Deployment Architecture, Requirements and Solutions for Multicast over MPLS-based Core

Zafar Ali, IJsbrand Wijnands and John Evans

Cisco Systems

Futurenet 2008

April 2008.

Agenda

- Multicast Service Requirements
- Multicast Solutions Space
  - P-Tree Building
  - Exchanging Customer mcast routes
  - Auto-discovering peering PE-es
  - Encapsulation
- Migrating Path to Label Switched Multicast Core
- Summary

Diversity, Diversity, and Diversity!

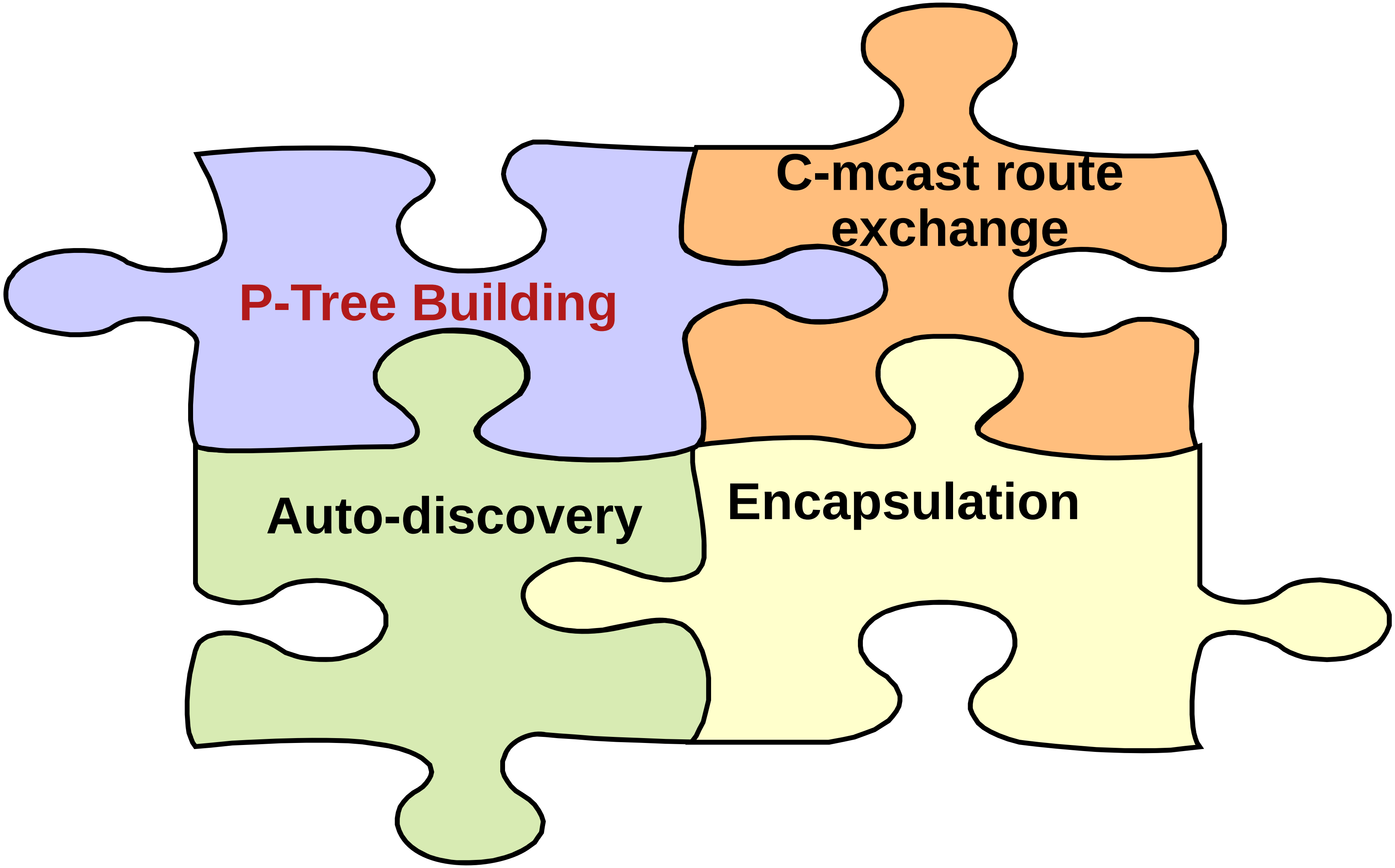
- Diverse applications for label switched multicast with diverse requirements
- Some typical applications are:
  - Video transport (Contribution and Primary Distribution)
  - Secondary Video Distribution, e.g., IPTV
    - IP multicast distribution from centralized servers
  - Managed Enterprise mVPN Services
- Diverse requirements within the same application, depending on deployment specifics.
- Stringent video SLAs

How requirement diversity influences the solution space?

Agenda

- Multicast Service Requirements
- Multicast Solutions Space
- Migrating Path to Label Switched Multicast Core
- Summary

Components of Multicast Solutions Space



P-Tree Building

C-mcast route exchange

Auto-discovery

Encapsulation

P-Tree Building Tool Kit

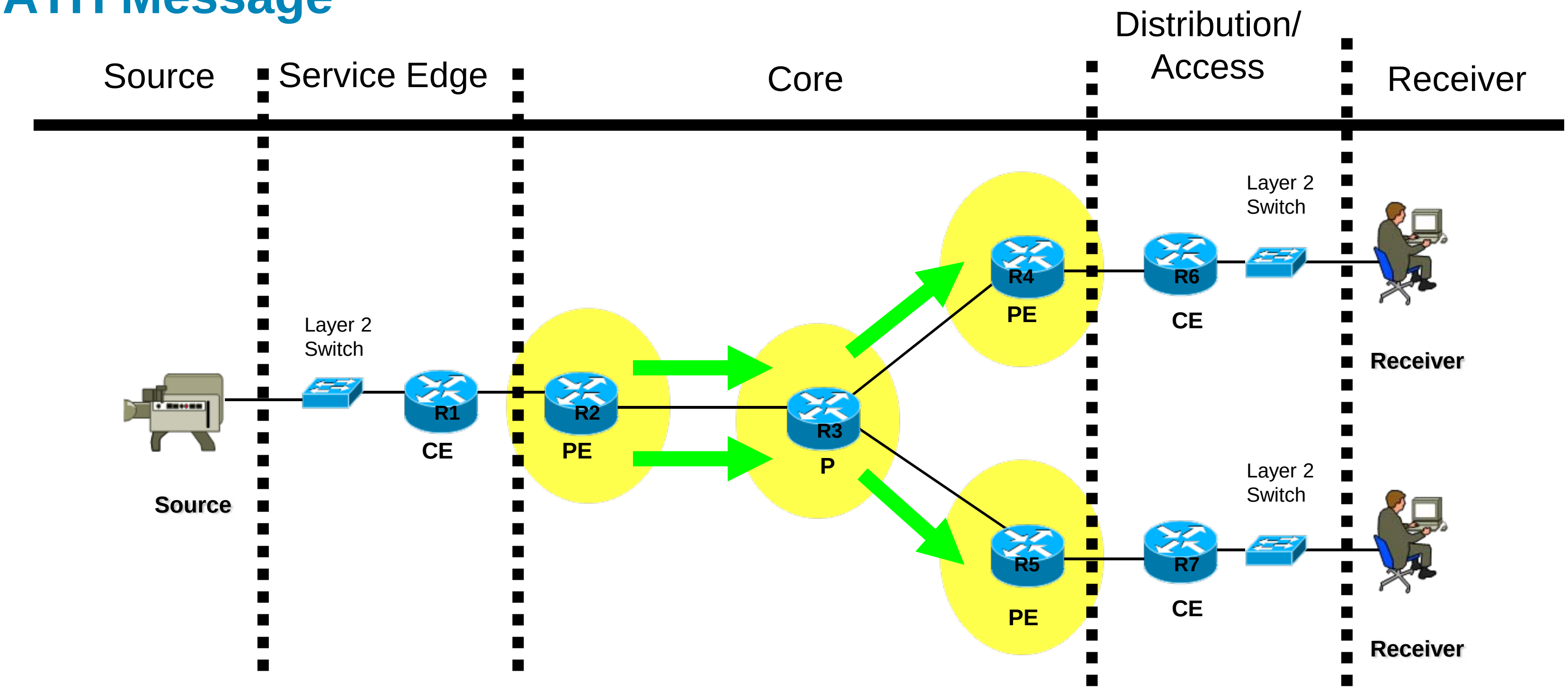
P-Tree Types

- Point-to-Multi Point (P2MP)
- Multi Point-to-Multi Point (MP2MP)

P-Tree Building Protocols

- RSVP-TE
  - Extension to RSVP-TE to build P2MP trees
  - Source Driven (unlike PIM)
  - Supports Traffic Engineering
- Multicast LDP (mLDP)
  - Extension to LDP to build P2MP and MP2MP Trees
  - Very similar to PIM
    - Receiver Driven
- PIM (Not focus of this presentation)

P2MP Tunnel Setup (RSVP-TE Non-Aggregated Mode): PATH Message



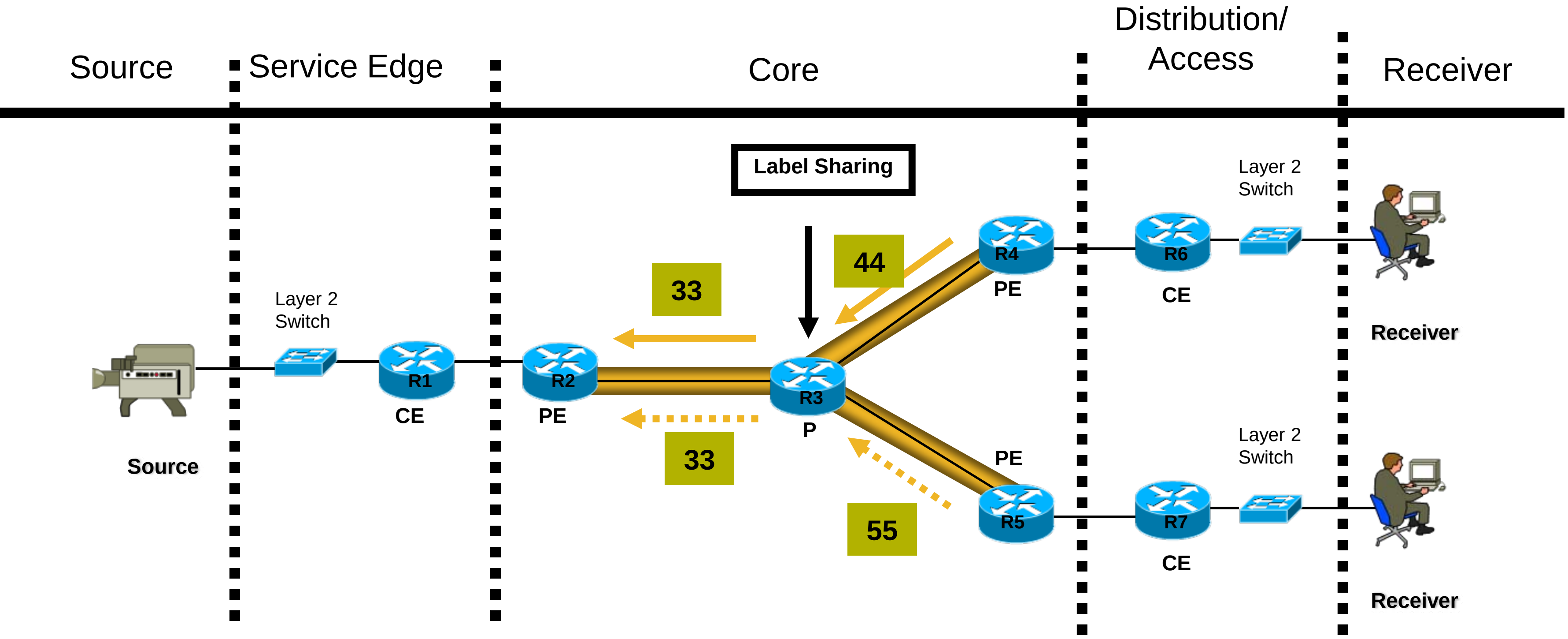
Non-Aggregated Mode: Headend sends one PATH message per destination.

P2MP 1st sub-lsp PATH : ERO: R2-R3-R4

P2MP 2nd sub-lsp PATH : ERO: R2-R3-R5

- RSVP-TE also supports aggregated mode, where a single Path message can carry all sub-LSP information for all destinations.

P2MP Tunnel Setup (RSVP-TE Non-Aggregated Mode): RESV Message



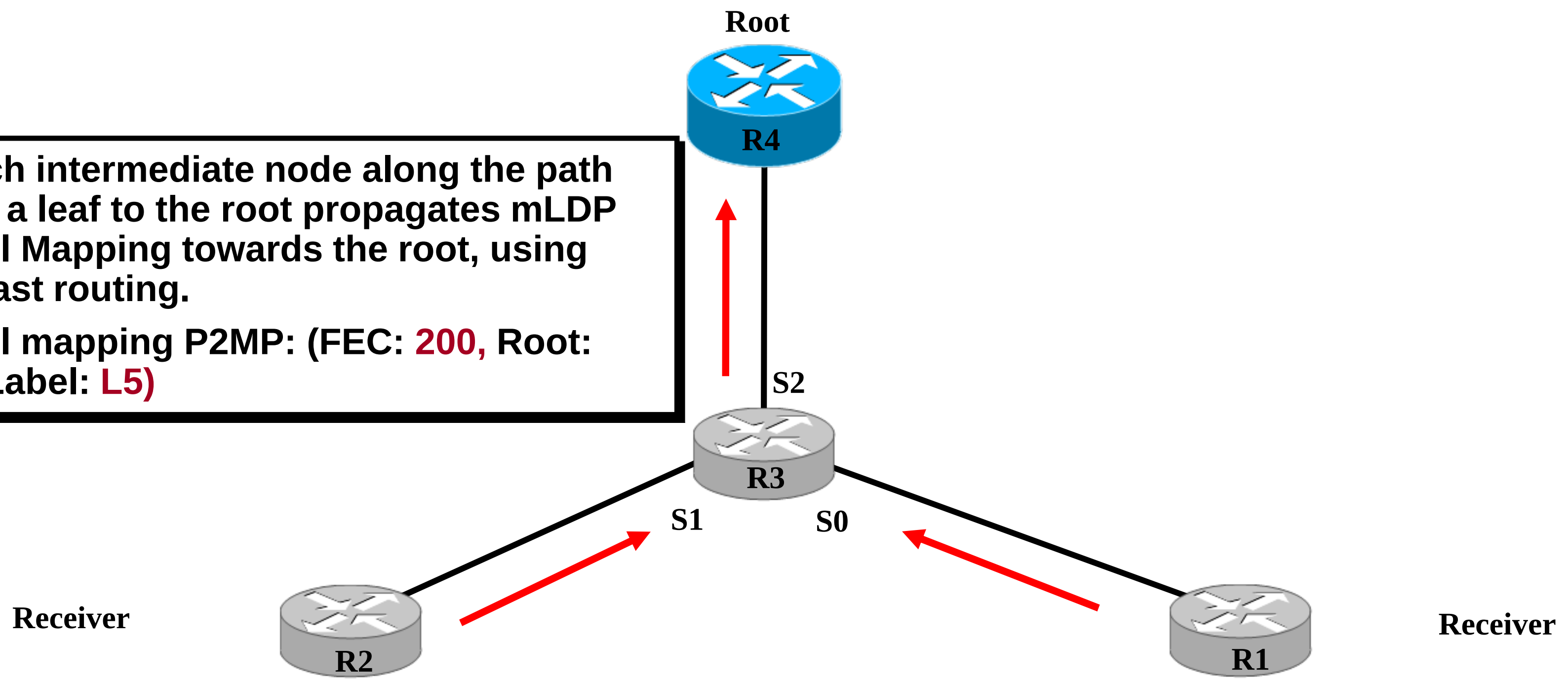
<u>RESV Messages</u> are sent by Tailend routers; Communicates labels & reserves BW on each link

← RESV Msg Initiated by R4

⇠ RESV Msg Initiated by R5

55 Label Advertisement carries in the RESV Message

P2MP LSP setup using mLDP



- Each intermediate node along the path from a leaf to the root propagates mLDP Label Mapping towards the root, using unicast routing.

Label mapping P2MP: (FEC: 200, Root: R4, Label: L5)

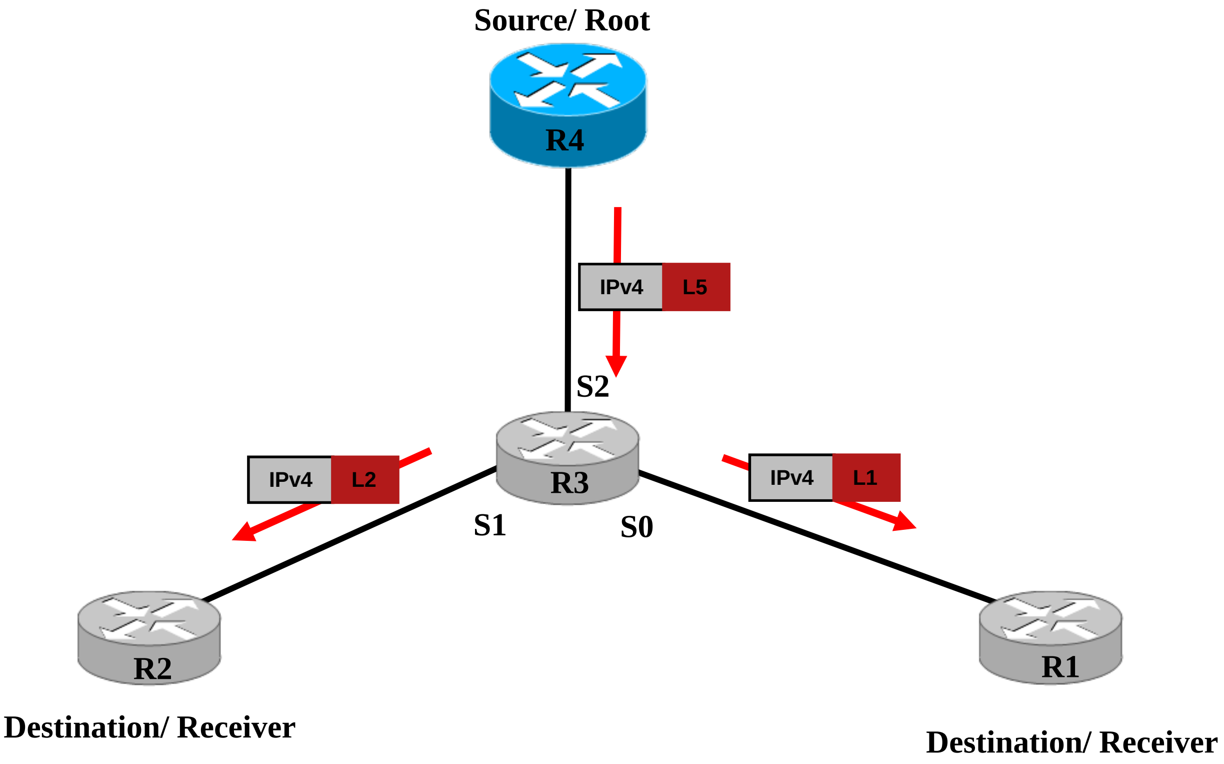
- Each leaf node initiates P2MP LSP setup by sending mLDP Label Mapping message towards the root, using unicast routes.
- Label Mapping message carries the identity of the LSP, encoded as P2MP FEC.

Label mapping P2MP: (FEC: 200, Root: R4, Label: L2)

- Each leaf node initiates P2MP LSP setup by sending mLDP Label Mapping message towards the root, using unicast routes.
- Label Mapping message carries the identity of the LSP, encoded as P2MP FEC.

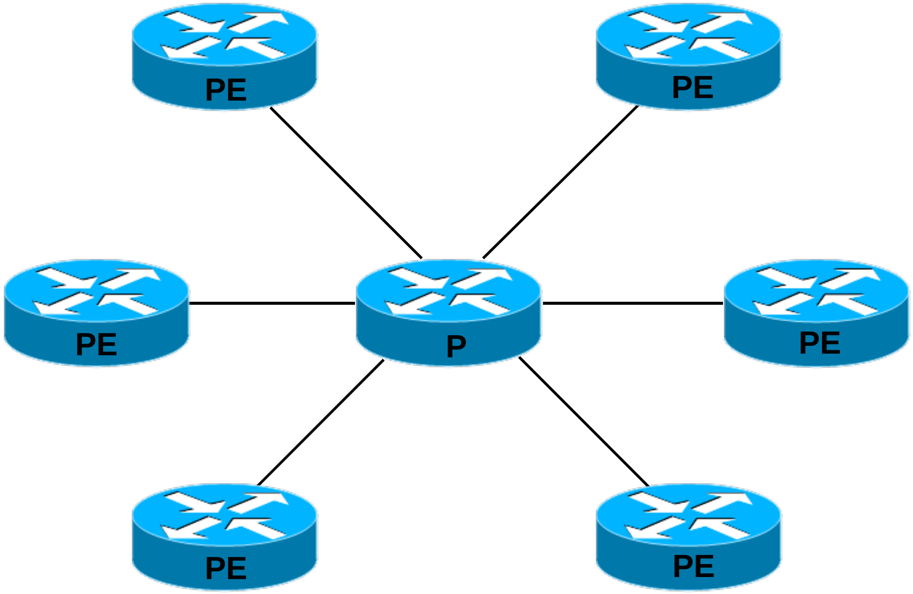
Label mapping P2MP, FEC: 200, Root: R4, Label: L1)

P2MP LSP (Data Plane)



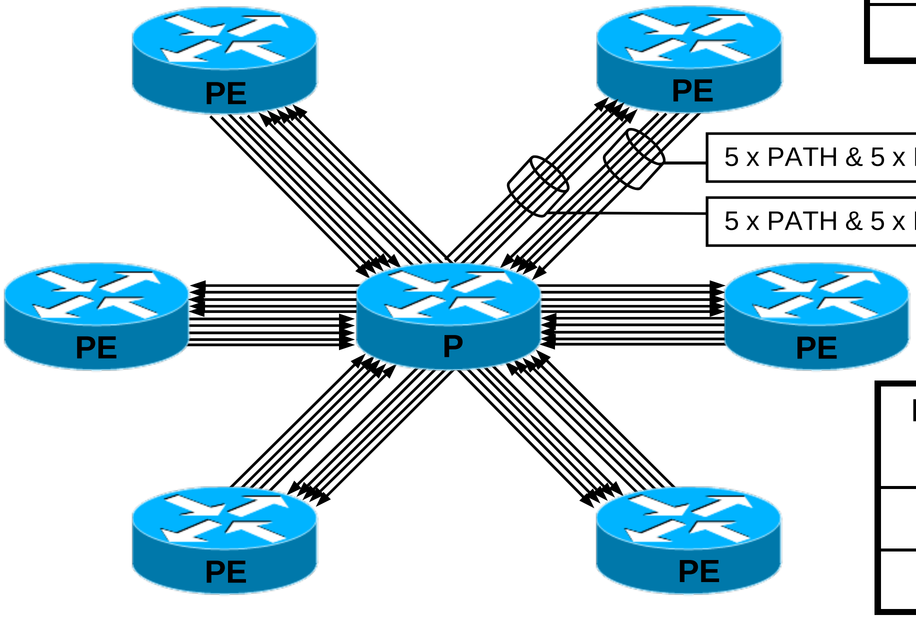
Comparison Basis for P-Tree Type and Protocol

- Suppose we are building a emulated LAN between 6 PE routers.
- To compare we connect the 6 PE's via a single core router, we see how much protocol updates, state and labels are need to build the E-LAN.
- Note, in real life there will probably be more then one P router and the amount of state will be distributed across multiple P routers.
- It should be noted that big-O scaling characteristics remains same for different tree types.



Full Mesh P2MP RSVP-TE

- Head-end driven tree setup
- Assuming non-aggregated signaling.



5 x PATH & 5 x RECV messages to other PE's

5 x PATH & 5 x RESV received

6 PE Routers

RSVP-TE P2MP	State Sub-LSP	Local Labels	Protocol msg IN/OUT
PE	5	5	10/10
P	30	6	60/60

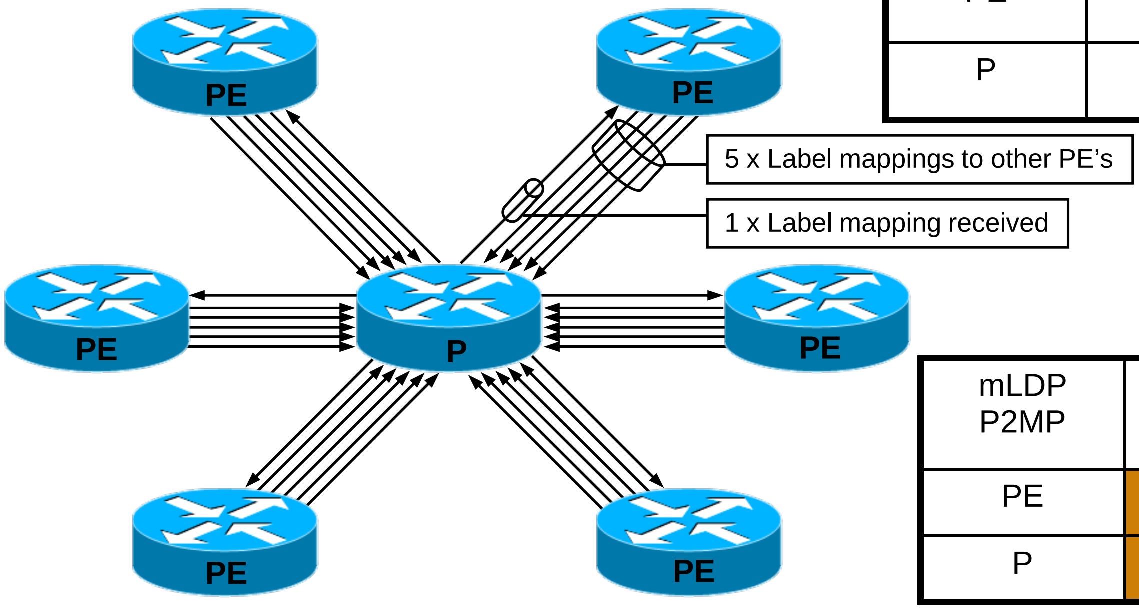
1K PE Routers

RSVP-TE P2MP	State Sub-LSP	Local Labels	Protocol msg IN/OUT
PE	1K	~1K	~2K/~2K
P	~1M	1K	~2M/~2M

- **O(PE^2) Control Plane States**
- **O(PE) Data Plane States**
- **O(PE^2) Protocol Messaging**
- **These asymptotic characteristics are independent of Tree Type.**

Full Mesh P2MP mLDP

- Receiver driven tree setup



5 x Label mappings to other PE's

1 x Label mapping received

6 PE Routers

mLDP P2MP	State FEC	Local Labels	Protocol msg IN/OUT
PE	6	5	1/5
P	6	6	30/6

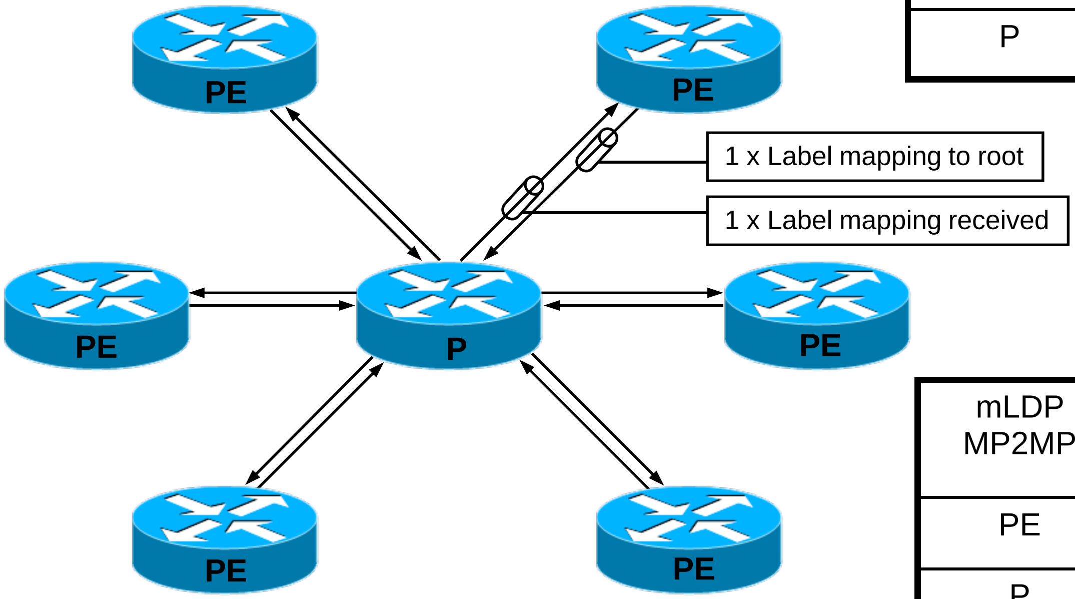
1K PE Routers

mLDP P2MP	State FEC	Local Labels	Protocol msg IN/OUT
PE	1K	~1K	1/~1K
P	1K	1K	1M/1K

- **O(PE) Control Plane States**
- **O(PE) Data Plane States**
- **O(PE^2) Protocol Messaging**
- **These asymptotic characteristics are independent of Tree Type.**

Single MP2MP mLDP

- Receiver driven tree setup
- P is the root of the MP2MP LSP



1 x Label mapping to root

1 x Label mapping received

6 PE Routers

mLDP MP2MP	State FEC	Local Labels	Protocol msg IN/OUT
PE	1	1	1/1
P	1	6	6/6

1K PE Routers

mLDP MP2MP	State FEC	Local Labels	Protocol msg IN/OUT
PE	1	1	1/1
P	1	1K	1K/1K

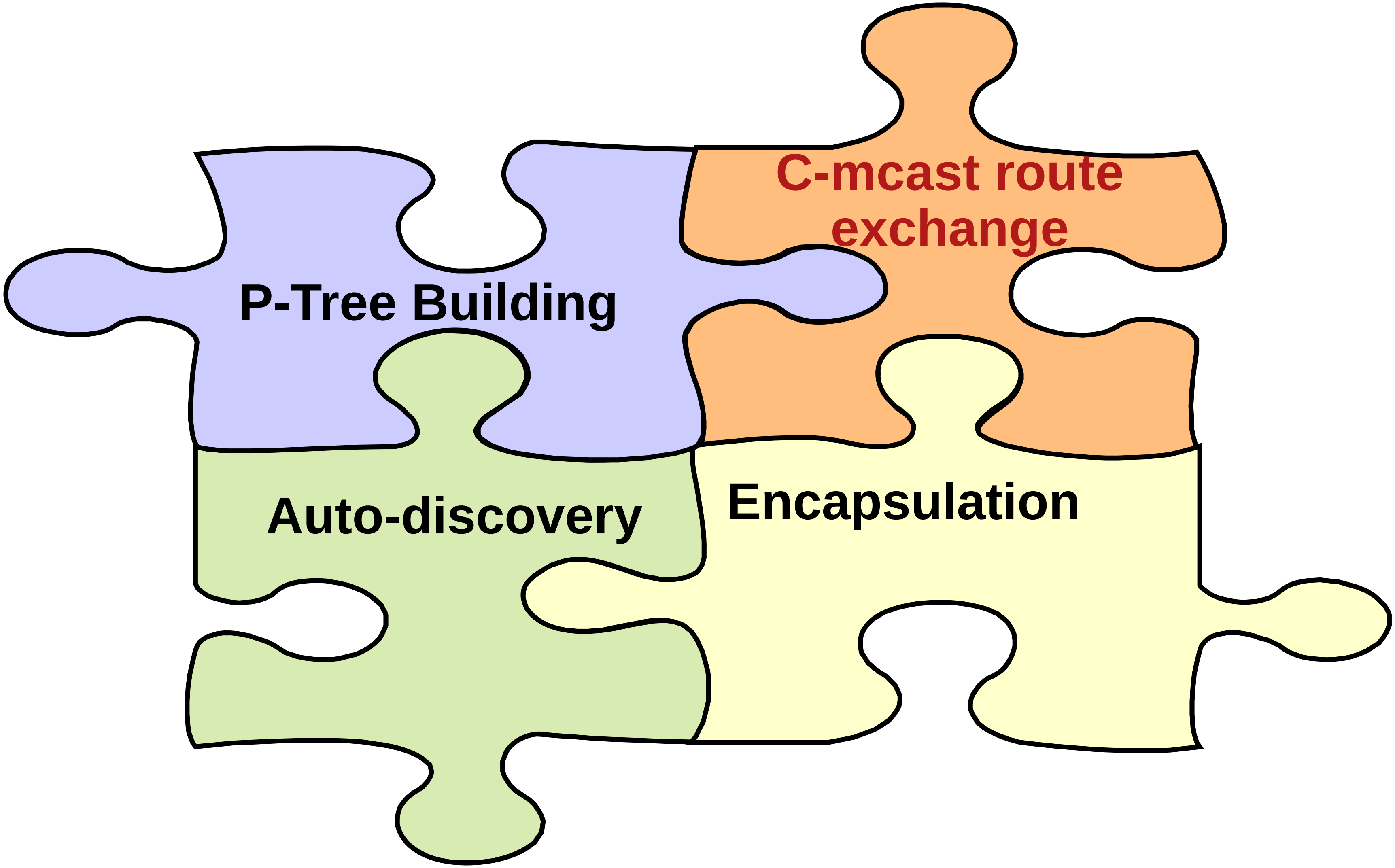
- **O(1) Control Plane States**
- **O(1) Data Plane States**
- **O(PE) Protocol Messaging**

Core Tree Protocol Selection

- mLDP is more scalable protocol then RSVP-TE (even if RSVP-TE aggregated signaling mode is used).
- RSVP-TE provides Traffic Engineering functionality.
- MP2MP trees are more scalable than P2MP trees.
- mLDP supports signaling for MP2MP trees.
- RSVP-TE does not supports signaling for MP2MP trees.
- Grafting and pruning operations are more expensive in RSVP-TE, then in mLDP.

• No one size fit all.
• Use of RSVP or mLDP depends on application requirements

Components of Multicast Solutions Space



Multicast Signaling (Exchanging Customer mcast routes)

- Mechanics used for customer mcast routes exchange is independent of core tree building and auto discovery methods.
- In draft-ietf-l3vpn-2547bis-mcast-06 two options are specified:
    - PIM
    - BGP

Use of PIM for exchanging customer mcast routes

- Used for PIM for exchanging c-mcast routes does not require PIM in the core.
- Currently deployed, proven.

Use of BGP for exchanging customer mcast routes

- New addition to multicast world, unproven for this application.
- Even when BGP is used for exchanging c-mcast routes, PEs still run per-VPN PIM instance (PIM over PE-CE link).
- Translates customer PIM Join/Prunes to BGP by encoding PIM join and prune info in a new MVPN AFI/SAFI.
- RD is required in order to uniquely identify the <C-Source, C-Group> when different MVPNs have overlapping address spaces.
- Mechanics similar to RFC4364, e.g., Route Reflector may be used.
- New BGP procedures are needed to handle PIM-SM.

  BGP needs to emulate PIM sparse-mode!

BGP vs. PIM for C-mcast Route Exchange: Comparison Basis

How can we use PIM and BGP for exchanging customer routes, for the following types of trees?

- Emulated LAN (E-LAN) (all PEs to every PEs)
- Selective-PMSI (one PE to a select subset of PEs)
- Partitioned E-LAN

Emulated LAN (E-LAN) or MI-PMSI

- From all PEs to every PEs
- Known as Multidirectional Inclusive Provider Multicast Service Instance (MI-PMSI). Also known as default-MDT.
- May use a full mesh of P2MP LSPs or a single MP2MP LSP.

Selective-PMSI

- From one PE to a select subset of PEs.
- Also known as data-MDT.
- Uses a single P2MP LSP per ingress PE.

Partitioned E-LAN

- Combination between selective-PMSI and E-LAN.
- This is a is a dynamic version of the existing PIM based MVPN deployments using multicast domain model, as specified in draft-ietf-l3vpn-2547bis-mcast-06.
- We setup a tree per ingress PE!
- The tree is a MP2MP LSP, so bidirectional!
- The root of the MP2MP is the ingress PE.
- Supports Anycast sources.
- Supports bidirectional Multicast without the need of upstream assigned labels.

BGP vs. PIM for C-mcast Route Exchange Over E-LAN Tree

- C-mcast Route Exchange Over E-LAN (MI-PMSI) needs to support:
  - Customer PIM-SM, PIM-SSM, PIM-Bidir.
  - Resolve duplicate forwarders on the LAN.
  - Elect a Designated Forwarder on the LAN.

- No modifications necessary to PIM.
  - Solves duplicate forwarders using asserts
  - Solves DF using PIM DF election procedures.
- Supports PIM-SM, PIM-SSM and PIM-Bidir

- BGP needs to implement extensions in 2547bis-mcast.
- BGP needs to implement sparse-mode procedures to emulate PIM sparse-mode!
- BGP-SM has some differences from PIM-SM, impact remains to be seen.

BGP vs. PIM for C-mcast Route Exchange Over Partitioned E-LAN Tree

- Multicast signalling over Partitioned E-LAN needs to support:

  Customer PIM-SM, PIM-SSM, PIM-Bidir.

  No duplicate forwarder detection necessary.

  No PIM DF election necessary, the root is the DF.

- No modifications necessary to PIM.
- Supports PIM-SM, PIM-SSM and PIM-Bidir

- BGP needs to implement extensions in 2547bis-mcast.
- BGP needs to implement sparse-mode procedures to emulate PIM sparse-mode!
- BGP-SM has some differences from PIM-SM, impact remains to be seen.

BGP vs. PIM for C-mcast Route Exchange Over Selective-PMSI Tree

- Multicast signalling over Selective-PMSI needs to support:

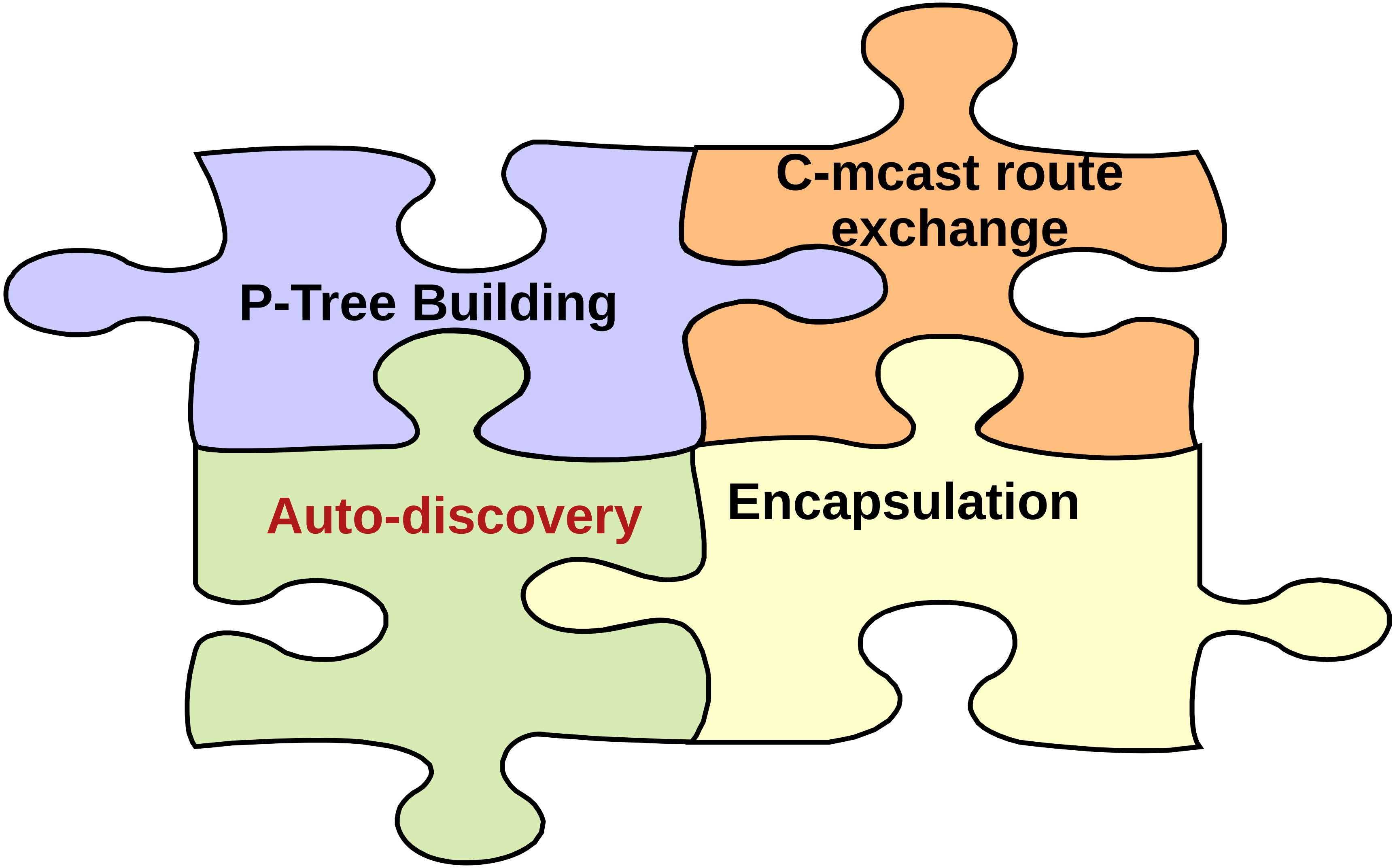
  Bidirectional multicast is not supported

  No duplicate forwarder detection necessary.

- As this is a uni-directional tree, PIM cannot run without some modifications.
- The required modifications that are being discussed in IETF.

- BGP needs to implement 2547bis-mcast.
- BGP needs to implement sparse-mode procedures to emulate PIM sparse-mode! BGP-SM has some differences from PIM-SM, impact remains to be seen.

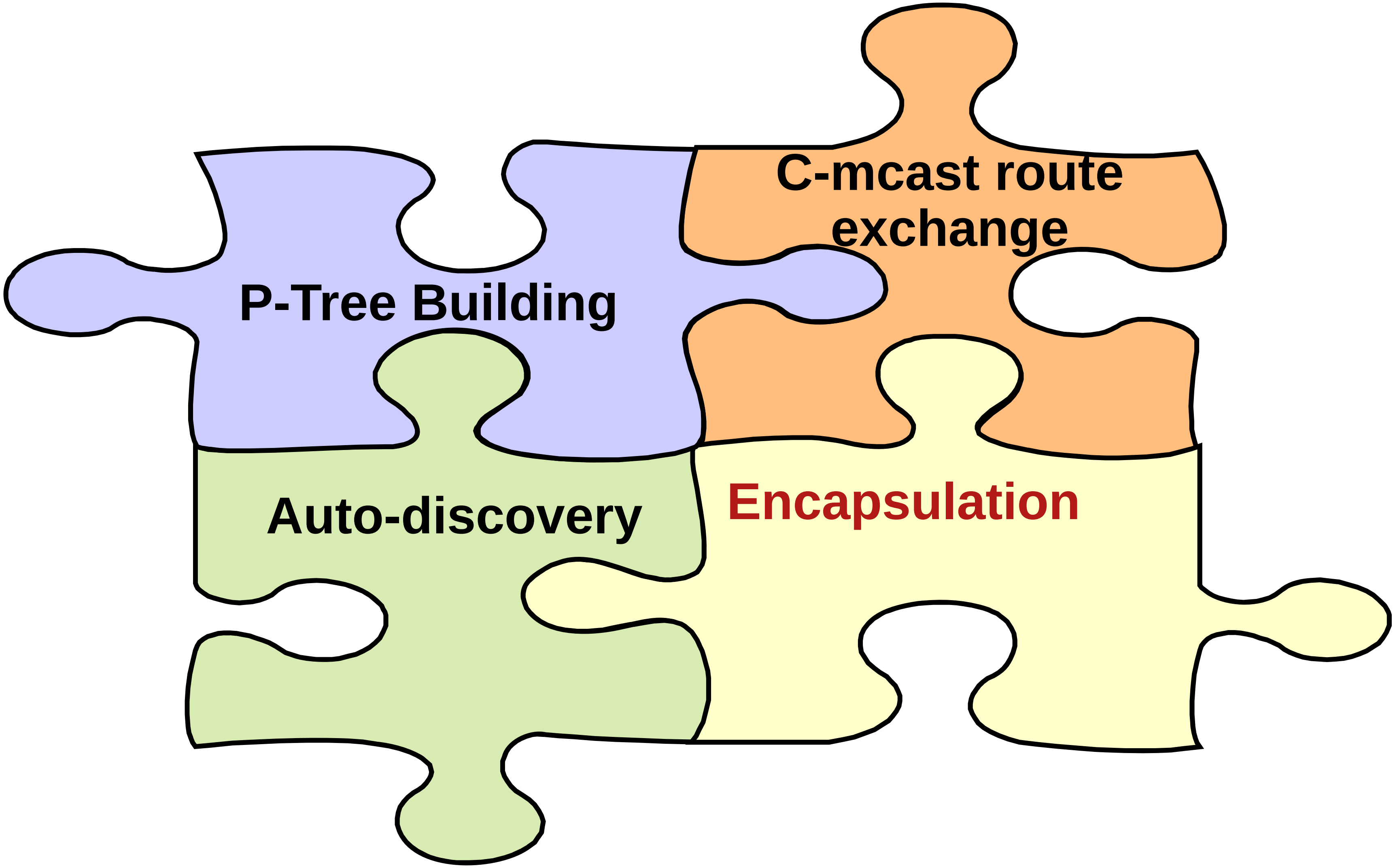
Components of Multicast Solutions Space



Auto Discovering Peering PE-es

- Auto Discovery is a process of discovering which PEs support which VPNs.
- Again, auto discovery mechanism is independent of core tree building and customer mcast routes exchange methods.
- Candidate protocols are PIM and BGP.
- If PIM is also P-Tree building protocol, it makes sense to use it also for auto discovery (as PIM is leave driven).
- BGP is also good for auto discovery for future deployments, where there is no PIM in the core.

Components of Multicast Solutions Space



Encapsulation

- There are 2 tunnel encapsulation options:
  - GRE (Currently Deployed)
  - MPLS (Focus of this presentation)

Agenda

- Multicast Solutions Space

  P-Tree Building

  Exchanging Customer mcast routes

  Auto-discovering peering PE-es

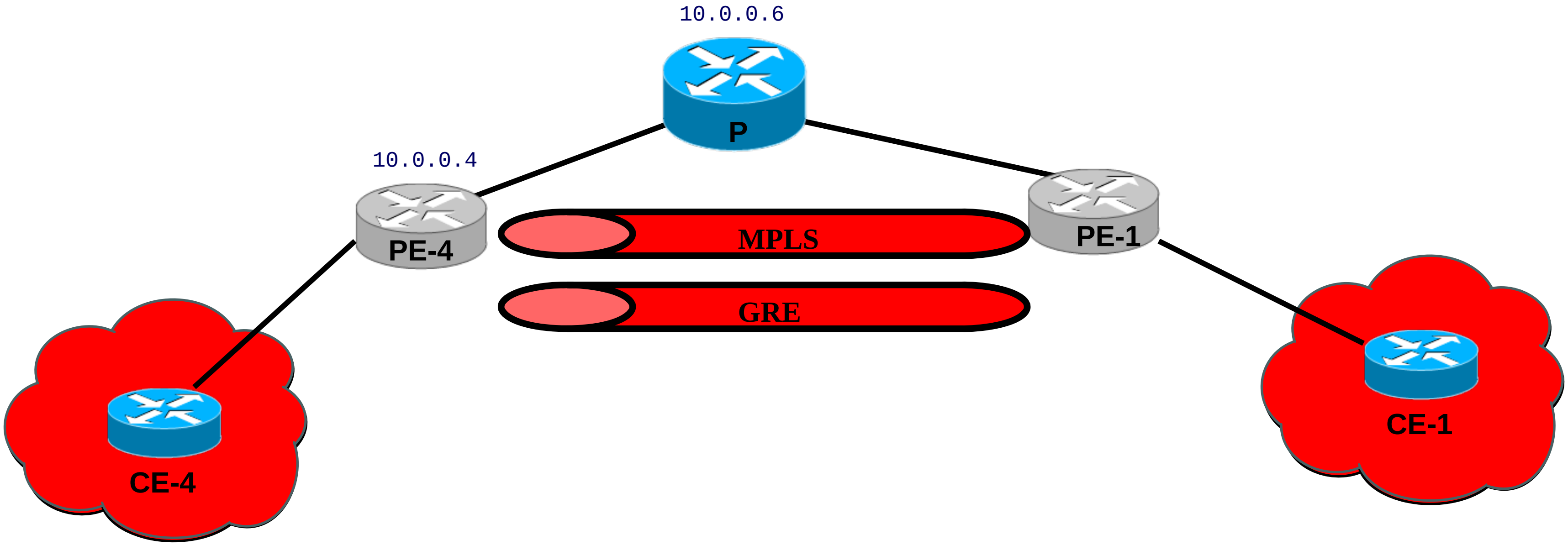
  Encapsulation

- Migrating Path to Label Switched Multicast Core
- Summary

What are we changing?

- To understand migration path, we need to understand what are we changing?
  - Changing encapsulation (GRE to MPLS)
  - P-tree building protocol (from PIM to mLDP or RSVP-TE)
- Change in Tree building Protocol and encapsulation method does not require a change in method used today to exchange c-mcast routes (which is PIM).
- PE routers still need to run PIM (Even when P routers become PIM-free).

MVPN During Migration



- To facilitate migration, MPLS and GRE tunnels can co-exists side-by-side.
- PE’s will see same PIM neighbor over different Tunnels.
- PE’s may select the Tunnel of their preference.

Use of BGP: Summary

- New and experimental use of BGP
  - First use of BGP where BGP events are caused by end user actions rather than topology changes.
- Rate of change:
  - BGP is great for steady state, but not so great when there is high rate of change.
  - Many c-mcast exchange operations are transactional, which is not BGP’s strength.
- Strict “join latency” requirements does not suite BGP so well.
- BGP needs to implement sparse-mode procedures to emulate PIM sparse-mode! BGP-SM has some differences from PIM-SM, impact remains to be seen.
- Impact on non-multicast use of BGP.
- This adds complexity to BGP solution.
- Difficult to migrate from existing multicast deployments.

• BGP is good for auto-discovery (when P routers become PIM-free).
• Use of BGP for c-mcast route exchange during migration to label switched multicast core is neither desirable nor required.

Use of PIM: Summary

- Already deployed and proven.
- Offers easiest migration path from existing deployments.
- Works without any changes (in most cases).
- Work is also in progress to support PIM over Selective-PMSI trees.
- Being soft-state, scaling is a limitation.
  - We have not seen these limitations in current deployments.
  - Work is in progress at IETF to address PIM scalability, e.g., PIM over TCP proposal.

• Use of PIM for c-mcast route exchange during migration to label switched multicast core is provides easiest migration path.

Agenda

- Multicast Solutions Space
  - P-Tree Building
  - Exchanging Customer mcast routes
  - Auto-discovering peering PE-es
  - Encapsulation
- Migrating Path to Label Switched Multicast Core
- Summary

Summary

- Multicast service requirements are extremely diverse.
- No one size fits all, applies here.
- Many factors need to be consider in selecting a specific solution, including:

    Application requirements.

    Capitalizing on current deployment mVPN experience.

    Finding easiest migration path to label switched multicast core.