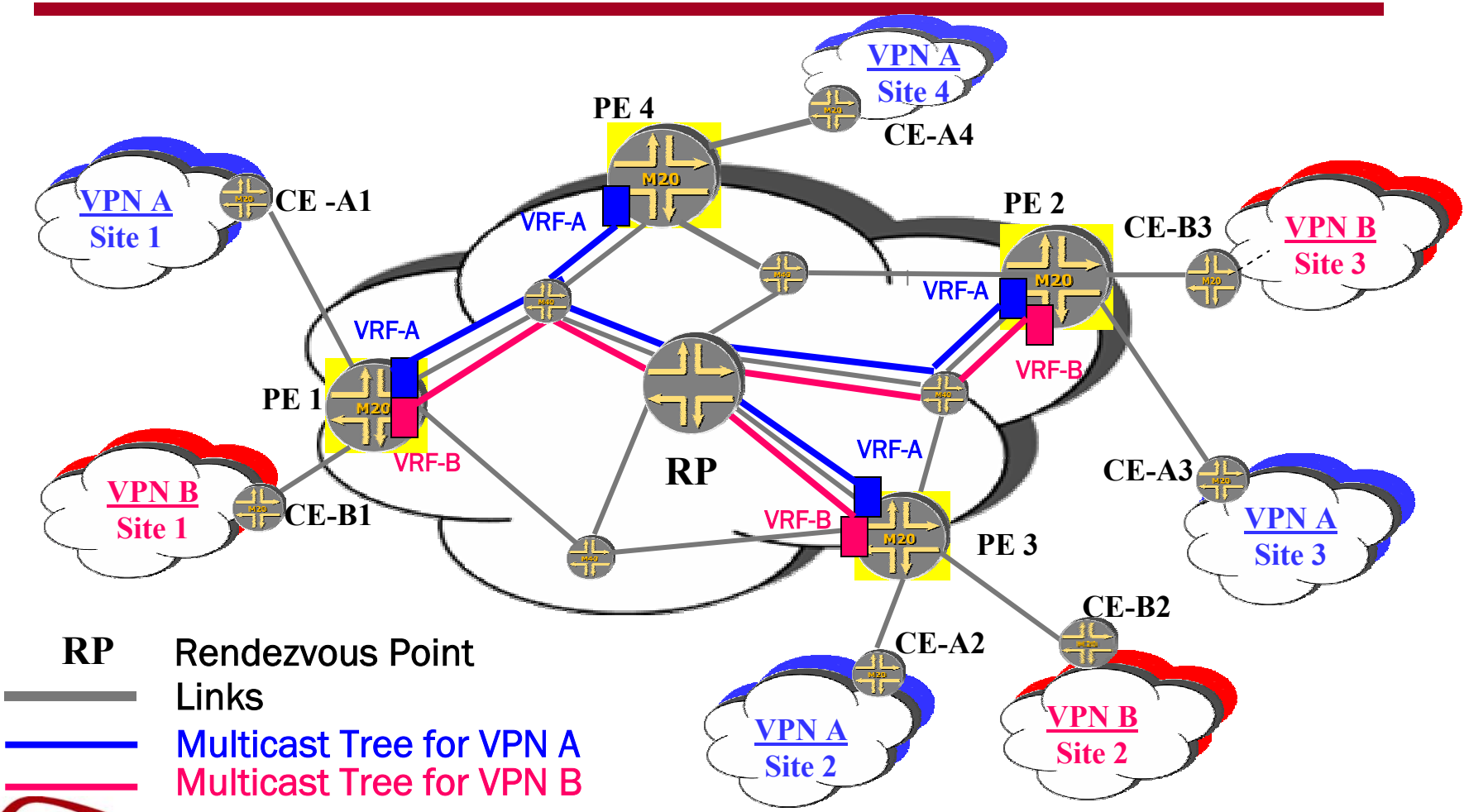
Implications on the state maintenance overhead

- Number of routing peers (PIM neighbors) PE router has to maintain is dominated by the number of directly connected VPNs/CEs **times average number of sites per VPN**
  - E.g., assume PE router with 1,000 CEs, each of these CEs is part of a distinct VPN, each VPN has on average 100 sites, PE router has to maintain **~100,000 PIM neighbors** ☹
- Amount of PIM traffic PE router has to handle is dominated by the PIM traffic between PE and all other PEs that have at least one VPN in common
  - E.g., assume PE with 1,000 CEs, each CE is part of a distinct VPN, each VPN has on average 100 sites, PE router has to process **~3,300 PIM Hellos per second** ☹
  - **PIM Join adds more PIM traffic** ☹

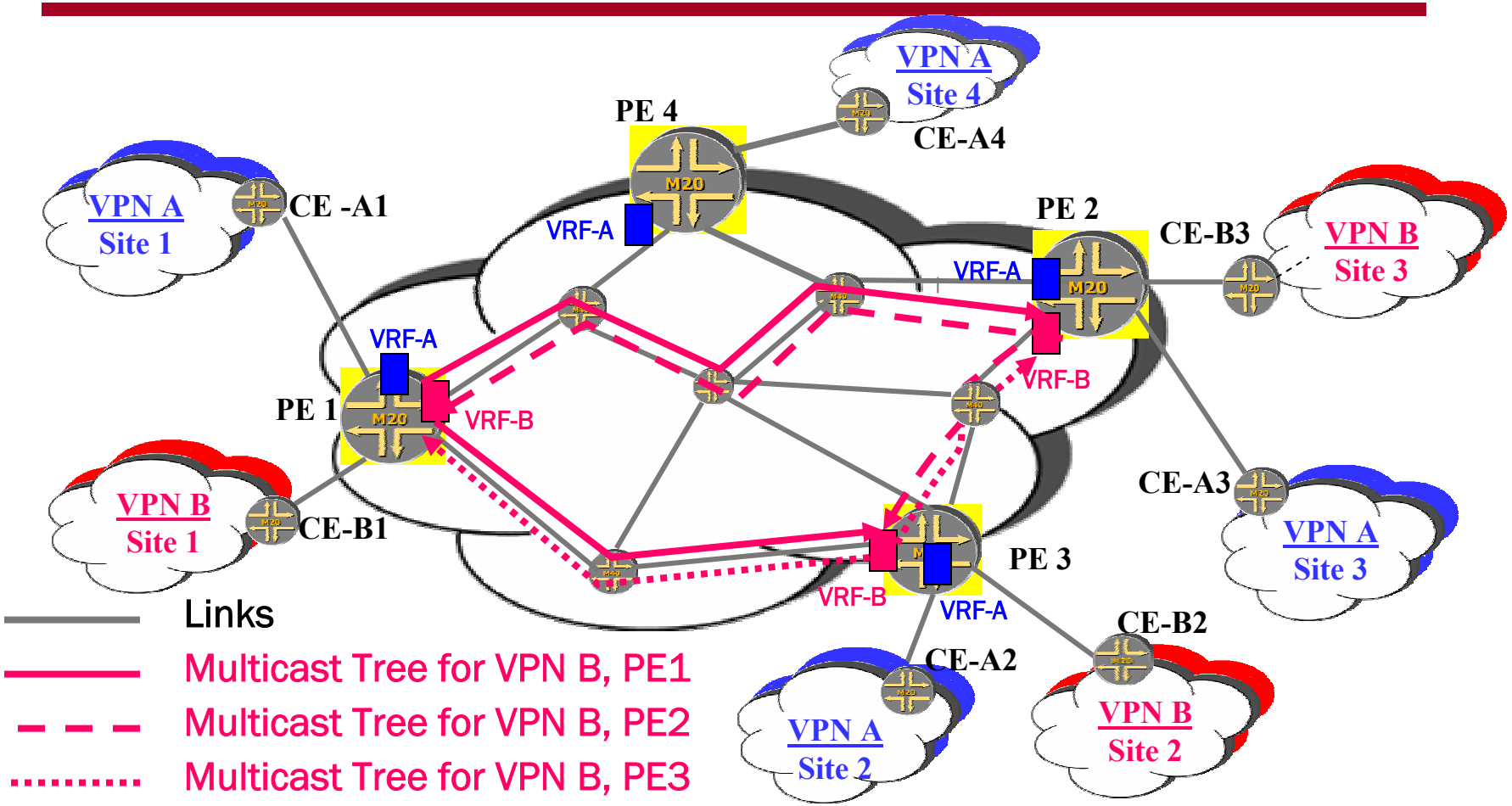
Forwarding VPN multicast traffic within the service provider – Multicast Tunnels

- **Option 1:** Distinct Service Provider multicast tree **per VPN** (PIM-SM or Bidirectional PIM) – emulate Multicast Tunnel via a single tree
  - **Use of PIM-SM assumes no switch to source-specific tree**
- **Option 2:** Distinct Service Provider multicast tree **per VPN per PE** (PIM-SSM) – emulate Multicast Tunnel via a collection of trees
- **Option 3:** Distinct Service Provider tree **per VPN** (PIM-SM or Bidirectional PIM), plus distinct Service Provider tree **per (S, G)** for subset of (S,G) of a given VPN (PIM-SSM)
- **Option 4**: Distinct Service Provider tree **per VPN per PE** (PIM-SSM), plus distinct Service Provider tree **per (S, G)** for subset of (S, G) of a given VPN (PIM-SSM)

Option 1: distinct multicast tree per VPN (PIM-SM) - example

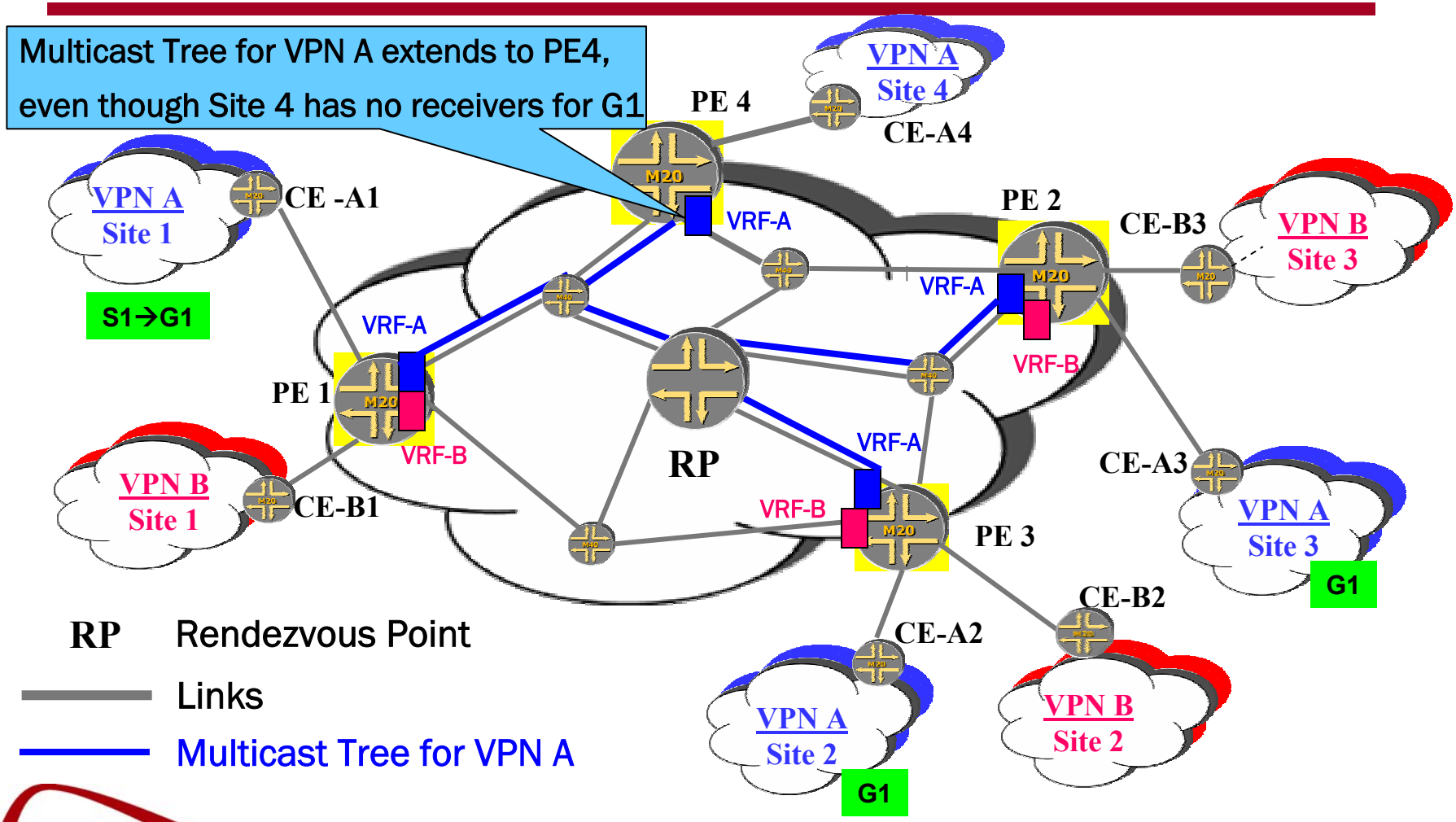


Option 2: distinct multicast tree per VPN per PE (PIM-SSM) - example

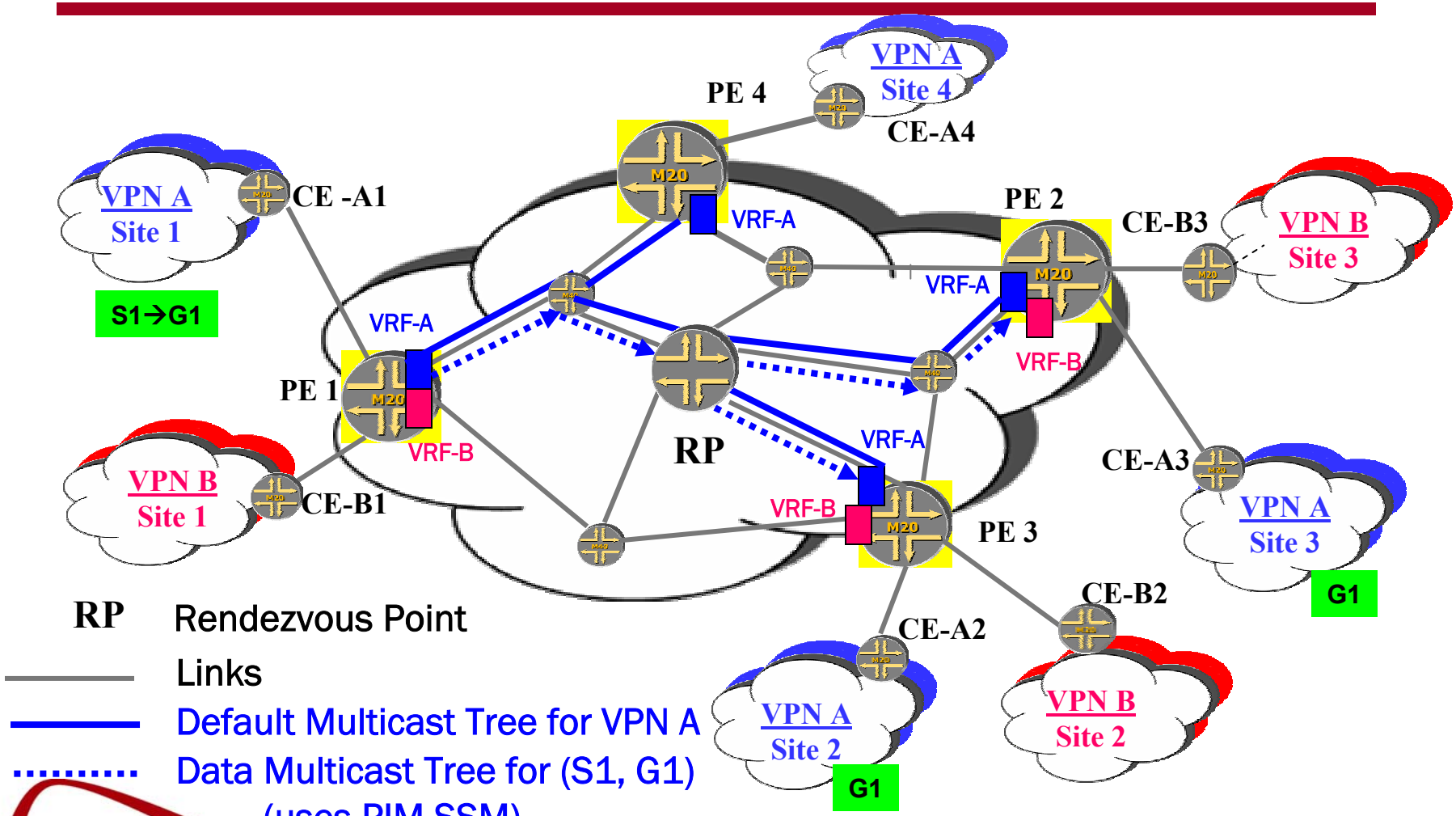


3 times as many trees as with the previous example !

Default Multicast Tunnel: sending multicast traffic to PEs with no receivers



Option 3: avoid sending traffic to PEs with no receivers - Data Multicast Tunnel



Implications on the amount of state in the service provider routers: unicast vs multicast

2547 Unicast:

- No VPN related state on P routers
- VPN-related state only on PE routers, and only for the VPNs directly connected to that PE
- Amount of state on P routers is bounded by the total number of PEs
  - Total number of LSPs with LDP is O(#PEs)
  - Total number of LSPs with RSVP is O(#PEs^2)

2547 Multicast:

- **Option 1 - distinct multicast tree per VPN:** provider has to maintain #trees = #VPNs
  - E.g. assume 10,000 VPNs per service provider requires **10,000** multicast trees
- **Option 2 - distinct multicast tree per VPN per PE:** provider has to maintain #trees = #VPNs times average number of PEs per VPN
  - E.g., assume 10,000 VPNs per service provider with each VPN present on average on 100 PEs requires **1,000,000** multicast trees ☹
- **Options 3, 4** (Data Multicast Tunnels)**:** adds even more state to maintain ☹

Forwarding VPN traffic: comparison

	A given customer (multicast) packet should traverse a given Service Provider link at most once	Delivering customer multicast traffic along the "optimal" paths within the Service Provider (from ingress PE to the egress PEs)	Delivering customer multicast traffic to only the PEs that have (active) (customer) receivers for that traffic	Additional state within the service provider network
Distinct Service Provider tree per VPN (PIM-SM)	yes, except for the traffic to Rendezvous Point from source PEs	no (as all traffic has to go through Rendezvous Point)	no	1 tree per VPN
Distinct Service Provider tree per site per VPN (PIM-SSM)	yes	yes, if "optimal" means min delay no, if "optimal" means min bandwidth	no	1 tree per VPN per PE
Distinct Service Provider tree per VPN (PIM-SM), plus distinct Service Provider tree per (S, G) for subset of (S,G) of a given VPN (PIM-SSM)	yes, except for the traffic to Rendezvous Point from source PEs	yes, but only for some traffic, and if "optimal" means min delay no, if "optimal" means min bandwidth	yes, but only for some traffic	1 tree per VPN plus one tree per (S, G) for subset of (S, G) of a given VPN
Distinct Service Provider tree per VPN per PE (PIM-SSM), plus distinct Service Provider tree per (S, G) for subset of (S, G) of a given VPN (PIM-SSM)	yes	yes, if "optimal" means min delay no, if "optimal" means min bandwidth	yes, but only for some traffic	1 tree per VPN per PE, plus one tree per (S, G) for subset of (S,G) of a given VPN

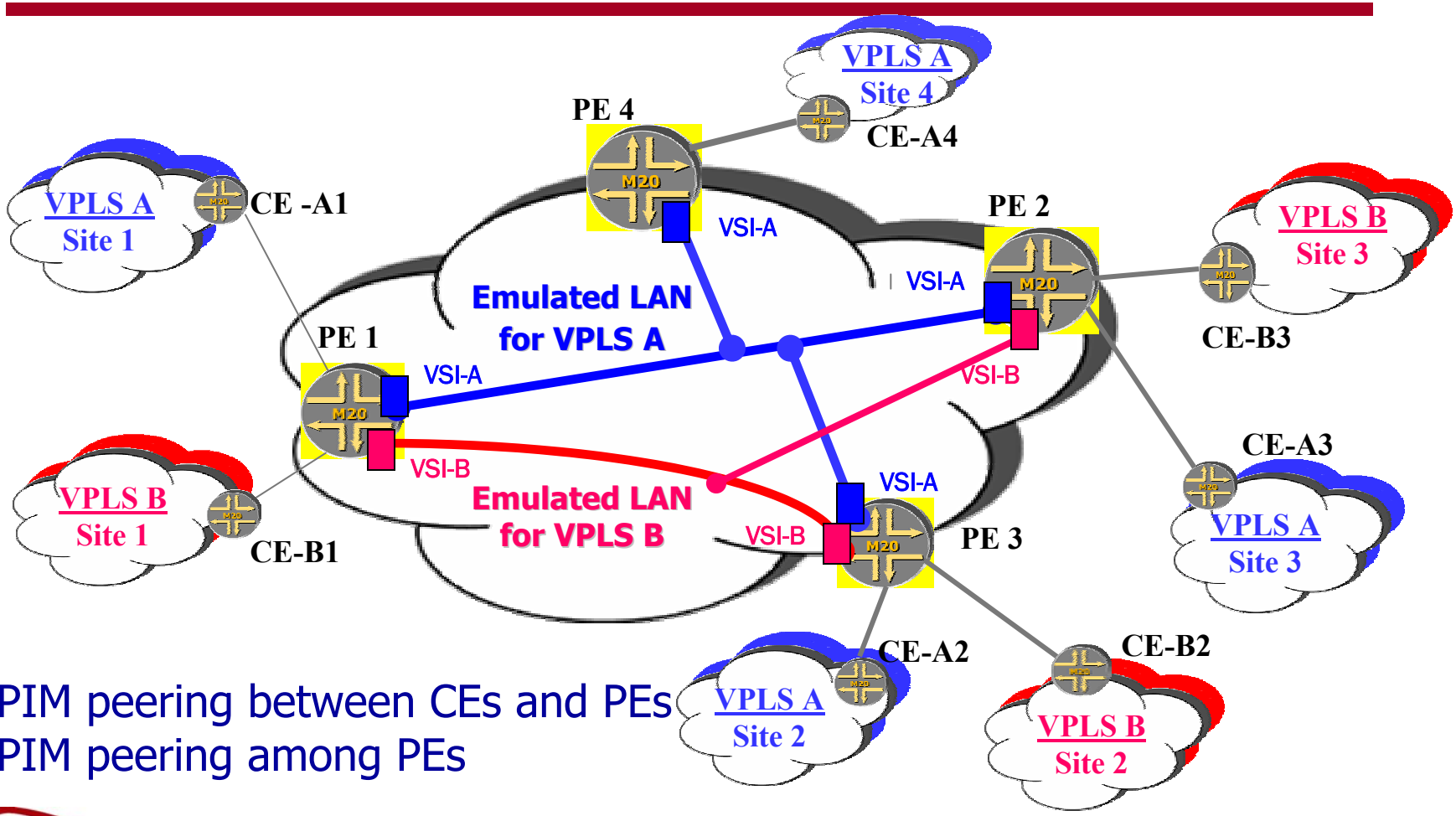
Agenda

- What are the (desirable) goals
- Support for multicast in 2547 VPNs - current proposals and their shortcomings
- **Support for multicast in VPLS - current proposals and their shortcomings**
- Final remarks

Current proposals:

- "Virtual Private LAN Service" (draft-ietf-l2vpn-vpls-bgp-02.txt)
- "Virtual Private LAN Services over MPLS" (draft-ietf-l2vpn-vpls-ldp-05.txt)

VPLS Reference Model



No PIM peering between CEs and PEs
No PIM peering among PEs

VSI – Virtual Switch Instance

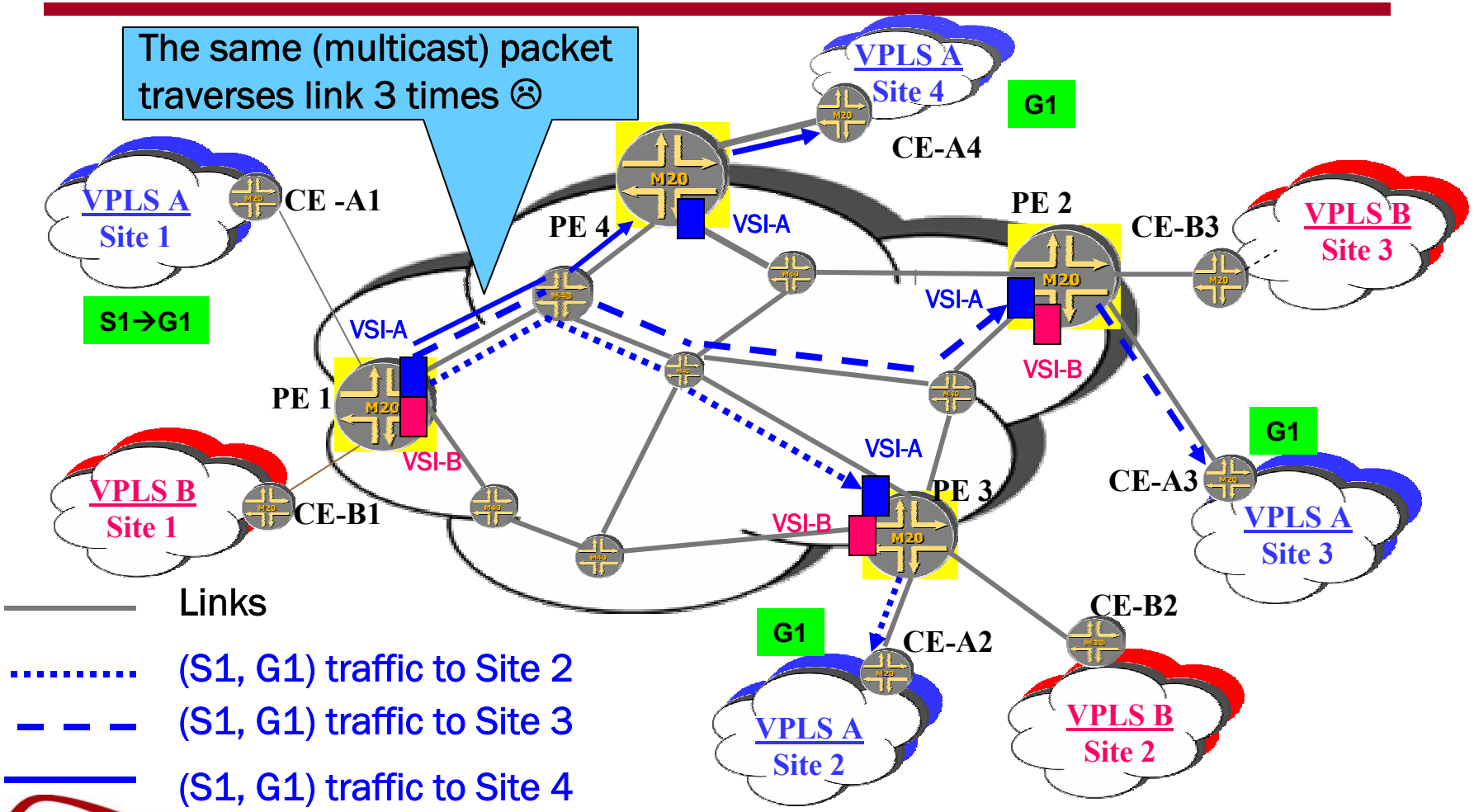
Implications on the state maintenance overhead on PE routers

- No PIM state in support of VPLS on PEs as:
  - No PIM peering between CEs and PEs
  - No PIM peering among PEs
- The state and the overhead of maintaining the state on PE routers in order to support IP multicast with VPLS is **insignificant** relative to the state and the overhead of maintaining the state in order to support unicast with VPLS
  - At least as long as optimizing bandwidth usage by avoiding sending VPLS multicast traffic to sites with no receivers is a non-goal
    - More on this later (see Slide 26)…

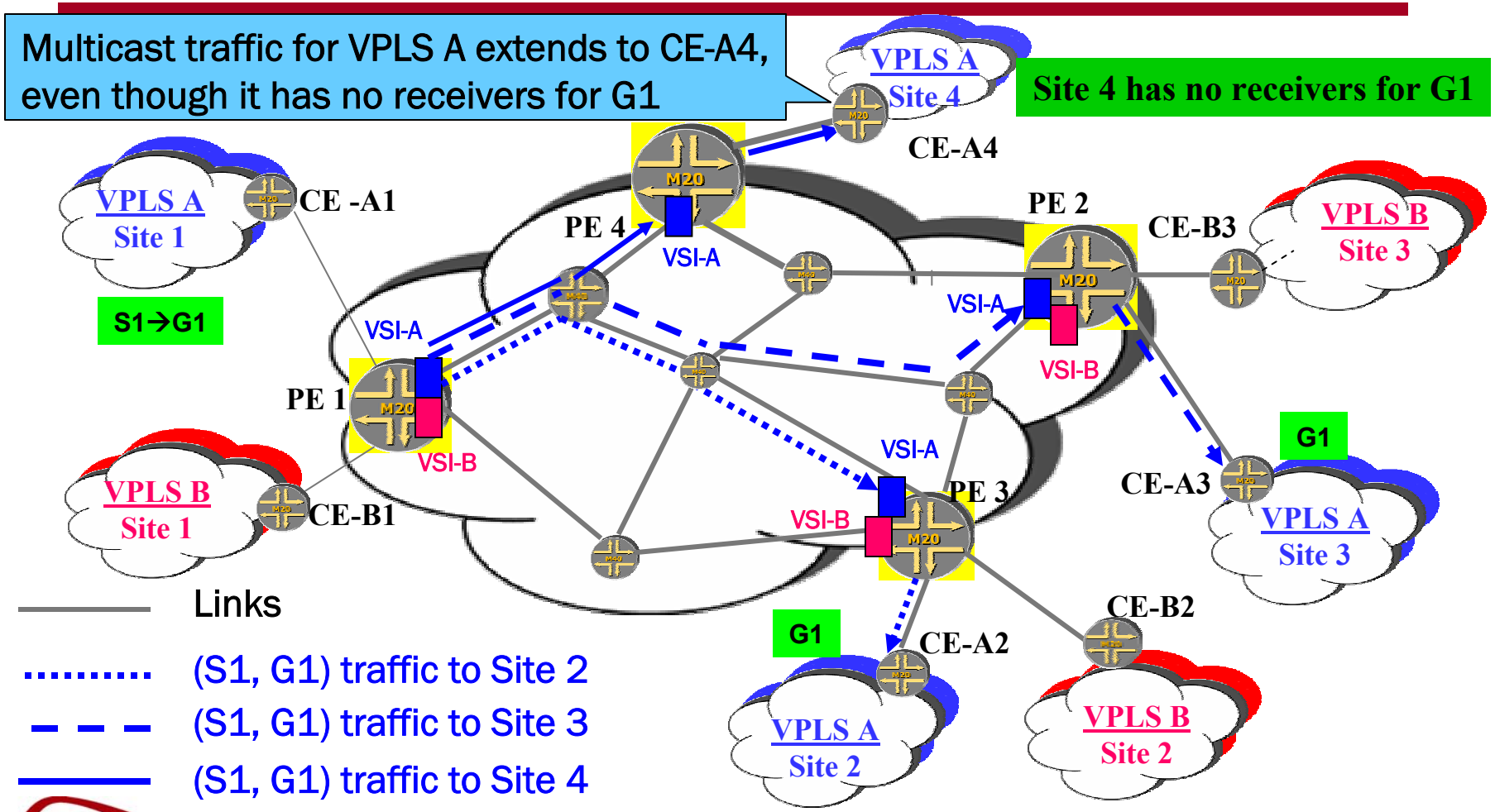
Forwarding VPLS multicast traffic within the service provider – Emulated LAN

- When an (ingress) PE receives an IP multicast packet from CE that belongs to a given VPLS, the PE sends the packet to all the other (egress) PEs that have sites of the same VPLS connected to them
- The packet is sent over the Emulated LAN associated with the VPLS
- Emulated LAN is realized by ingress replication – use collection of the existing (unicast) LSPs
  - From ingress PE to egress PEs
  - No additional state (beyond what is require by unicast) on P routers ☺
  - **May result in sending multiple copies of the same multicast packet over a given service provider link** ☹

Emulated LAN Ingress Replication: example



Sending multicast traffic to sites with no receivers: example



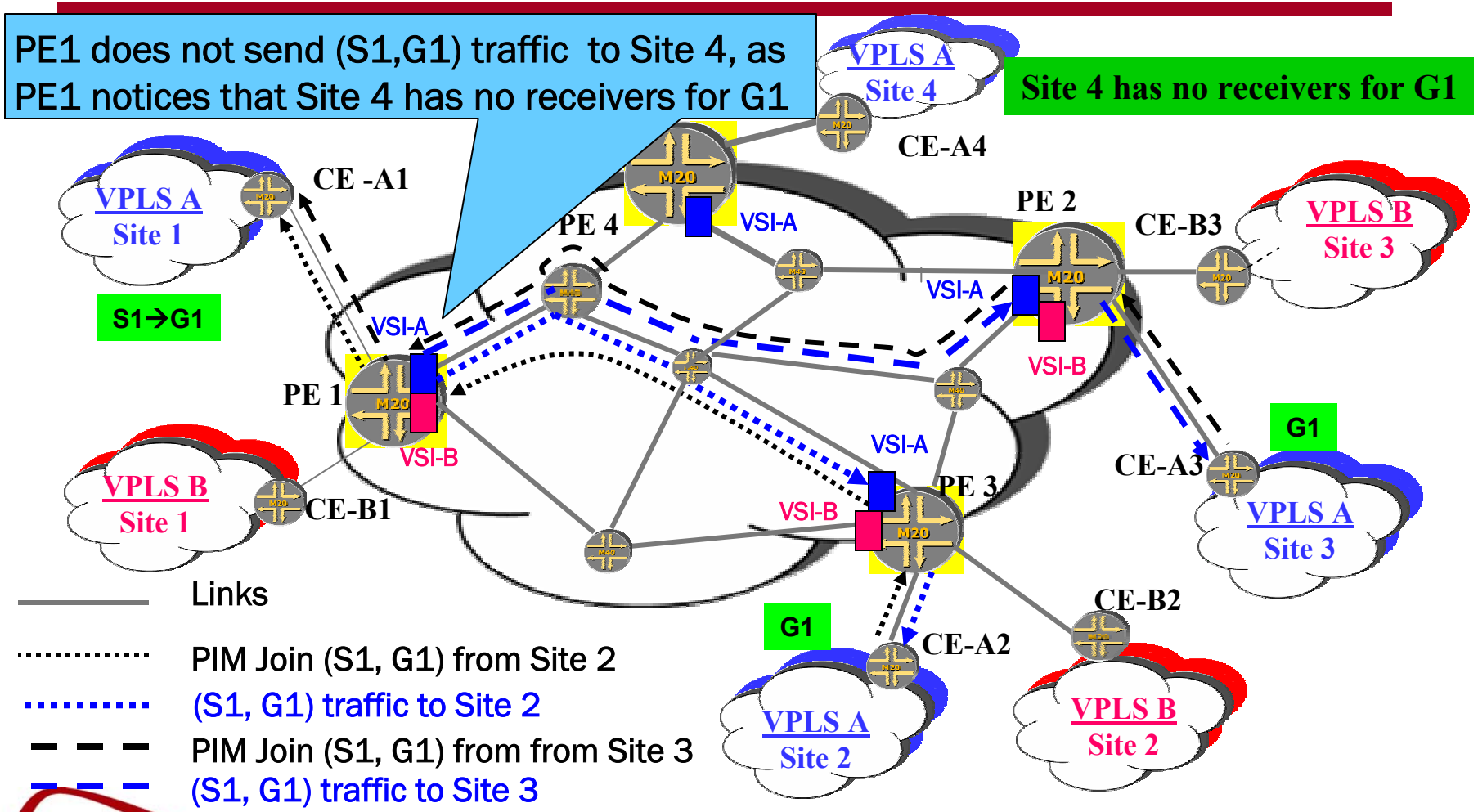
Sending multicast traffic to sites with no receivers

- As long as PE does not keep track of IP multicast receivers within each site of a given VPLS, PE has to send IP multicast traffic to all the sites within that VPLS
- As long as the ingress PE sends (multicast) traffic to all the sites within a VPLS, it is possible that the traffic will be delivered to the sites of that VPLS that have no receivers for the traffic
- **Suboptimal use of the service provider bandwidth due to sending IP multicast traffic to sites with no receivers is further compounded by the use of ingress replication for Emulated LAN ☹**

How to avoid sending multicast traffic to sites with no receivers – PIM/IGMP snooping

- Well-known approach used by Ethernet switches
  - An Ethernet switch determines whether a particular port has receivers for a given (S,G) by snooping on the PIM/IGMP messages received over that port
    - Requires to disable PIM Join suppression
- In the context of VPLS, PE has to snoop on PIM/IGMP messages received from:
  - all sites of that VPLS (directly) connected to the PE, AND
  - all the remote PEs that have members of that VPLS
- Just like with Ethernet switches, PIM/IGMP snooping in the context of VPLS requires to disable PIM Join suppression by VPLS customers

PIM snooping : example



PIM snooping – implications on the state maintenance on PE routers

- PE router has to maintain (S, G) state **at least** for all the (S,G) received from all the local CEs
  - E.g., assume PE with 1,000 CEs/sites, each VPLS site has at any given point in time on average receivers for 3 groups, PE has to maintain at least **3,000 (S,G) entries**
- PE router maintains (S, G) state by processing PIM Join messages received from (a) all sites of VPLSs connected to that PE, AND (b) all the remote PEs that have members of these VPLSs
  - E.g., assume PE router with 1,000 CEs/sites, each VPLS site has at any given point in time on average receivers for 3 group, each group is present on average in 10 sites, PE router has to **process ~300 PIM Join per second,** and **~900 (S, G) entries per second in a steady state** ☹
    - due to periodic PIM Join and PIM Join suppression

Agenda

- What are the (desirable) goals
- Support for multicast in 2547 VPNs - current proposals and their shortcomings
- Support for multicast in VPLS - current proposals and their shortcomings
- **Final remarks**

Multicast in 2547 VPNs vs multicast in VPLS

2547 VPNs:

- Focus on minimizing service provider bandwidth usage by (a) minimizing the amount of (multicast) traffic replication within the service provider, and by (b) avoiding sending traffic to the PE routers with no receivers
  - At the expense of additional state within the service provider

VPLS:

- Focus on minimizing state in the service provider routers by eliminating any multicast-related state in the P routers
  - At the expense of additional bandwidth usage within the service provider

Why the tradeoffs for multicast in 2547 VPNs are NOT the same as the tradeoffs for multicast in VPLS ?

Major shortcomings of the current proposals (not a complete list !!!)

- PIM neighbor state, and the overhead of its maintenance on PE routers (2547 VPNs):
  - e.g., **100,000 PIM neighbors per PE**, with **~3,300 PIM Hellos per sec** is problematic
- PIM state on P routers (2547 VPNs):
  - e.g., **1,000,000 multicast trees** per service provider with PIM-SSM is problematic
- Inefficient bandwidth use due to ingress replication as the ONLY mechanism to realize emulated LAN (VPLS):
  - although ingress replication is quite viable in the scenarios where the bandwidth of the multicast traffic is low or/and there is a sparse distribution of receiving PEs, such that the number of replications performed by the ingress PE on each outgoing interface for a particular customer multicast data packet is small
- Overhead of multicast routing state maintenance on PE with PIM (VPLS with PIM snooping):
  - e.g., need to process in a steady state **300 PIM Join per second,** and **900 (S, G) entries per second** is problematic

On the subject of thrust

- Is it possible to build and operate a system that would handle the control and data plane overhead of the current solutions to multicast in 2547 VPN and VPLS ?
- "With enough thrust pigs will fly"
- Does it mean that we should be content with the solutions that require plenty of "thrust" ?
  - Are such solutions Good Enough ?
- If not, what else ?

On being constructive

- Settle on a "stop-gap" solution
  - Needed to address immediate demand for multicast support in 2547 VPN and VPLS
- Work on addressing major shortcomings of the current proposals
- Pay attention to operational complexity
- Make sure the benefits of supporting multicast in 2547 VPNs and VPLS justify the cost
- Explore (and take advantage of) commonalities between supporting multicast for 2547 VPNs and for VPLS
  - As the problems are similar

Suggested reading:

- "Overview of IP Multicast in a Multi-Protocol Label Switching (MPLS) Environment" (RFC3353)
- "Multicast in MPLS/BGP IP VPNs" (draft-rosen-vpn-mcast-07.txt)
- "Base Specification for Multicast in BGP/MPLS VPNs"(draft-raggarwa-l3vpn-2547-mvpn-00.txt)
- "Framework for Layer 2 Virtual Private Networks (L2VPNs)" (draft-ietf-l2vpn-l2-framework-05.txt)
- "Virtual Private LAN Service" (draft-ietf-l2vpn-vpls-bgp-02.txt)
- "Virtual Private LAN Services over MPLS" (draft-ietf-l2vpn-vpls-ldp-04.txt)
- "Multicast in BGP/MPLS VPNs and VPLS" (draft-raggarwa-l3vpn-mvpn-vpls-mcast-00.txt)