MPLS 2008
11TH ANNUAL CONFERENCE

# Multicast in MPLS: Considerations in Migrating to a Label Switched Multicast Core

**Zafar Ali, IJsbrand Wijnands and John Evans**
**Cisco Systems**
**{zali, ice, joevans}@cisco.com**

**www.mpls2007.com**


# 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!

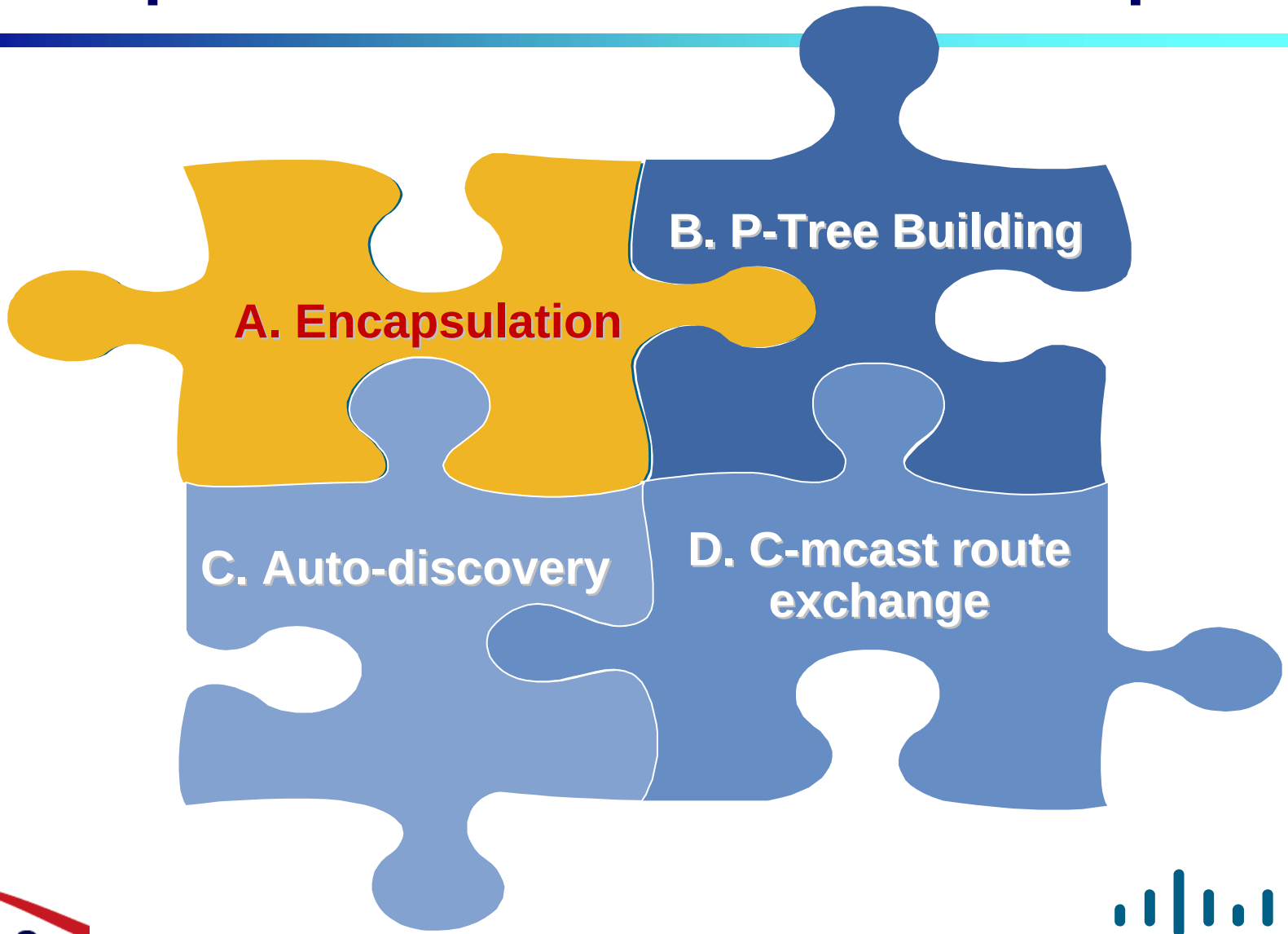
- Diverse applications for label switched multicast with diverse requirements.
- Some typical applications are:
  - Video transport (Contribution and Primary Distribution)
  - Secondary Video Distribution, e.g., IPTV
  - Video contribution, e.g. studio to studio
  - Managed Enterprise mVPN Services
- 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

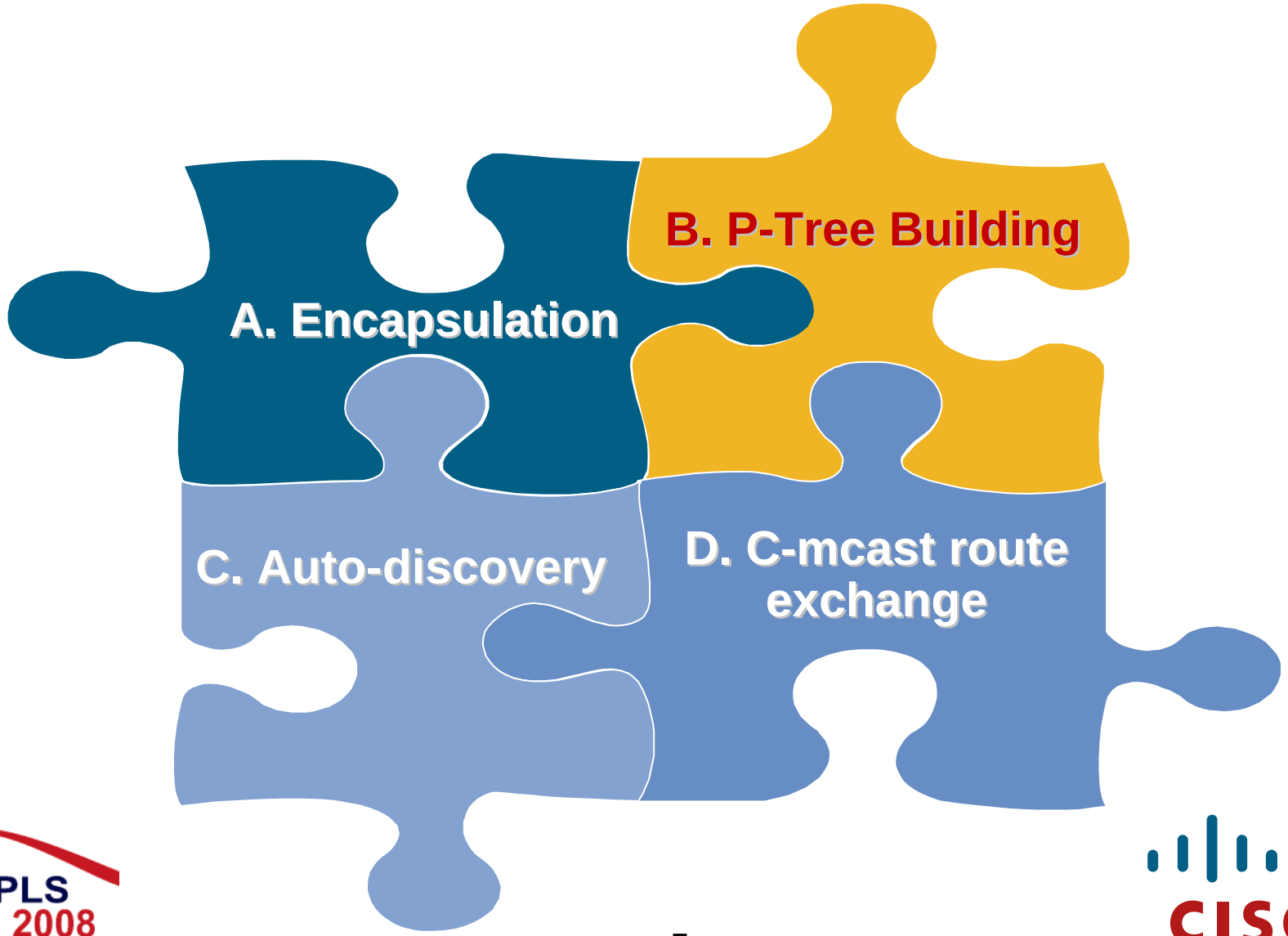
# Components of Multicast Solutions Space



# Encapsulation

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

# Components of Multicast Solutions Space



# P-Tree Building Tool Kit

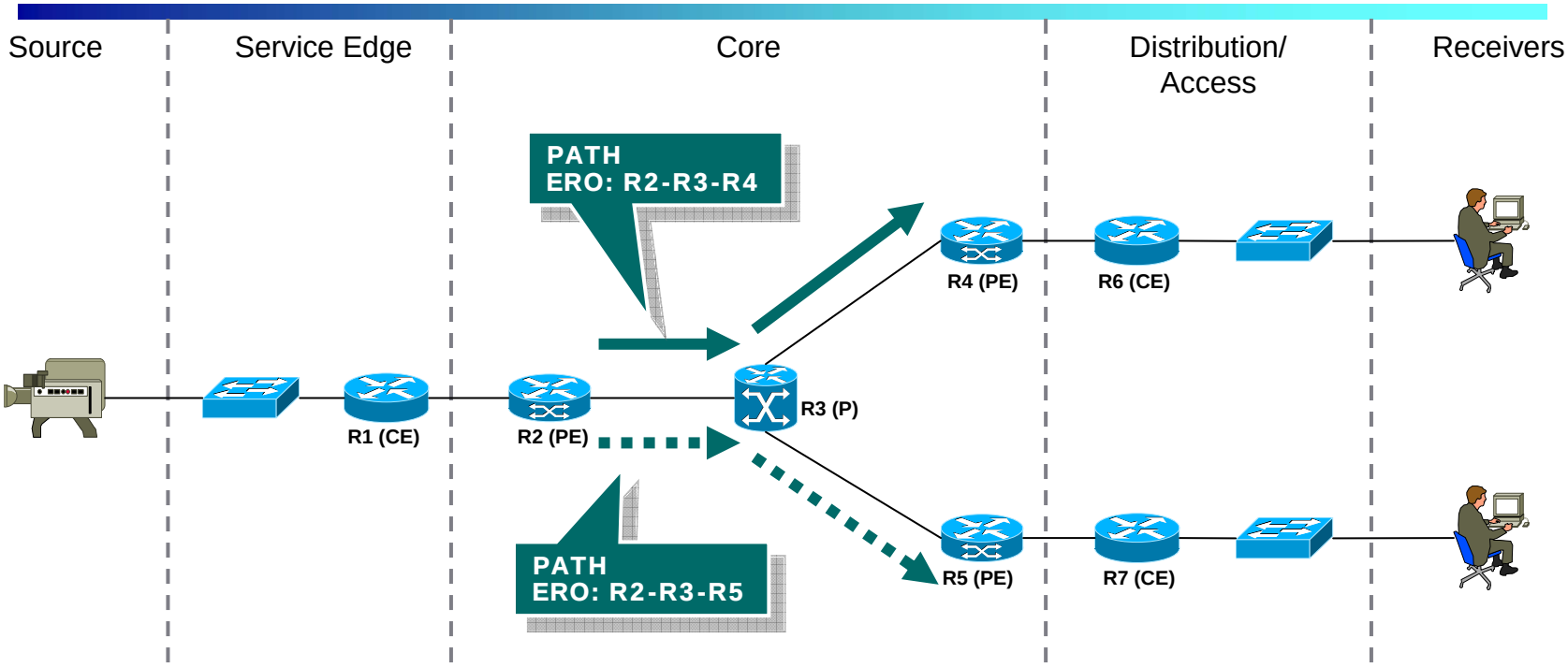
## P-Tree Types

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

## P-Tree Building Protocols

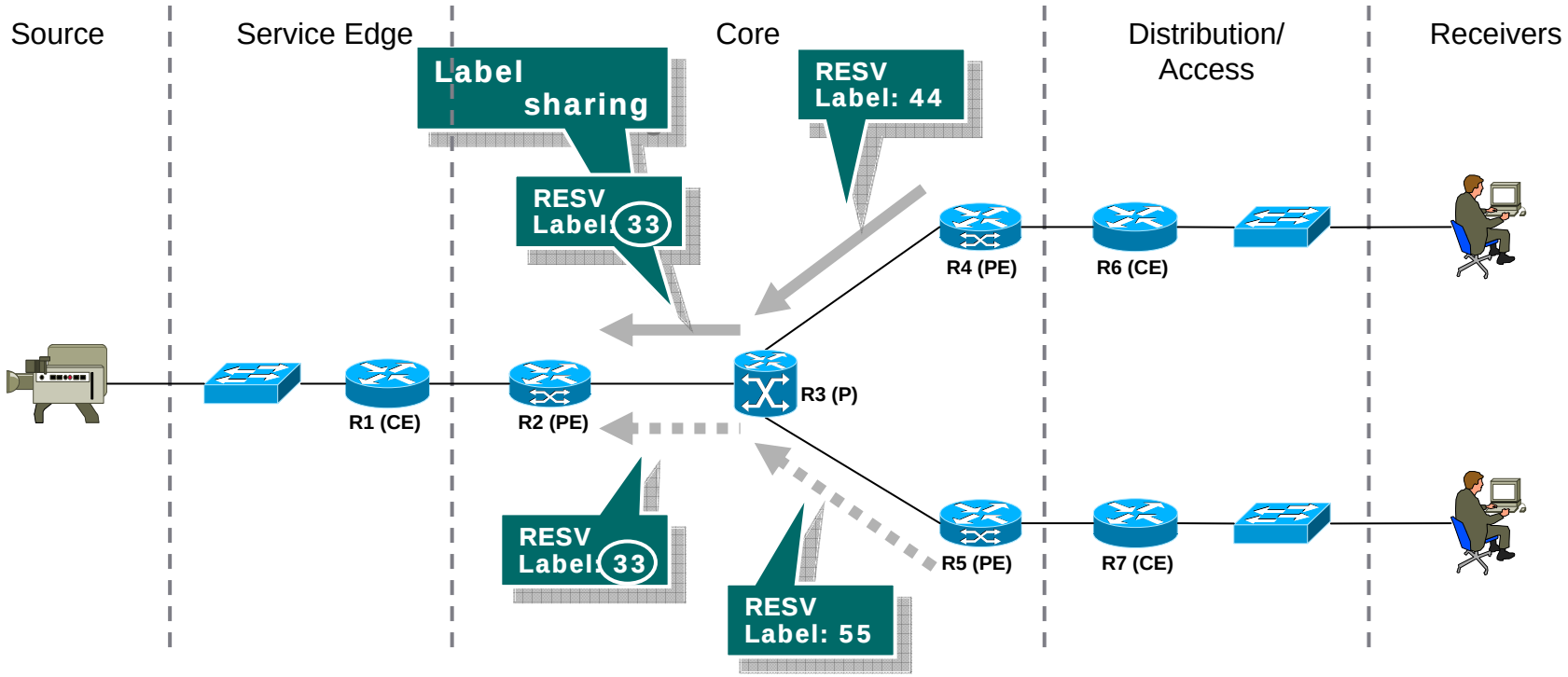
- RSVP-TE
  - Extension to RSVP-TE to build P2MP trees
  - Source Driven (unlike PIM)
  - Supports Traffic Engineering
- Multicast LDP (mLDP)
  - Extension to LDP to build P2MP and MP2MP Trees
  - Very similar to PIM
    - Receiver Driven
- PIM (Not focus of this presentation)

# P2MP Tunnel Setup – RSVP-TE Non-Aggregated Mode: PATH Message



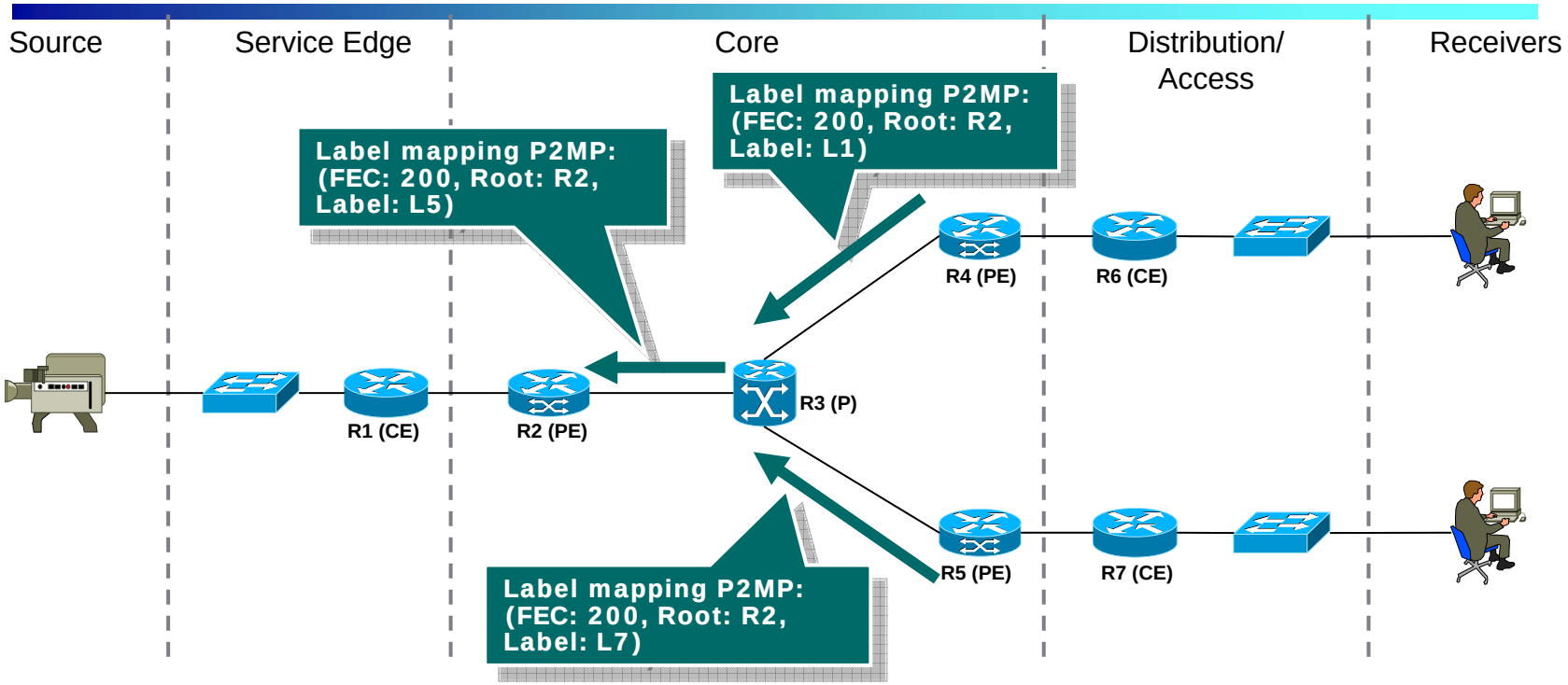

- Non-Aggregated Mode: Headend sends one PATH message per destination
- RSVP-TE also supports aggregated mode, where a single PATH message can carry all sub-LSP information for all destinations

# P2MP Tunnel Setup – RSVP-TE Non-Aggregated Mode: RESV Message




- RESV messages are sent by tailend routers – communicates labels and reserves BW on each link
- Label Advertisement carried in the RESV message

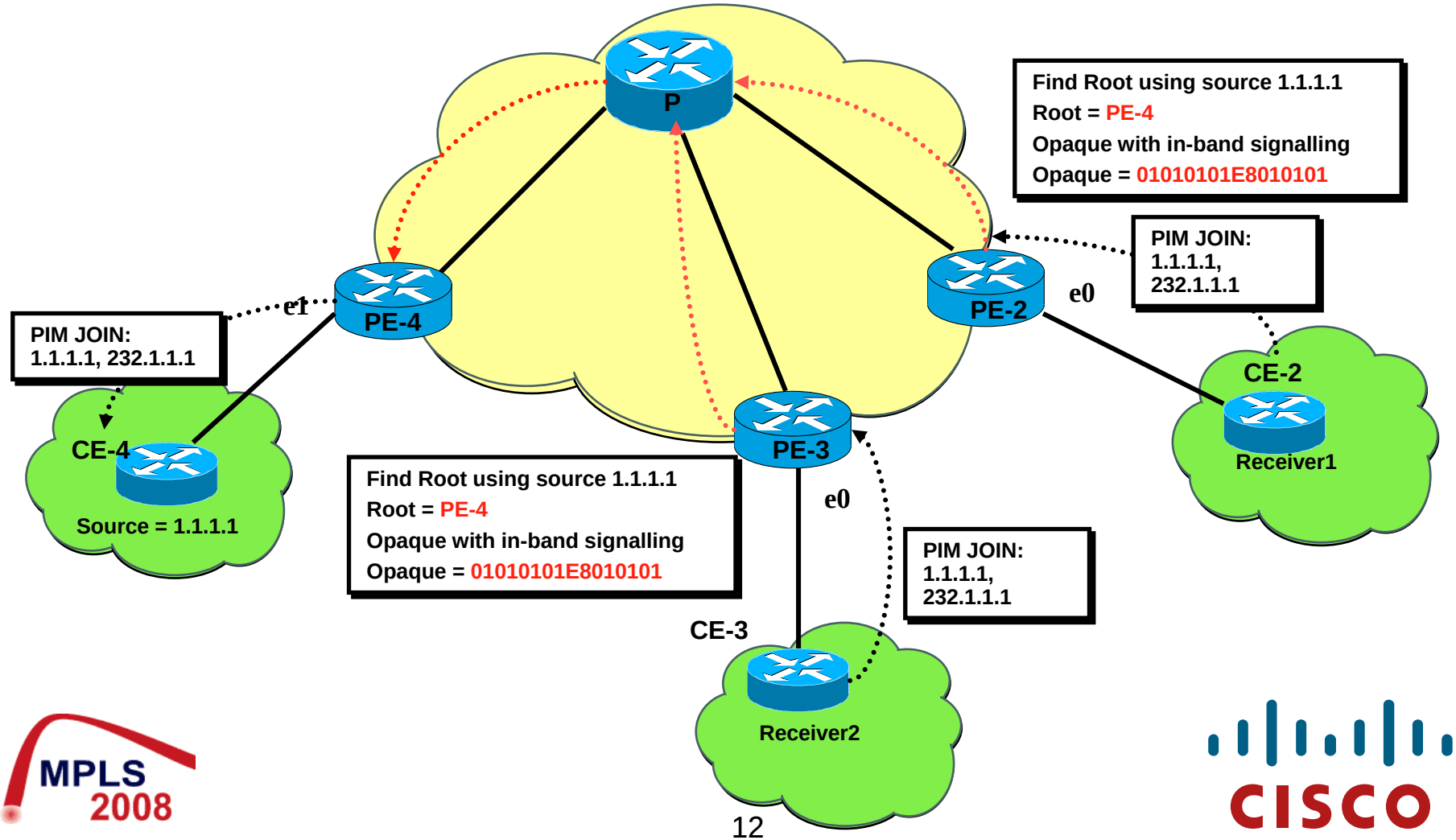
# P2MP Tunnel Setup – mLDP



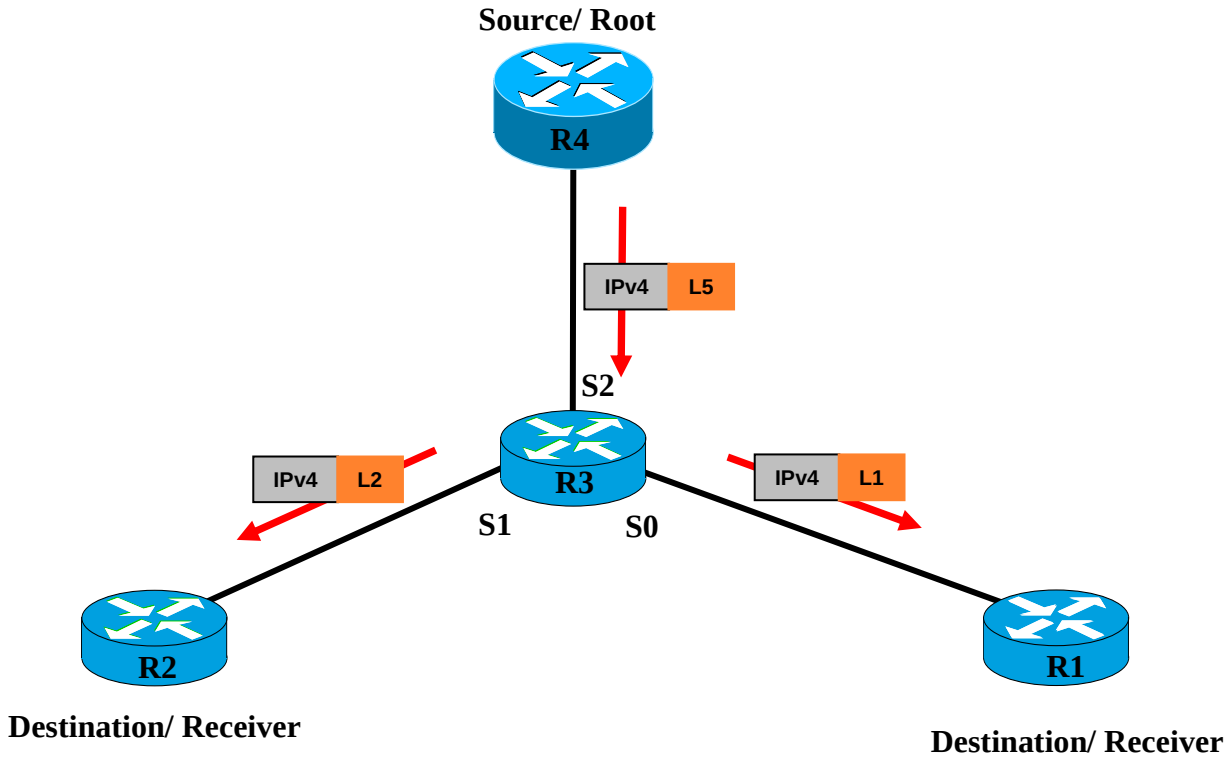

- Each leaf node initiates P2MP LSP setup by sending mLDP Label Mapping message towards the root, using unicast routing
- Label Mapping message carries the identity of the LSP, encoded as P2MP FEC
- Each intermediate node along the path from a leaf to the root propagates mLDP Label Mapping towards the root, using unicast routing

# Transiting PIM SSM (in-band) using mLDP

## mLDP supports in-band transiting of PIM SSM

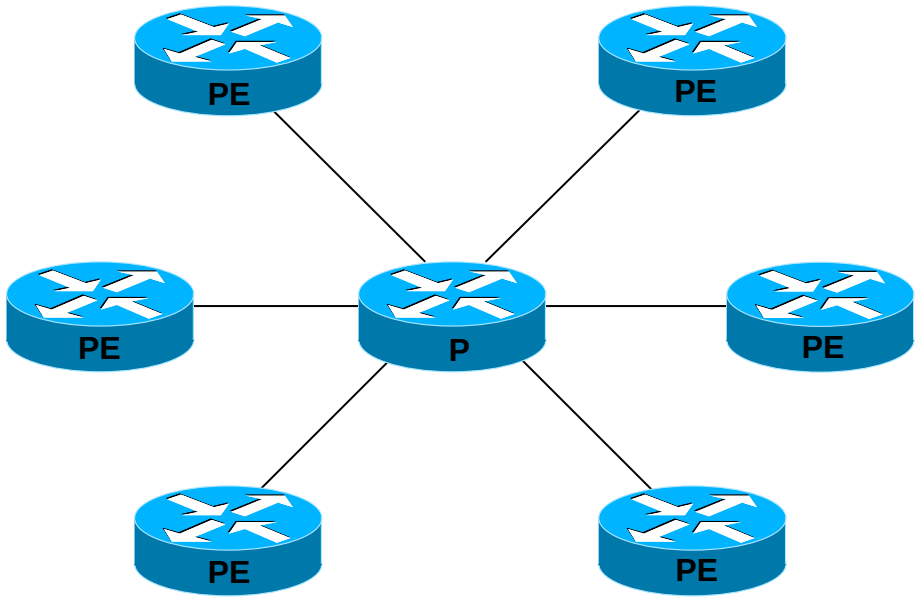


# P2MP LSP (Data Plane)



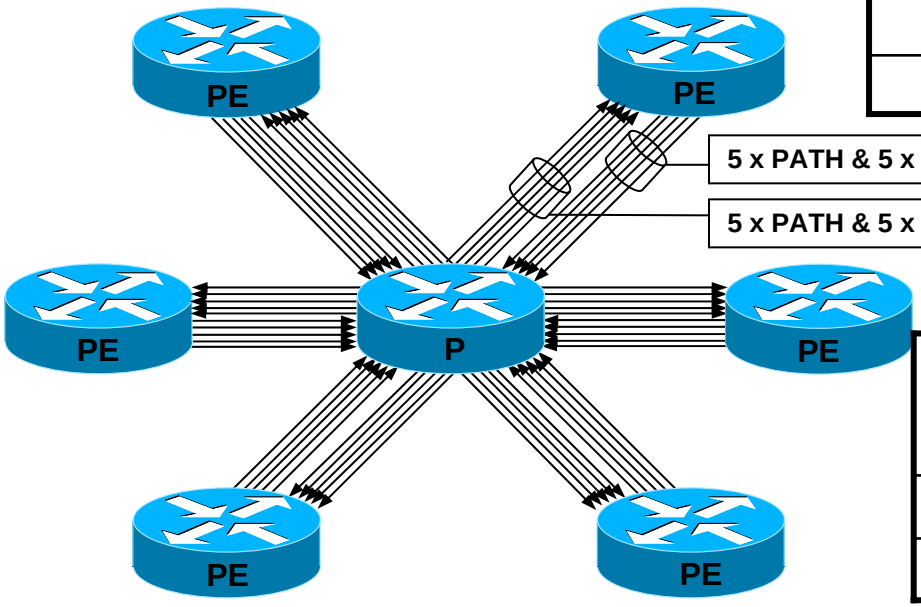
# Comparison Basis for P-Tree Type and Protocol

- Suppose we are building an (MI/MS)-PMSI 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 (MI/MS)-PMSI.
- 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.




**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 |

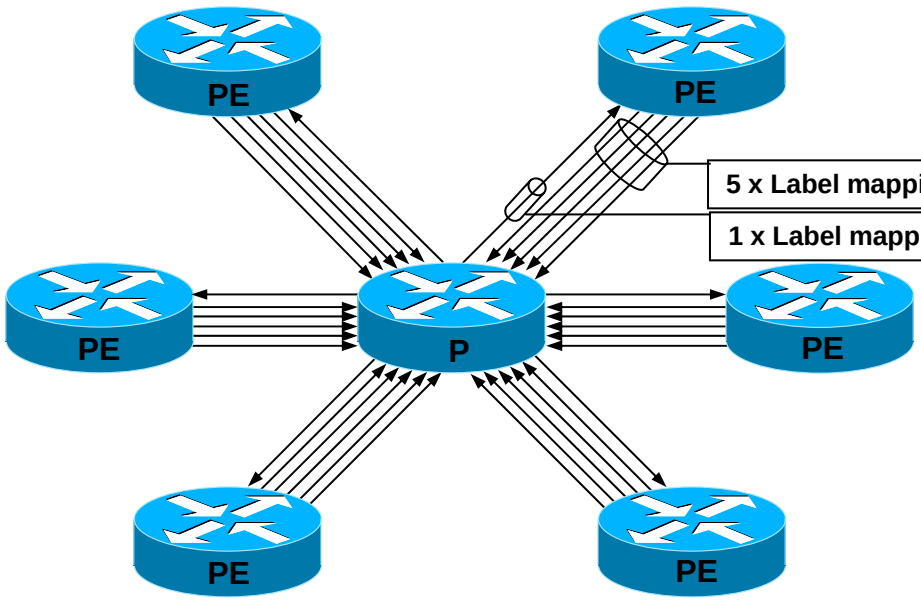
**1K PE Routers**

| 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




**6 PE Routers**

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

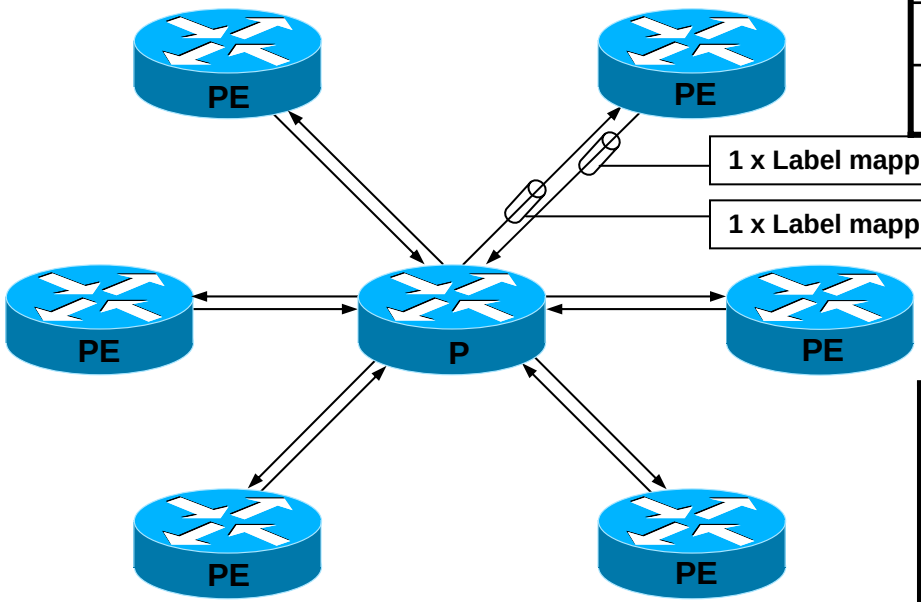
**1K PE Routers**

| 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




**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 |

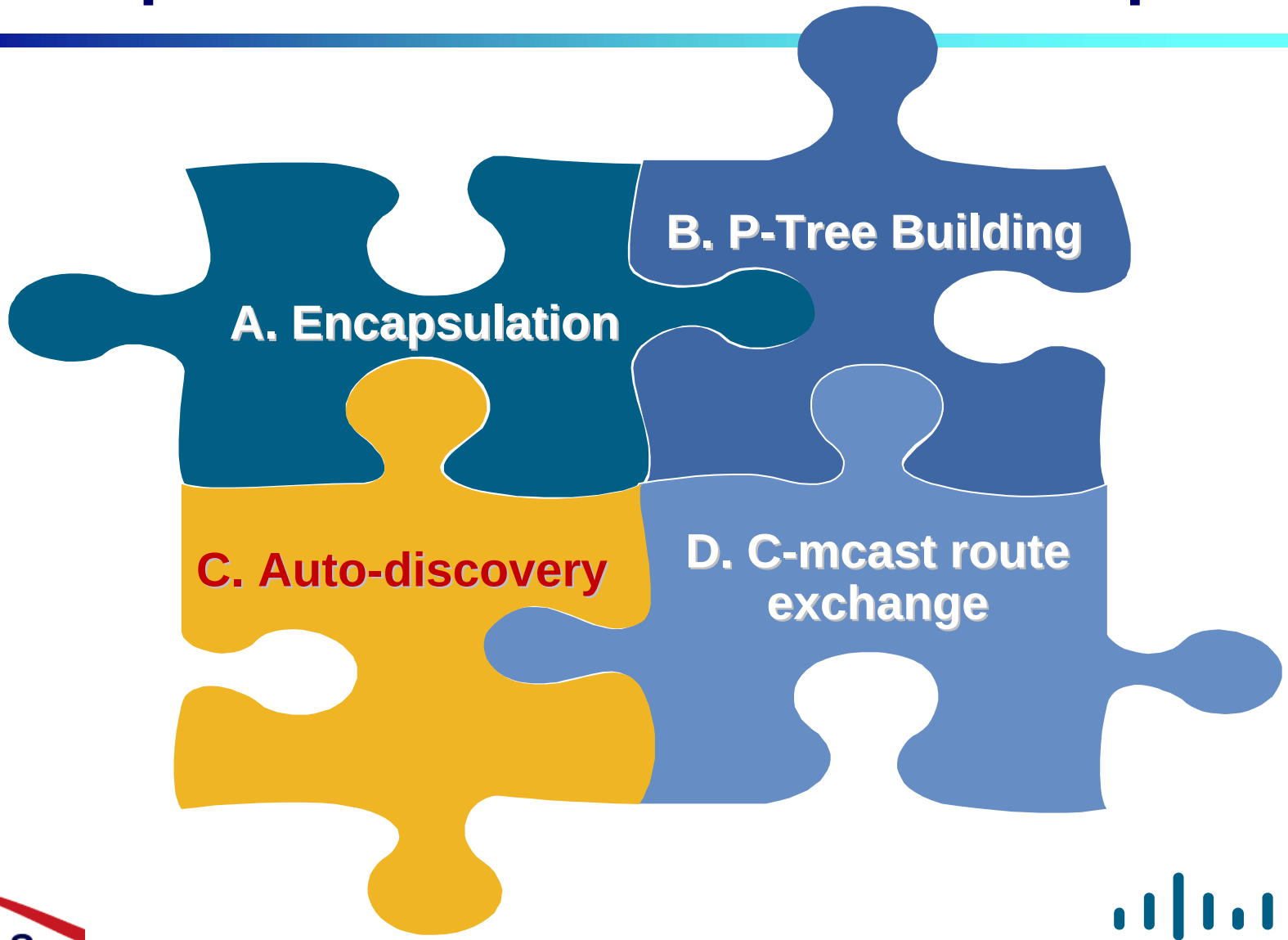
- **O(1) Control Plane States**
- **O(1) Data Plane States**
- **O(PE) Protocol Messaging**

# 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 for MI-PMSI and MS-PMSI.
- 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.
- mLDP supports in-band transiting of PIM SSM.

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

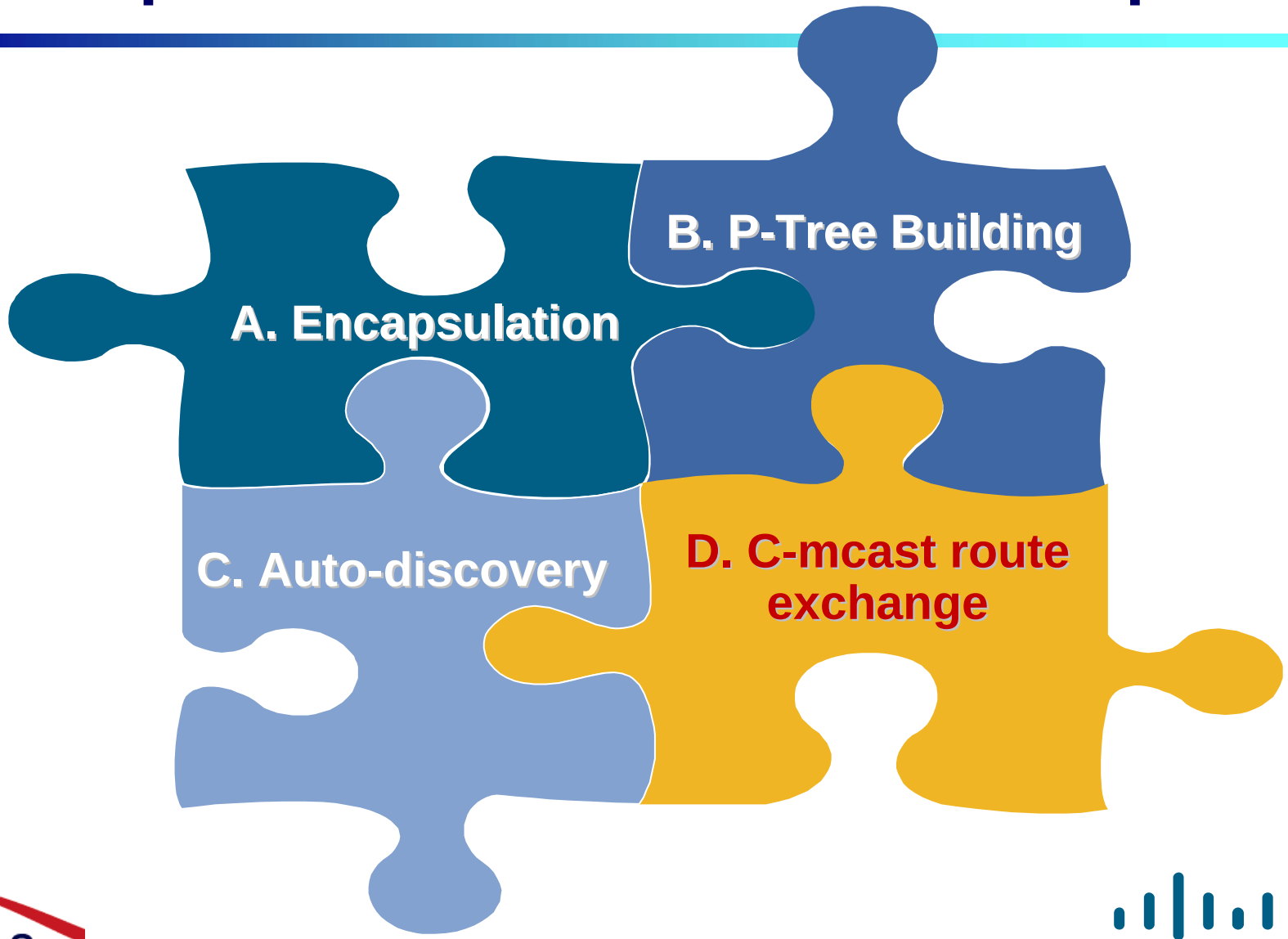
# Components of Multicast Solutions Space



# 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



# Multicast Signaling (Exchanging Customer mcast routes)

- Mechanics used for customer mcast route exchange is independent of core tree building and auto discovery methods
- In draft-ietf-l3vpn-2547bis-mcast-06 two options are specified:
  - **Option 1: Per-mVPN PIM peering among the PEs**
    - This is deployed today (draft-ietf-l3vpn-2547bis-mcast-06, a.k.a draft-rosen)
  - **Option 2: BGP**
    - Analogous to RFC4364 exchange of VPN-IPv4 routes, but with new MVPN AFI/SAFI

# 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.

# 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
  - Advertisement contains essentially the same info as a PIM join or prune (source, group, PE sending the message)
- 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?**

- MI-PMSI (E-LAN) (all PEs to every PEs)
- S-PMSI (one PE to a select subset of PEs)
- MS-PMSI (Partitioned E-LAN)

# 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.

# S-PMSI

- From one PE to a select subset of PEs.
- Also known as data-MDT.
- An S-PSMI network is a per PE tree from one PE to a select subset of PEs
- Uses a single P2MP LSP per ingress PE.

# MS-PMSI

- Combination between S-PMSI and MI-PMSI.
- 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 MI-PMSI

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

- 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.
- If the E-LAN is a MP2MP, BGP needs context labels to solve the duplicate forwarder problem
- BGP-SM has some differences from PIM-SM, impact remains to be seen

# BGP vs. PIM for C-mcast Route Exchange Over MS-PMSI

- Multicast signalling over MS-PMSI 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 S-PMSI

- Multicast signalling over Selective-PMSI needs to support:
  - Bidirectional multicast is not supported
  - No duplicate forwarder detection necessary.

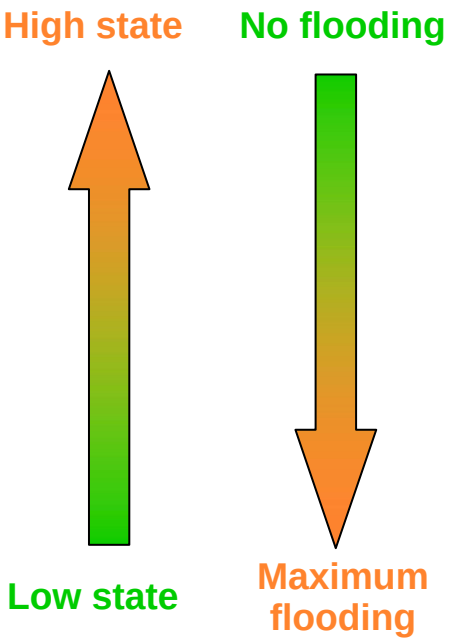
- As this is a uni-directional tree, PIM cannot run without some modifications.
- The required modifications that are being discussed in IETF
- PIM over Reliable Transport (PORT) is a good alternative
  - draft submitted to IETF

- 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.

# Aggregation policy

How much flooding you need to accept is dependent upon the amount of state the core tree protocol is able to handle.

- Scaling requires us to minimize P-router state, i.e., use as fewer trees as possible.
- Optimality demands us sending traffic only to PEs that needs it, i.e., use a tree for each customer multicast group.




1. A PE tree tree per Multicast stream
   - Most optimal customer multicast packet forwarding.
2. A PE tree per VPN
   - All VPN traffic source by this PE is aggregated on this tree.
3. An E-LAN per VPN
   - All traffic sourced by any PE in this VPN is aggregated on E-LAN.
4. Single PE Tree
   - All customer VPN traffic sourced by this PE is aggregated over a single PE tree.
5. Single E-LAN
   - All customer VPN traffic, sourced by any PE is aggregated over a single E-LAN in the core.

# Aggregation Solutions

- Solution: lots of options
  - draft-ietf-l3vpn-2547bis-mcast-06 (a.k.a draft-rosen) has one MDT per VPN, and optional data MDT for high BW or sparse customer groups
  - IETF WG draft retains those options; also allows a tunnel to be shared among multiple mVPNs
    - Better aggregation, less optimality
    - Requires a de-multiplexing field (e.g., MPLS label)
  - Utility of aggregation depends on amount of “congruence” and traffic volume

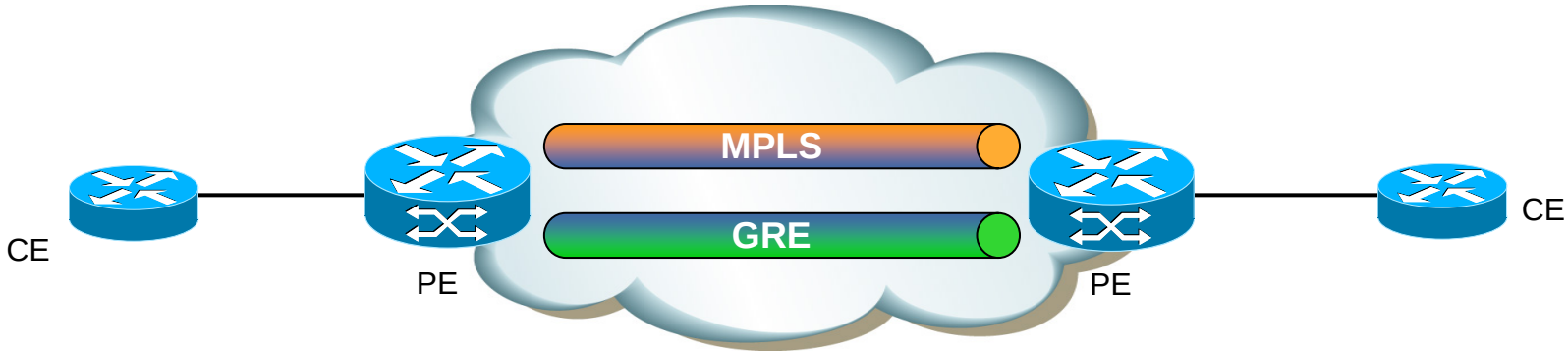
# Agenda

- 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




- Migrating from a GRE to an MPLS profile (either MI-PMSI or MS-PMSI) is easy because the PIM signaling and MVPN model (draft Rosen) does not change
- To facilitate migration, MPLS and GRE tunnels can co-exist side-by-side
- PE's will see same PIM neighbor over different tunnels
- PE's may select their preferred Tunnel
- Migration is just an RPF change for PIM
  - No interruption to customer multicast traffic
  - Phased migration possible

# 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.
  - MS-PMSI partitions MI-PSMI, which benefits scalability of PIM.

**• Use of PIM for c-mcast route exchange during migration to label switched multicast core 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

# Conclusion and Summary

- MPLS has a rich set of options for supporting multipoint services
- Richness derives from broad set of service demands
  - No one-size-fits-all answer
- MVPN solution space has various options:
  - Build P-trees with PIM, RSVP-TE or MLDP
  - Auto discover MVPN members with PIM or BGP
  - Exchange C-mroutes with PIM or BGP
- Many factors, such as rate of joins, customer topology drive tradeoffs, need to be consider in selecting a specific solution


CISCO