IP Multicast and Multipoint Design for IPTV Services

Mike McBride

Session Goal

To provide you with a thorough understanding of the end-to-end protocol, mechanics and service element of IP multicast technologies used in IPTV networks.

Agenda

- Introduction
- Architectural overview
- IP multicast primer (SSM)
- Transit Transport Design options

  Native (PIM), mLDP, RSVP-TE P2MP, L2/L3VPN, signaling
- Resiliency

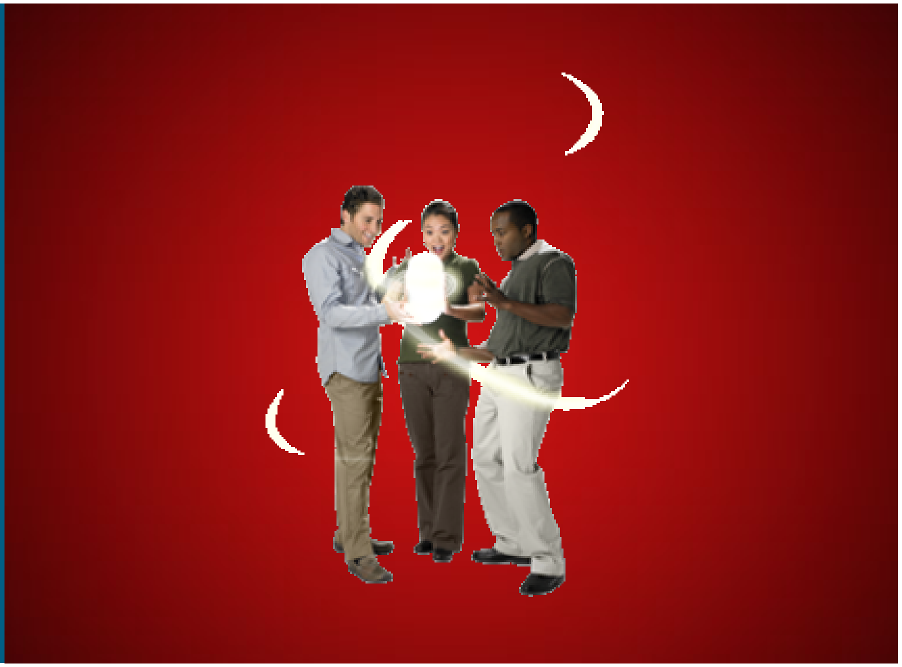
  Source redundancy, protected pseudowires, FRR, live-live, MoFRR
- Broadband Edge

  IGMP snooping, MVR, vVLAN, DSL, Cable, FTTH
- Path selection

  ECMP, multi topologies, RSVP-TE P2MP
- Admission control
- Channel changing

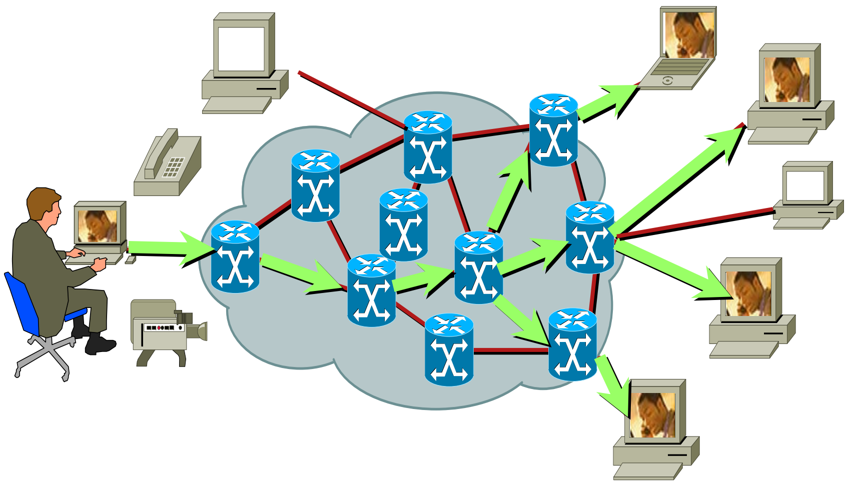
  Join/leave latency, static/dynamic forwarding, acceleration

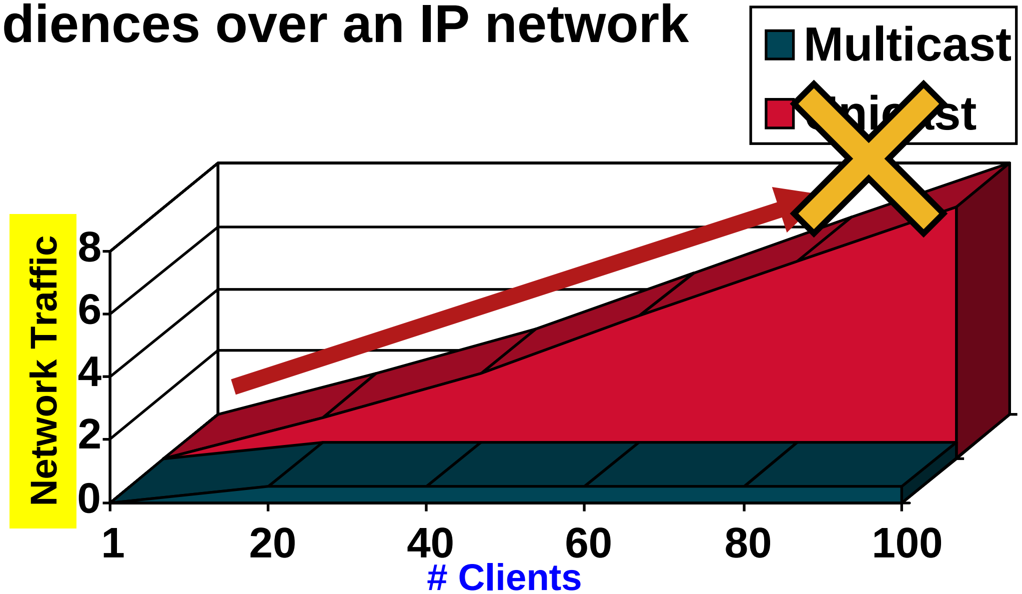
Introduction
IPTV and
IP multicast



Multicast for IPTV Delivery

Distribute information to large audiences over an IP network





Multicast

1. Efficiently Controls network traffic
2. Reduces server and CPU loads
3. Eliminates traffic redundancy
4. Makes Multipoint applications possible

Multicast Benefits

- **Increase Productivity & Save Cost**
- **Generate New Revenue Stream**

IP Multicast Is a Green Technology!!!!!!!!

Internet Protocol (IP) multicast is a bandwidth-conserving technology that reduces traffic by simultaneously delivering a single stream of information to thousands of corporate recipients and homes; applications that take advantage of multicast include videoconferencing, corporate communications, distance learning, and distribution of software, stock quotes, and news

Facts

- **Multicast reduces the number of servers required—Unicast uses many servers which must process individual requests for streaming media content from tens, hundreds or thousands of users and then send duplicated streams**
- **Reducing the number of network resources required not only saves capital costs and operating expenditure but also saves power which in turn reduces carbon footprint**
- **It requires 838 pounds of coal to power one PC for one year**

Broadcast IPTV = IP multicast

- …however transport network transits packets ..

  "Native IP multicast", MPLS, L2, optical
- IP multicast sources:

  Encoder, Transcoder, Groomer, Ad-Splicer, …
- IP multicast receivers:

  Transcoder, Groomer, Ad-Splicer, QAM, STB
- IP == IPv6 (Japan) or IPv4 (RotW *rest of the world*)

  No address exhaustion issue (SSM)

  No/slow move to IPv6 for IPTV in RotW

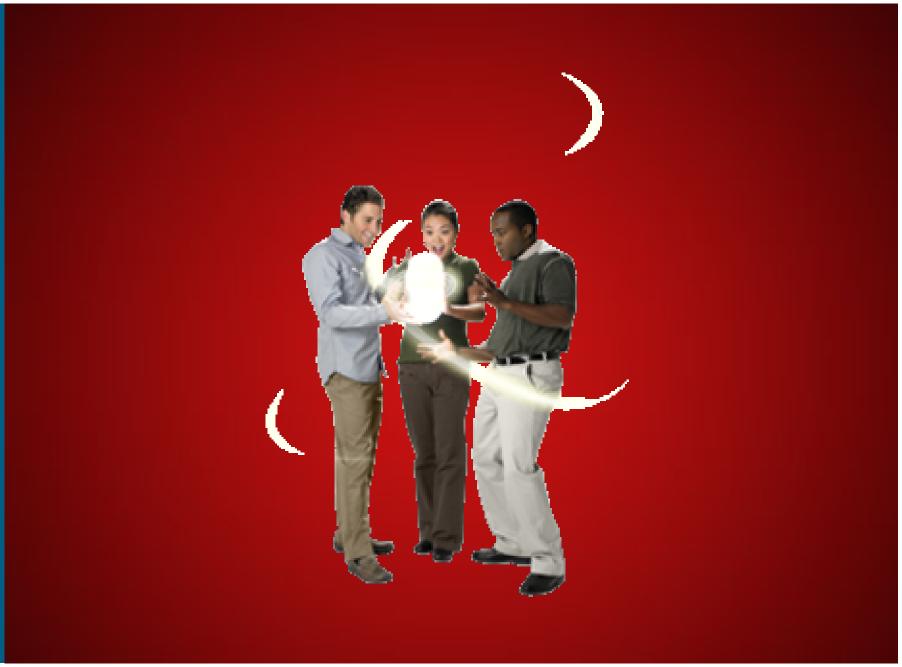
Deployment strategy
Overview, Recommendation

- **Network**
  - Add IP multicast to your network core
  - Choose transport methods based on SLA and operational requirements/preferences
    - Native IP multicast, MPLS, L2, mix
  - Solution should minimize involvement in provisioning of individual applications/services
- **IPTV services**
  - Start with traditional broadcast TV
  - Investigate extending IPTV and other (IP multicast) services
    - More RoI on network layer investment

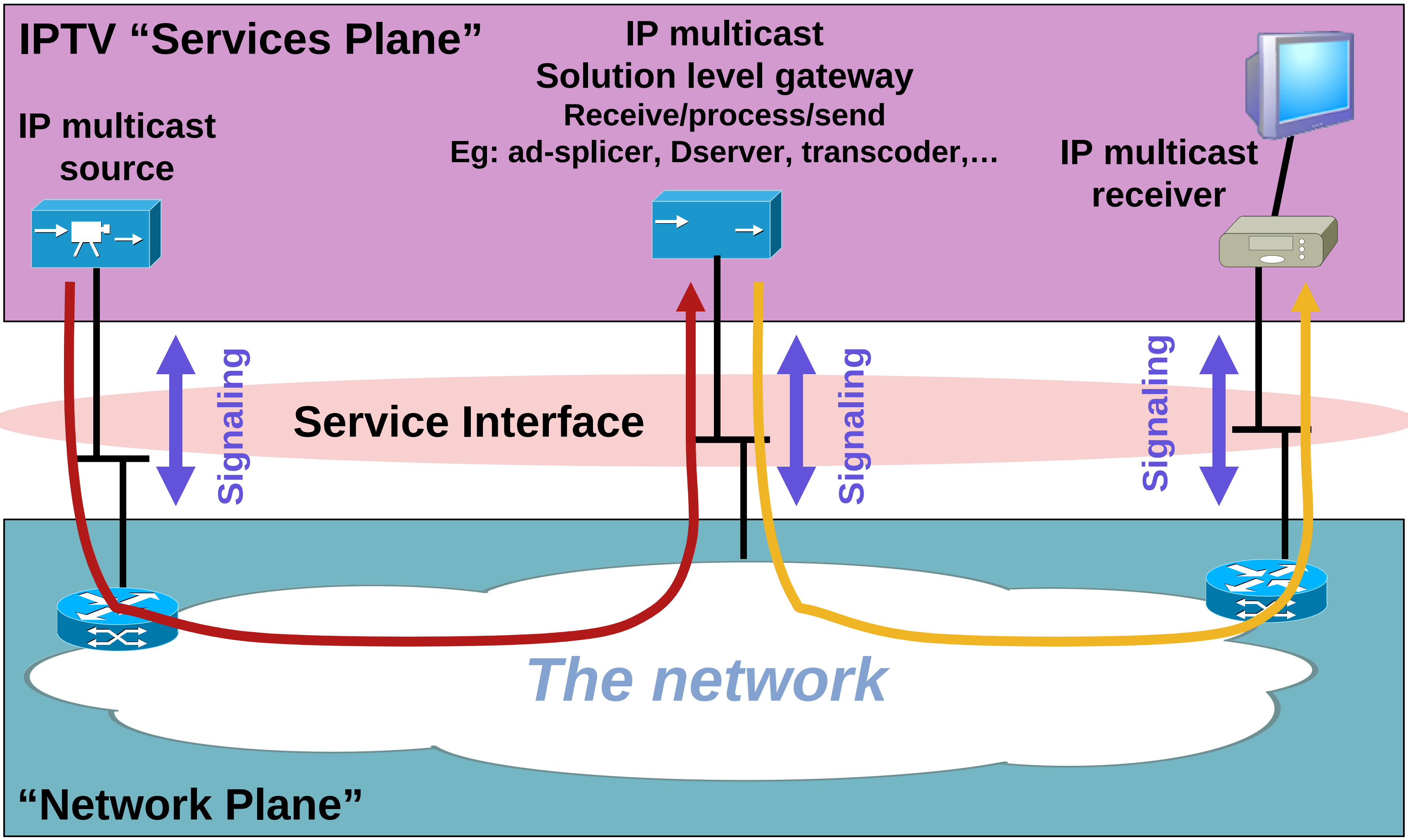
Additional service opportunities
Across common SSM IP multicast service

- **No need to change the IP multicast functionality in the network**
    - May want improvements on optional elements (RSVP, …)
- **Extending IPTV broadcast service**
    - Dynamic redundancy (regional to national)
    - Variety of reach of transmission (src->rcvr)
        - Groomer/transcoders, Add-Splicers
    - Switched Digital Video, oversubscription
    - Wholesale, dynamic, international channels
- **Other services**
    - Commercial (MVPN)
    - Content pre-provisioning to VoD server, STB
    - Multicast in Internet Service (eg: To PC)
    - Voice conferencing, gaming, surveillance, …

Architectural Overview



50,000 feet architecture
IPTV and multicast



50,000 feet architecture
Goals

- **Separate "network" and "services" plane**

    Network = shared infrastructure for all services

        Routers, switches, optical gear, NMS, …

    IPTV = encoders, groomers, splicers, VoD server, STB, …

        Often operated by different entity/group than network

- **IP multicast**

    Allow to attach solution plane devices (sourcing, receiving) anywhere – global, national, regional, local. Start/stop sending traffic dynamically, best utilize bandwidth only when needed.

    One network technology usable for all services (IPTV, MVPN, …)

    Enable network operator not to provision/worry about individual programming.

- **Service Interface**

    How network & service operator infrastructure interacts with each other

    SLA of IP multicast traffic sent/received

    Signaling used

Service Interface

Basic service description (recommended IP multicast for IPTV)

- **P2MP = SSM tree (traffic forwarding)**
  - Build trees from any individual source.
    - Easy to: Inject everywhere, receive everywhere (securely)
  - Best join/prune latencies
    - *Warning: fast network join is not same as fast solution join*!
  - Largest #trees supported,…
  - No coordination of tree addresses (SSM channels)
  - No spoofing of traffic across the tree
- **Redundancy**
  - Source redundancy: Anycast/Prioritycast
  - Optional live-live service (path separation)
    - (for up to 0 packet loss during network failure)

Service Interface
More features

- **Admission control**

  Per-flow bandwidth based admission control

    RSVP-TE, RSVP/UPnP-CAC at edge

    Router local admission control

- **More …**:

  (per subscriber) access control (eg: lineup), provisioning of subscriber policies, …

  Accounting (Radius, Netflow, …)

  Management, troubleshooting

  Not further covered in this presentation
  Lots of product specifics

Service Interface
Expectation against service devices

- **Mandatory:**

  SSM-tree building: IGMPv3/MLDv2 with SSM ‘joins’

    receivers needs to know (S,G) channels to join

  Send multicast packets with TTL > 1

- **Optional:**

  DSCP setting

  Signaling for source redundancy

  Send/receive traffic twice (redundancy and/or live-live)

  RSVP/UPnP-CAC – for admission control

- **Workarounds in network**

  Static building of multicast trees, SSM transition, DSCP marking, router based CAC, …

Network infrastructure

Only implicitly impacting services (resilience, security,..)

- **Preferred choice of transport:**

  IP (native multicast/PIM) or MPLS (mLDP and RSVP-TE P2MP)

- **Path selection**

  (dual path) – MoFRR or exposed to service

  Tree cost optimization

  Load-splitting:

    ECMP: PIM and mLDP

    Arbitrary: RSVP-TE (CSPF)

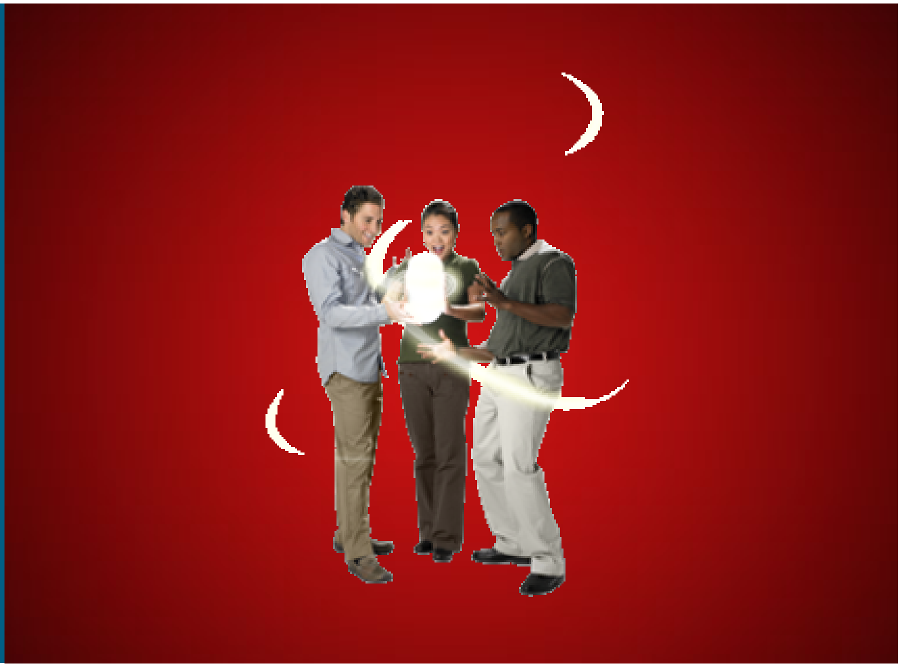
- **Preferred choice of virtualization**

  L2VPN, L3VPN context – or why not…

- …not complete list

IP multicast primer (SSM)

… as required for IPTV…



Protocols and Services …and IP multicast

- multicast / multipoint *protocols*

  Between routers, switches, ..

  “Only of interest to network operator”

  PIM-SM, MSDP, (M)BGP, AutoRP, BSR, mLDP, RSVP-TE, …), IGPs (OSPF, ISIS), …

- multicast *services*

  How end-devices can use IP multicast

  “Of interest to network and service operator”

  ASM, SSM (and protocols “IGMP/MLD”)

  Service operator just need to add SLA requirements!

IP multicast services

- ASM: "Any Source Multicast" (1990, rfc1112)
  - The "traditional IP multicast service" (collaborative)
  - Sources send packets to multicast groups
  - Receivers join to (G) groups, receive from any source
- SSM "source specific multicast" (~2000, rfc4607/4604)
  - The multicast variant for IPTV (or other "content distribution")
  - Unchanged: Sources send packets to multicast groups
  - Receivers subscribe (S,G) channels,
receive only traffic from S sent to G
  - *Primarily introduced (by IETF) for IPTV type services*
    - *Because of limitations of standard (protocol) model for ASM*

Standard protocol model for ASM

- What is the standard protocol model ?
- A1: MBone and DVMRP

  Please go back to your time machine and dial 1994

- A2: Native Multicast with:

  PIM-SM

    AutoRP, BSR or MSDP/Anycast-RP redundancy

    MSDP for Interdomain support

    Multiprotocol BGP for interdomain RPF selection

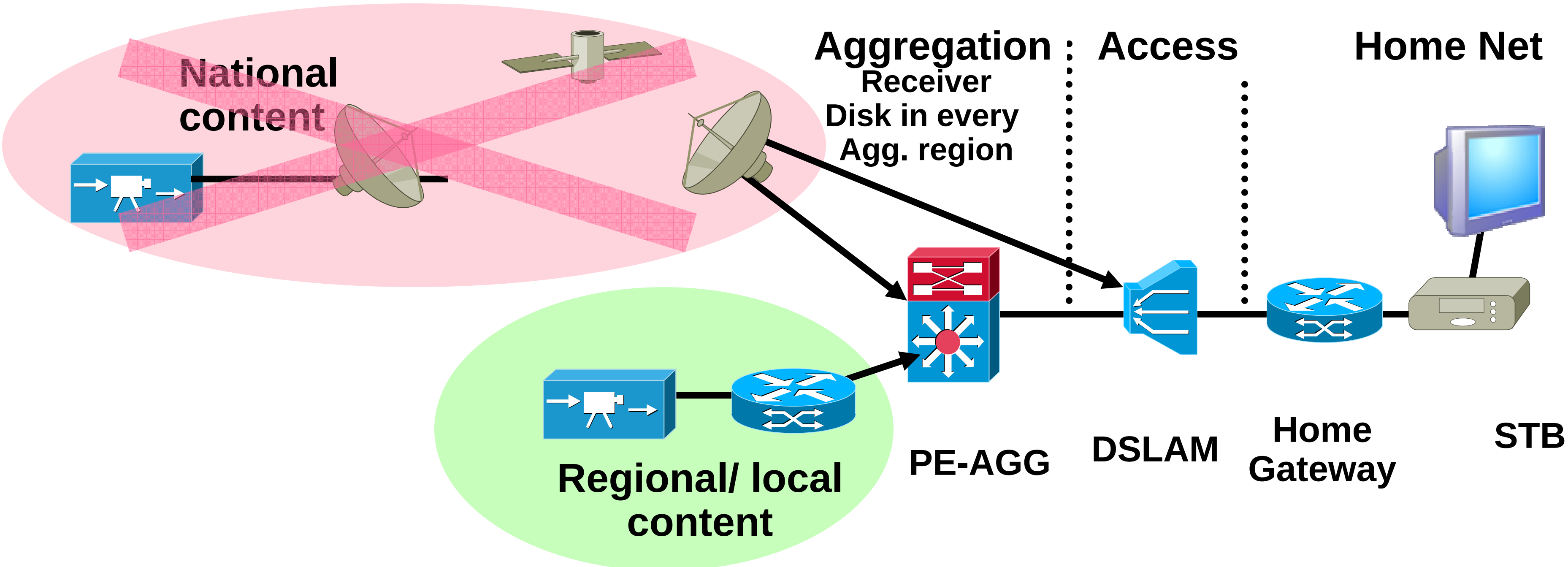
  Best available general purpose ASM protocol suite

    …but with issues

IP multicast services
Issues with ASM – resolved with SSM

- ASM
  - DoS attacks by unwanted sources
  - Address allocation
- Standard protocol suite
  - Complexity of protocol operations required
    - PIM-SM (RPT+SPT+Switchover), RP redundancy, announce, location
    - MSDP (RPF), BGP congruency,
    - Interactions with MPLS cores, bandwidth reservation, protection
  - Scalability, Speed of protocol operations (convergence)
    - RPT + SPT operations needed

End-to-end protocol view
Historic development



- Old designs: Use non-IP satellite distribution, inject regional / locally

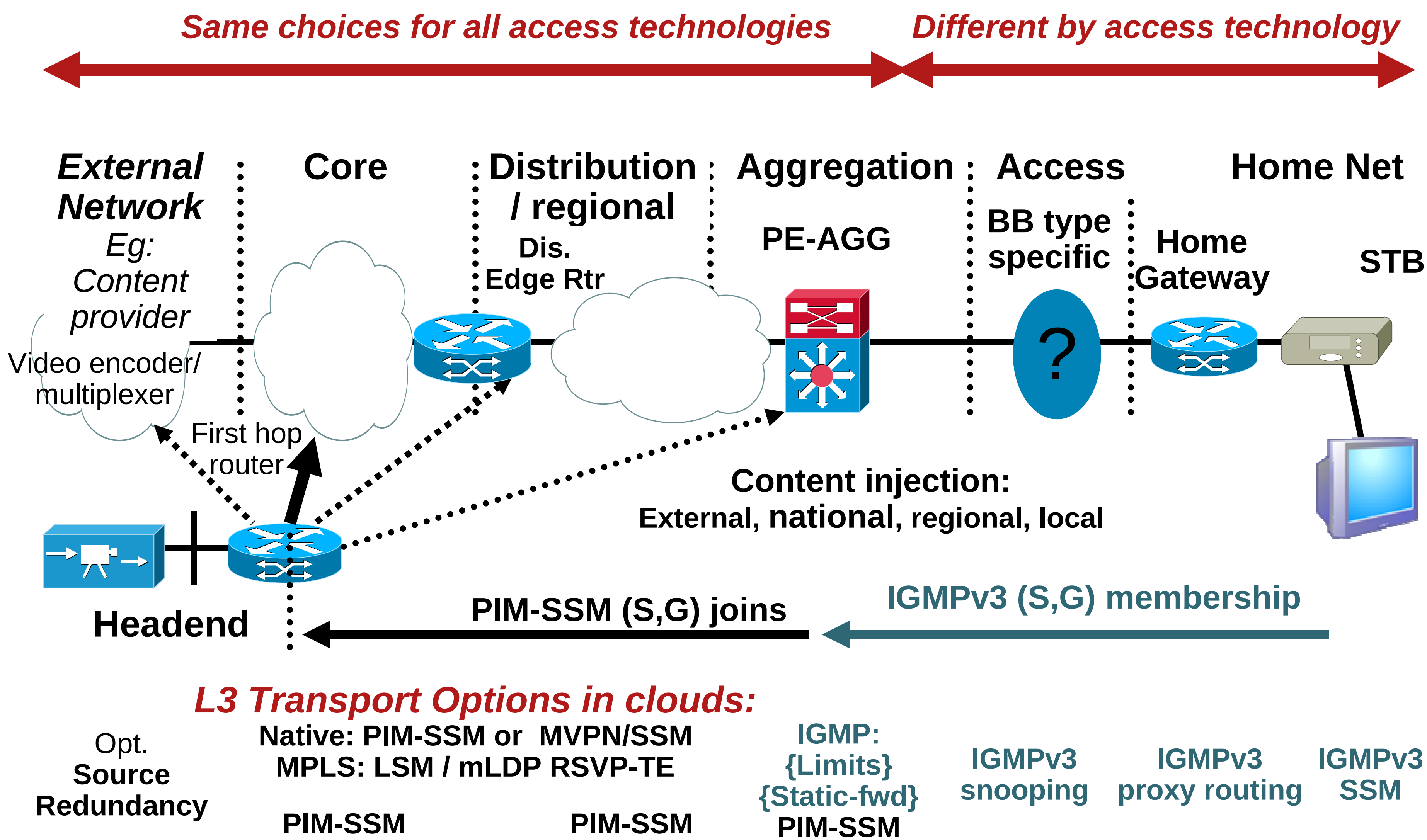
  “National IP network can not transport video (cost, function)”

- Current designs: use regional/local injection only for regional/local content

  The national core IP network can transport video perfectly

  May also want to feed local/region back across core (national redist)

End-to-end protocol view example: L3 aggregation



Transit Transport design options

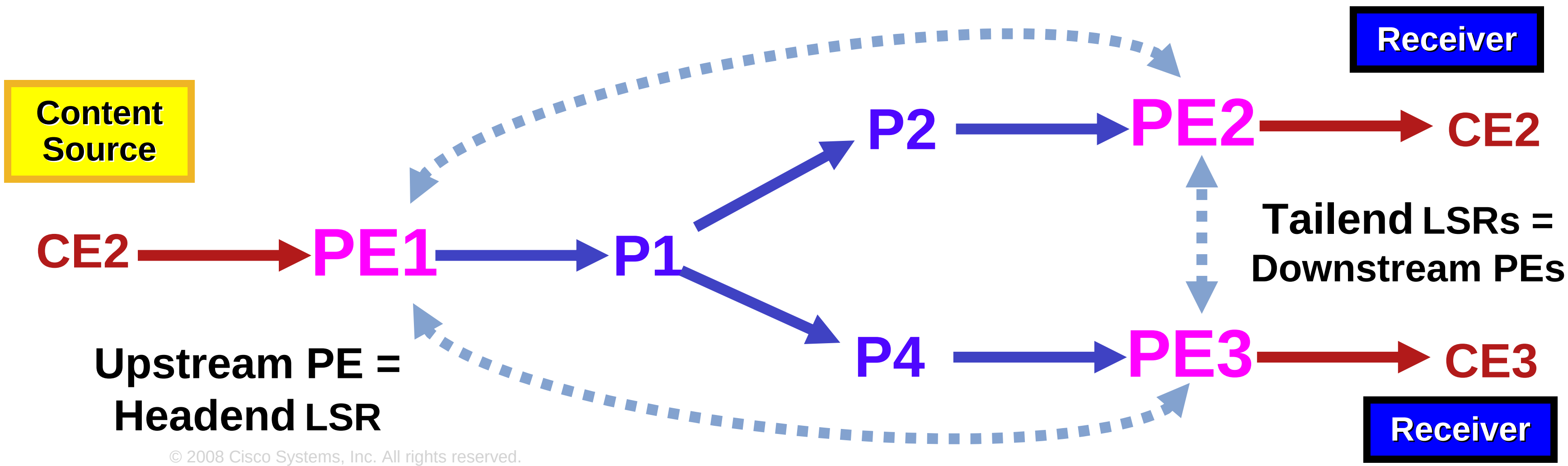
Transport architecture
Overview

- **Common deployments: Native PIM-SSM or MVPN**
- **Concentrate on futures / components**
  - Support for MPLS multicast (LSM)
    - Build P2MP / MP2MP label switched delivery trees
    - mLDP (P2MP, MP2MP), RSVP-TE P2MP
  - Put traffic into a VPN context
    - As a method of service isolation / multiplexing
  - Using L2 vs. L3 on PE nodes
    - To “integrate” better into an L2 service model
  - Redefine PE-PE signaling for MVPN

Overview

Elements of transport architecture for tree building

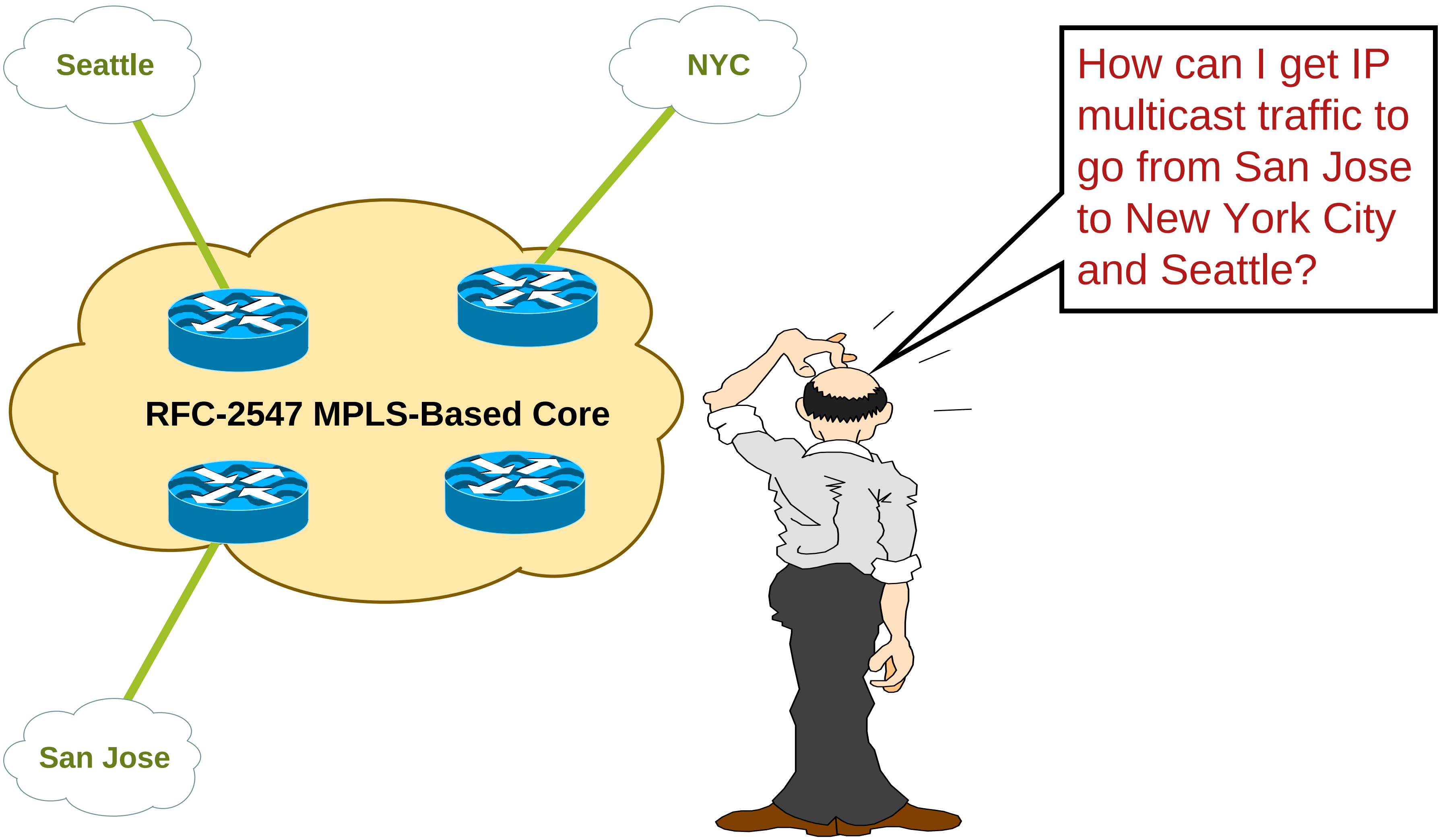
- C(ustomer)-tree building protocols
  IPTV: IGMPv3 / PIM-SSM
- P(rovider)-tree (PMSI) building protocols
  Native: PIM-SSM/SSM/Bidir, MPLS: mLDP, RSVP-TE
- PE mapping: C-tree(s) to P-tree
  1:1/N:1 (aggregation) ; ‘native’/VPN (L2, L3) ; static/dynamic
- PE-PE (“overlay”) tree signaling protocols
  Optional PIM or BGP (extensions)
  Not needed: native IPv4/IPv6, ‘direct-MDT’ mLDP, static mapping



Combinations with L3 on PE
Current widely deployed

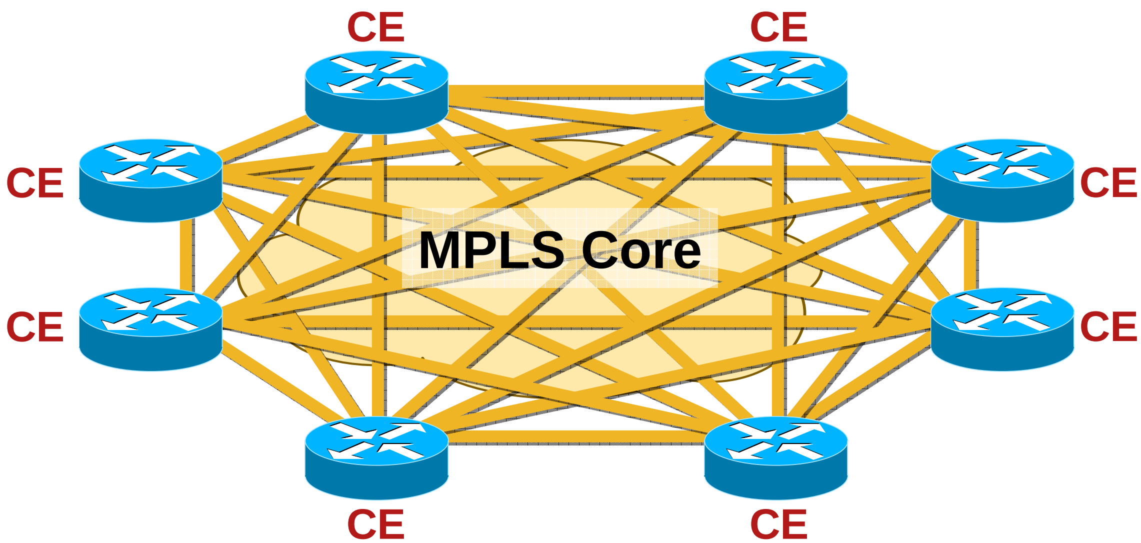
- **“Native IP multicast” (IPv4/IPv6)**
  - IPv4/IPv6 PIM-SSM in core
  - User side = core tree: No PE-PE signaling required.
  - “RPF-Vector” for “BGP free core”
- **“MVPN”(PIM)**
  - Carries traffic across RFC2547 compatible L3 VPN.
  - With aggregation
  - IPv4 PIM-SSM/SM/Bidir in core (IPv4)
  - RFC2547 BGP ; GRE encap/encap on PE
  - PE-PE signaling required
    - I-PMSI = Default-MDT ; SI-PIMSI = Data-MDT
    - BGP extensions for InterAS and SSM support

Deploying MPLS-Based L3 VPNs and…

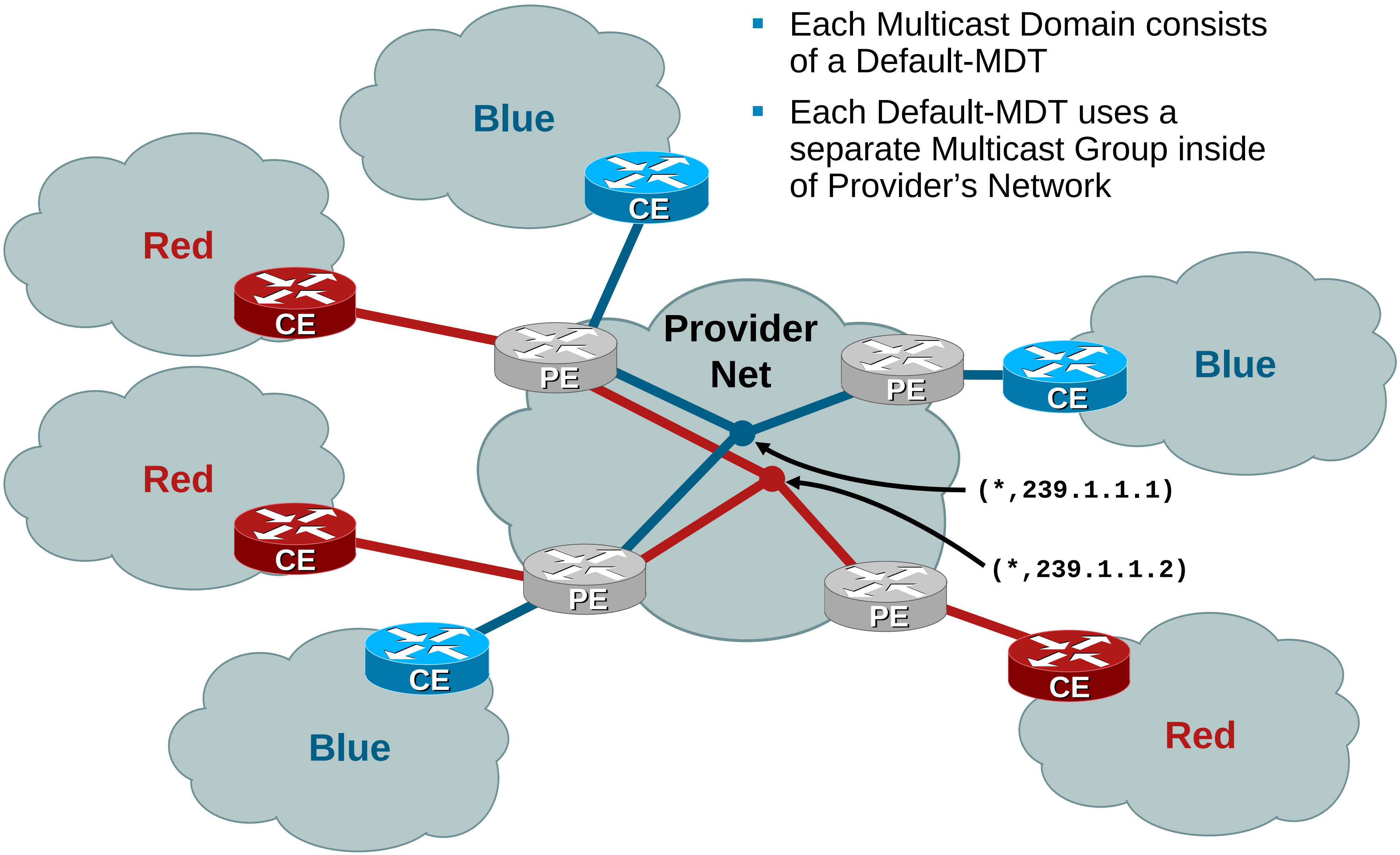


Multicast VPN: Challenges

- Multicast not originally supported with MPLS (RFC 2547)
- Workaround was point-to-point GRE tunnels from CE to CE
- Not scalable with many CE routers
  - Traffic overhead
  - Administration overhead



Multicast VPN: Overview



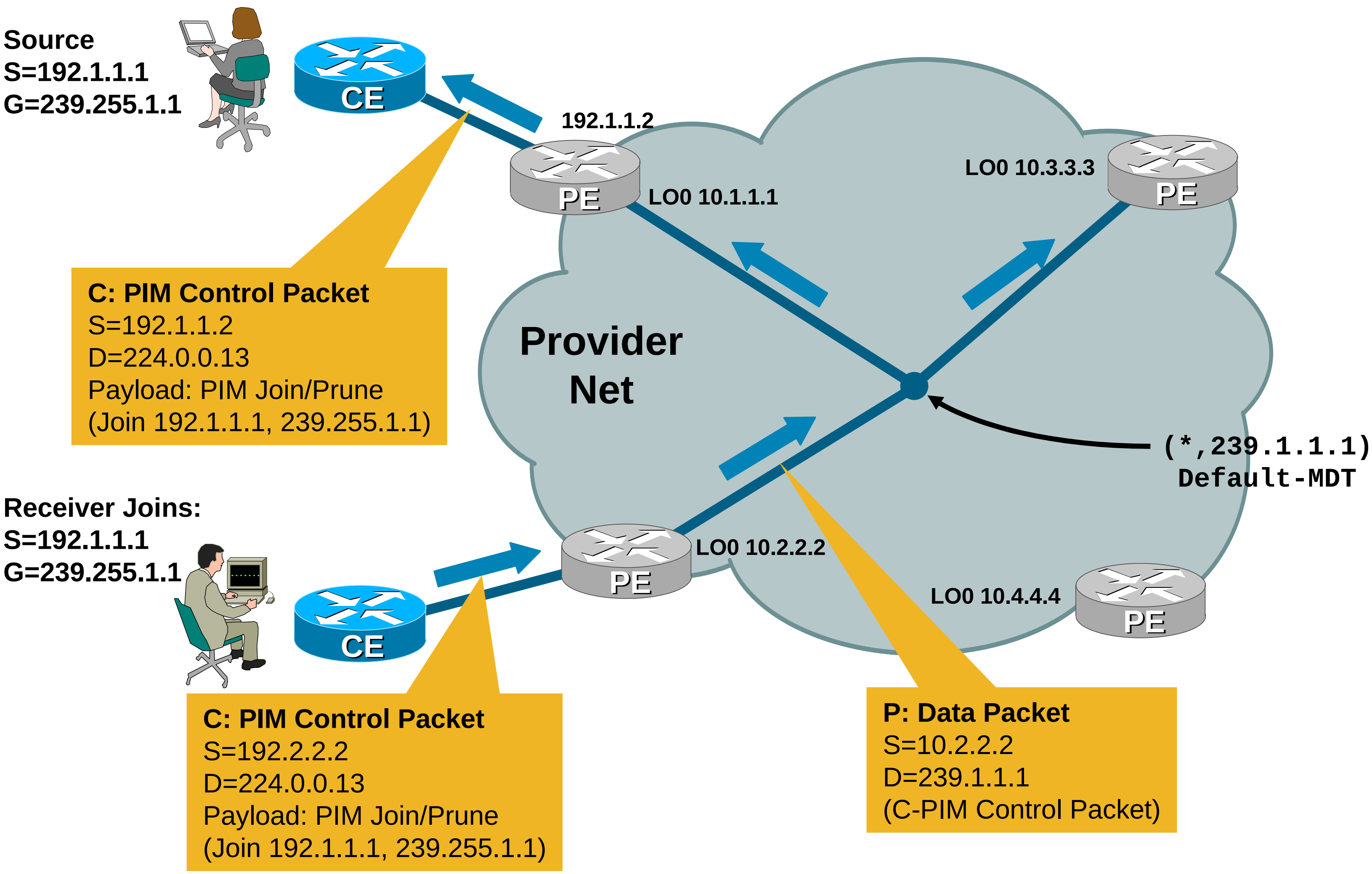
- Each Multicast Domain consists of a Default-MDT
- Each Default-MDT uses a separate Multicast Group inside of Provider's Network

Two Types of MDT Groups

- Default MDT Groups
  - Configured for every MVRF if MPLS or IP core network present
  - Used for PIM control traffic, low bandwidth sources, and flooding of dense-mode traffic
    - MI-PMSI (2547bis-mcast)
- Data MDT Groups
  - Optionally configured
  - Used for high bandwidth sources to reduce replication to uninterested PEs
    - S-PMSI (2547bis-mcast)

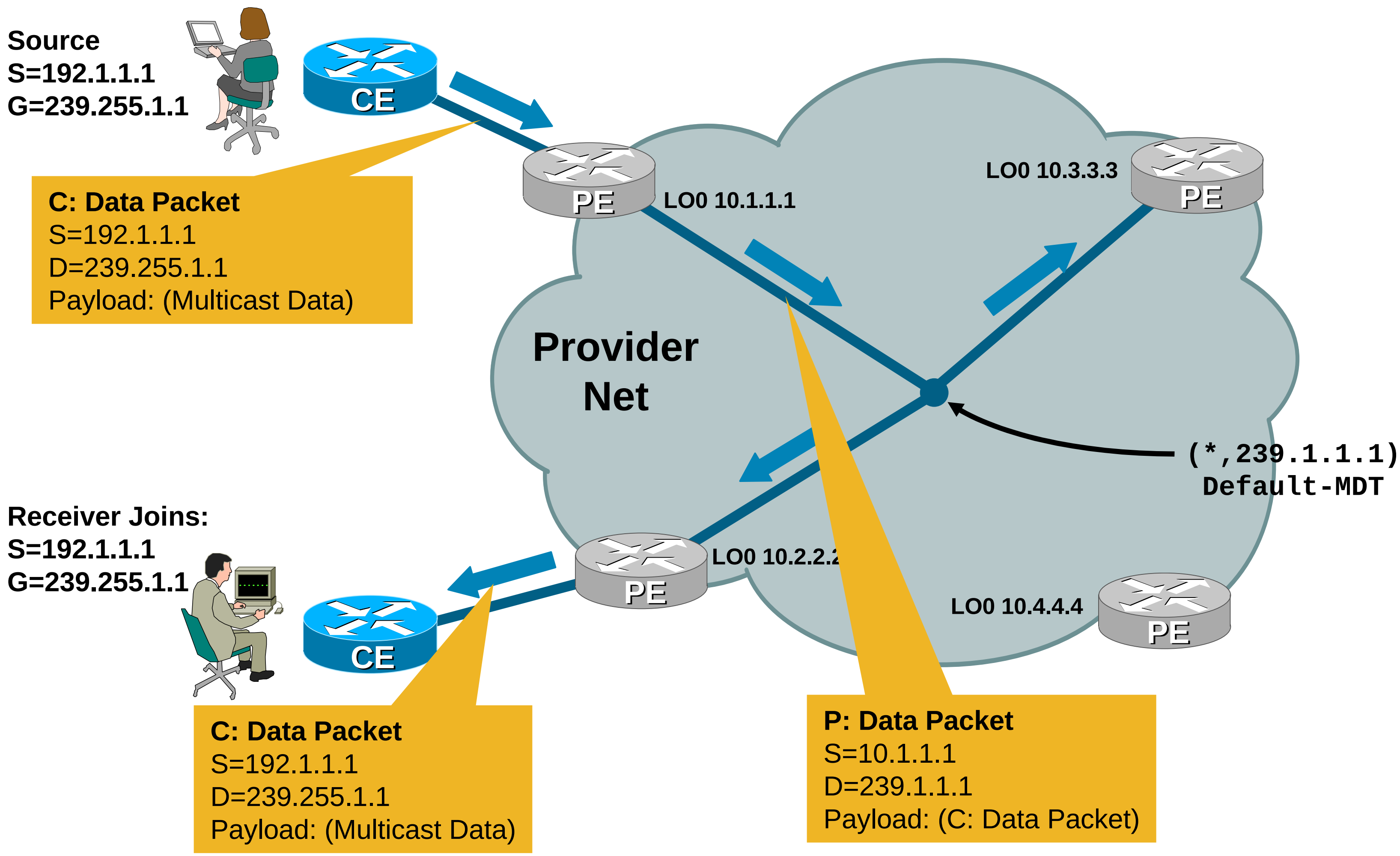
Default MDT: A Closer Look

PIM Control Traffic Flow



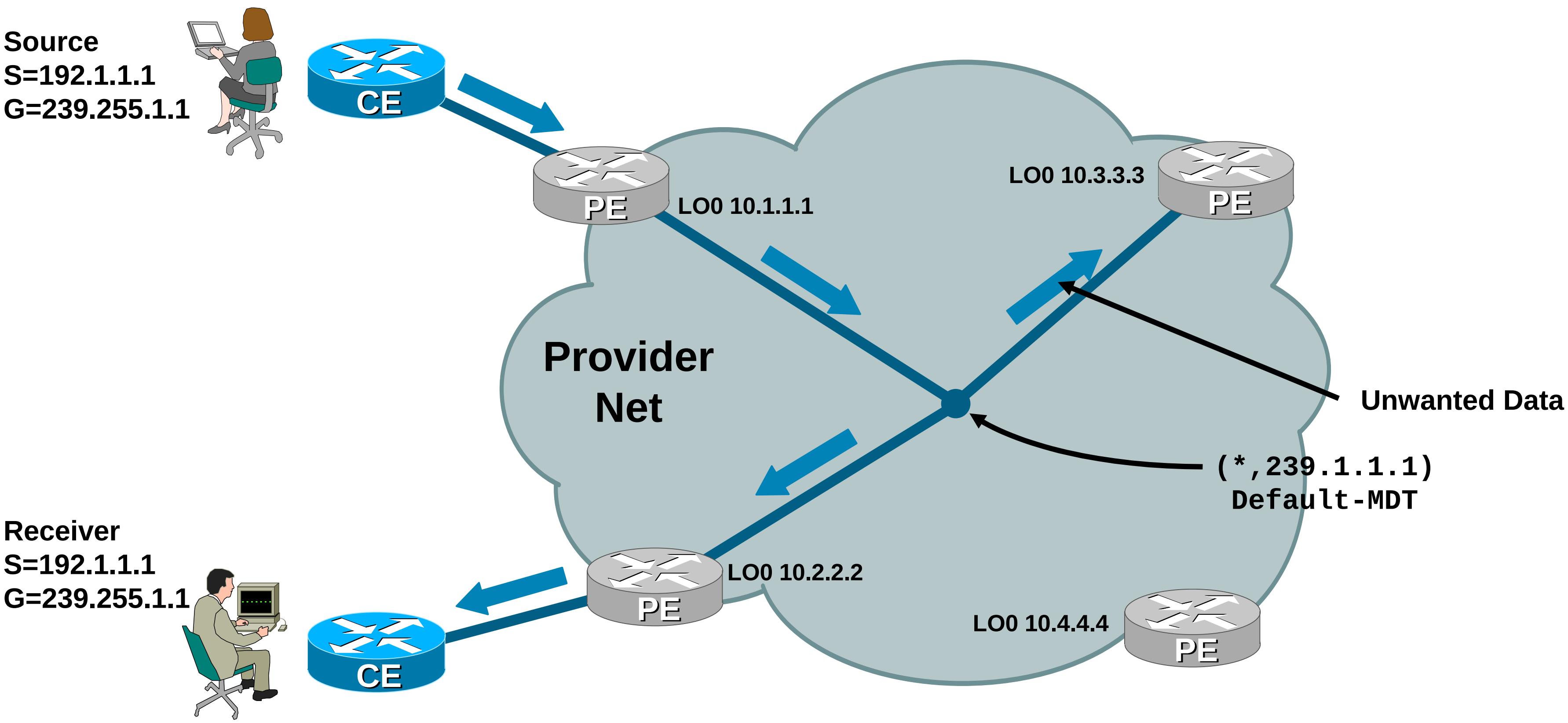
Default MDT: A Closer Look

Multicast Data Traffic Flow



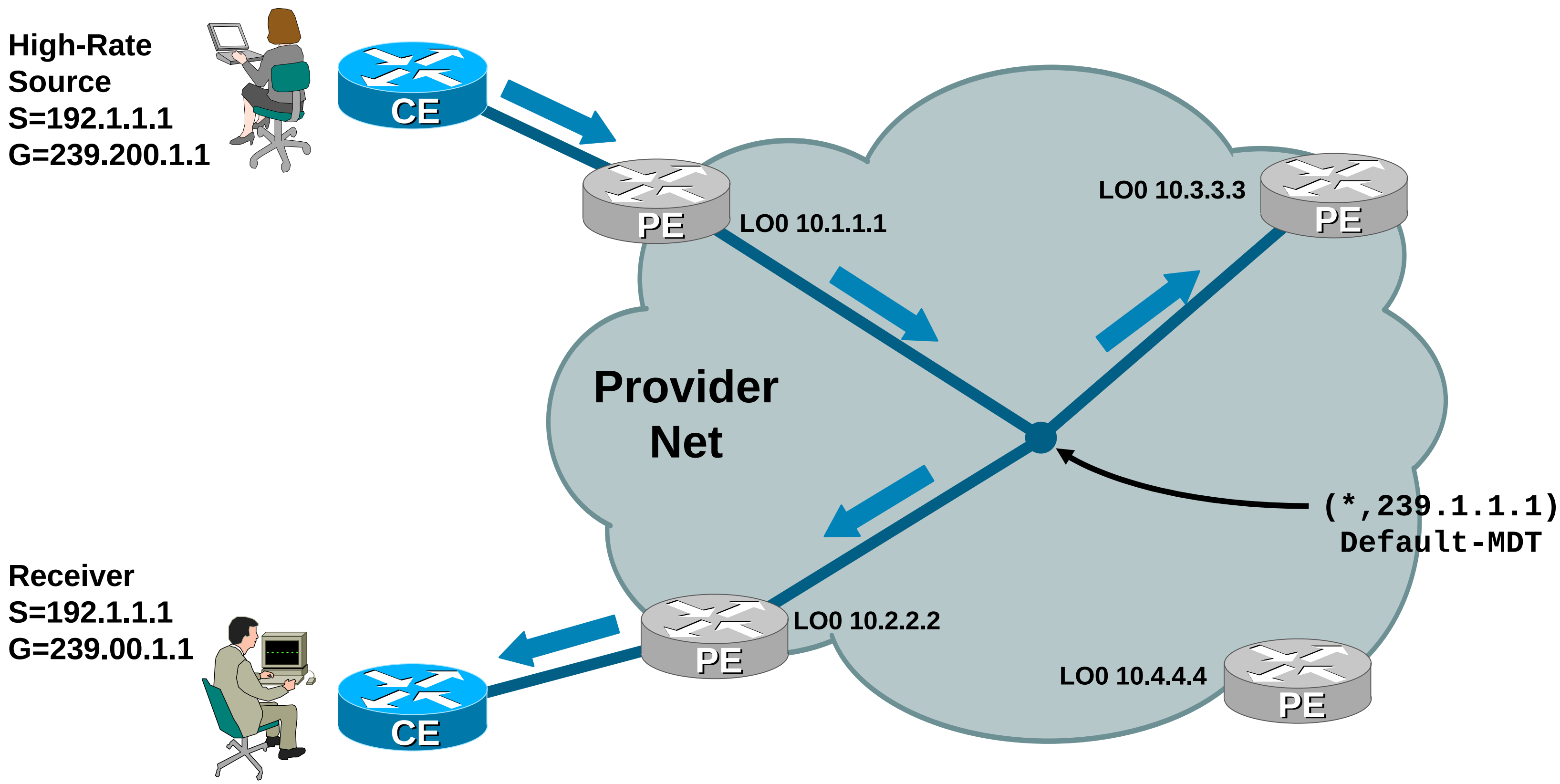
Default MDT: A Closer Look

Advantages and Disadvantages



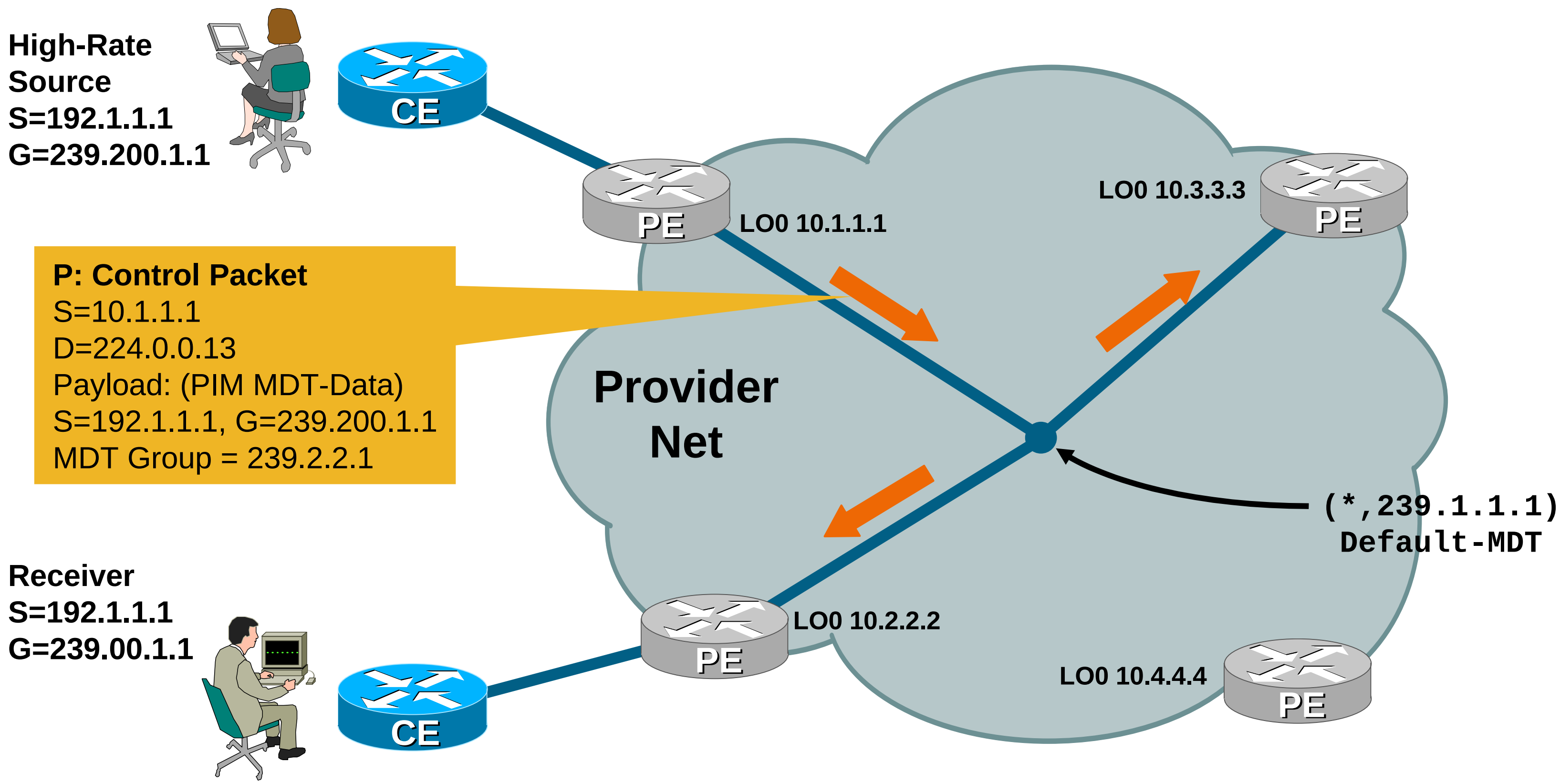
- **Advantage:** Reduces multicast state in the P routers in the core
- **Disadvantage:** Can result in wasted bandwidth
- **Solution:** Use separate Data-MDTs for high rate sources

Data MDTs: Concepts



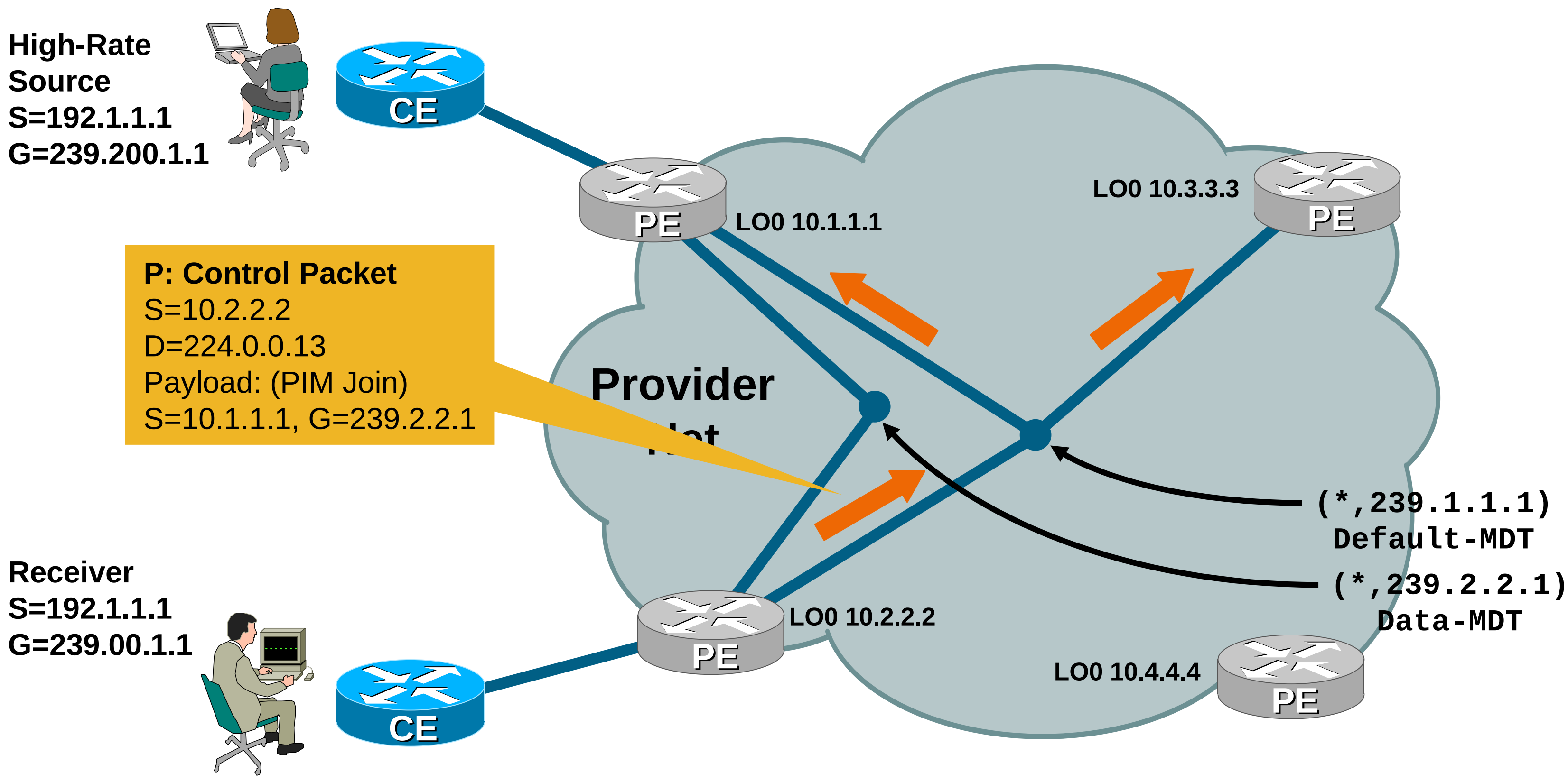
- Traffic exceeds Data-MDT threshold configured on PE router

Data MDTs: Concepts



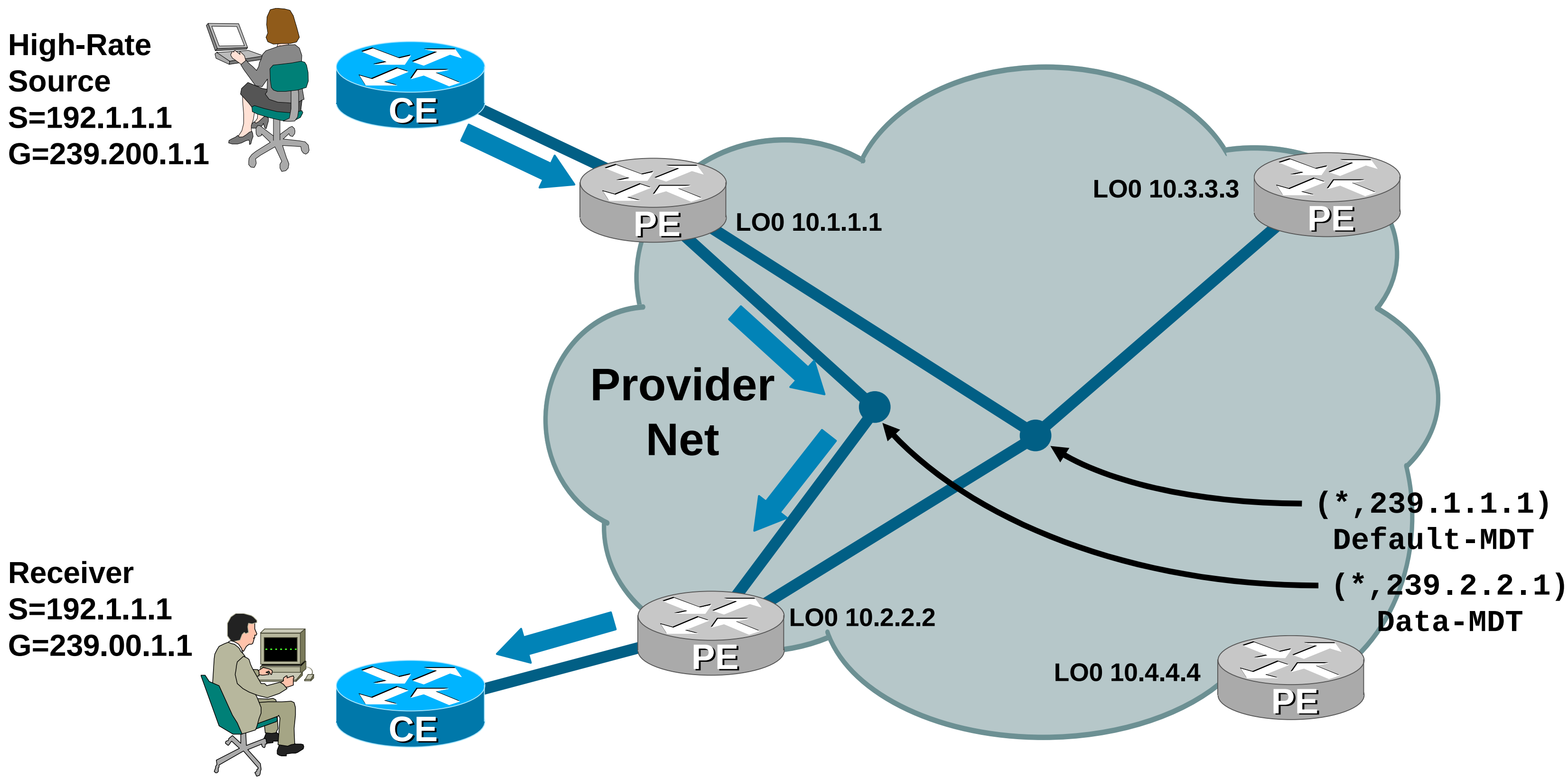
- PE router signals switch to Data-MDT using new group, 239.2.2.1

Data MDTs: Concepts



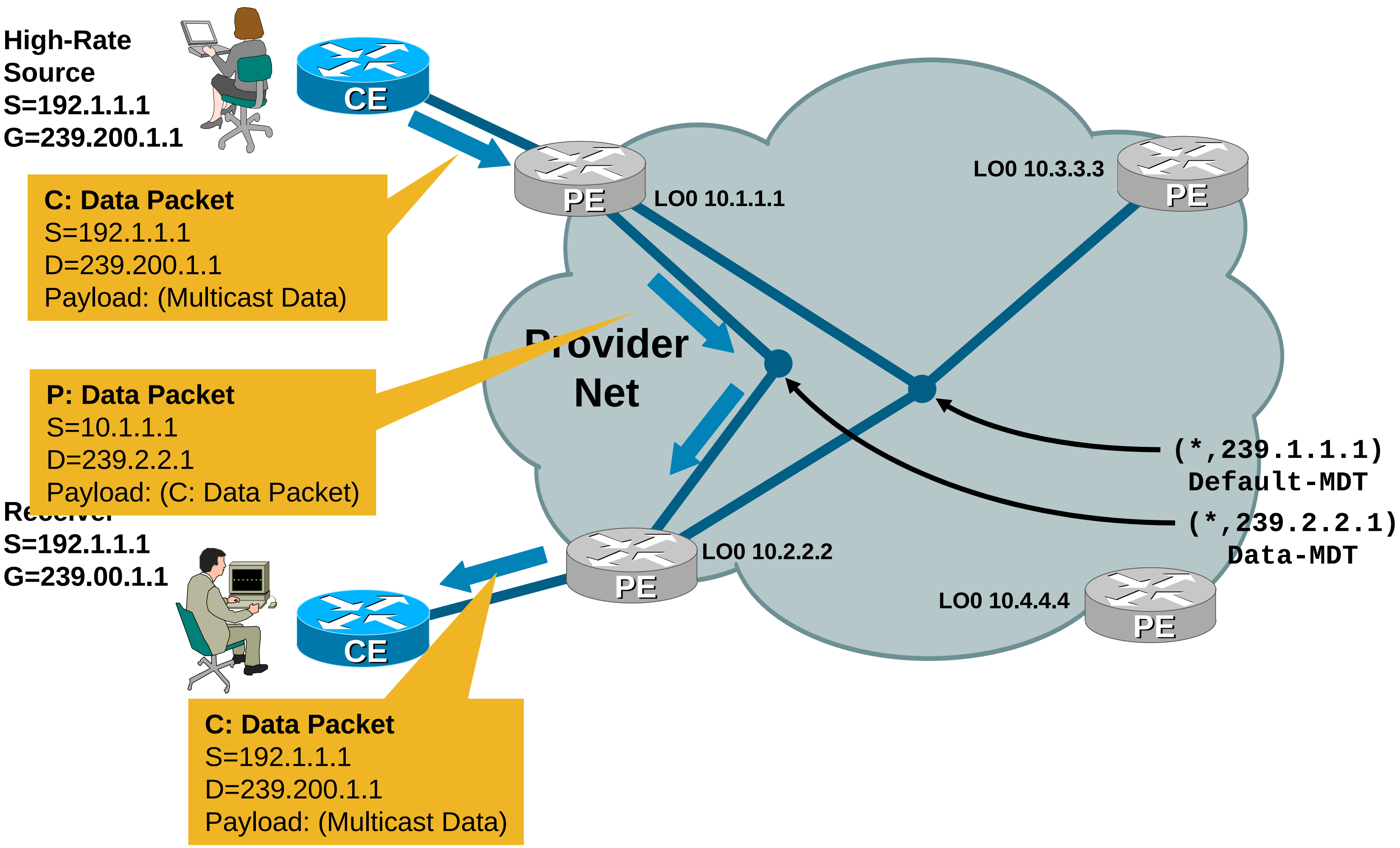
- PE routers with receivers sends Join to group 239.2.2.1
- Data-MDT is built using group 239.2.2.1

Data MDTs: Concepts



- High-rate data begins flowing via Data-MDT
- Data only goes to PE routers that have receivers

Data MDTs: Concepts

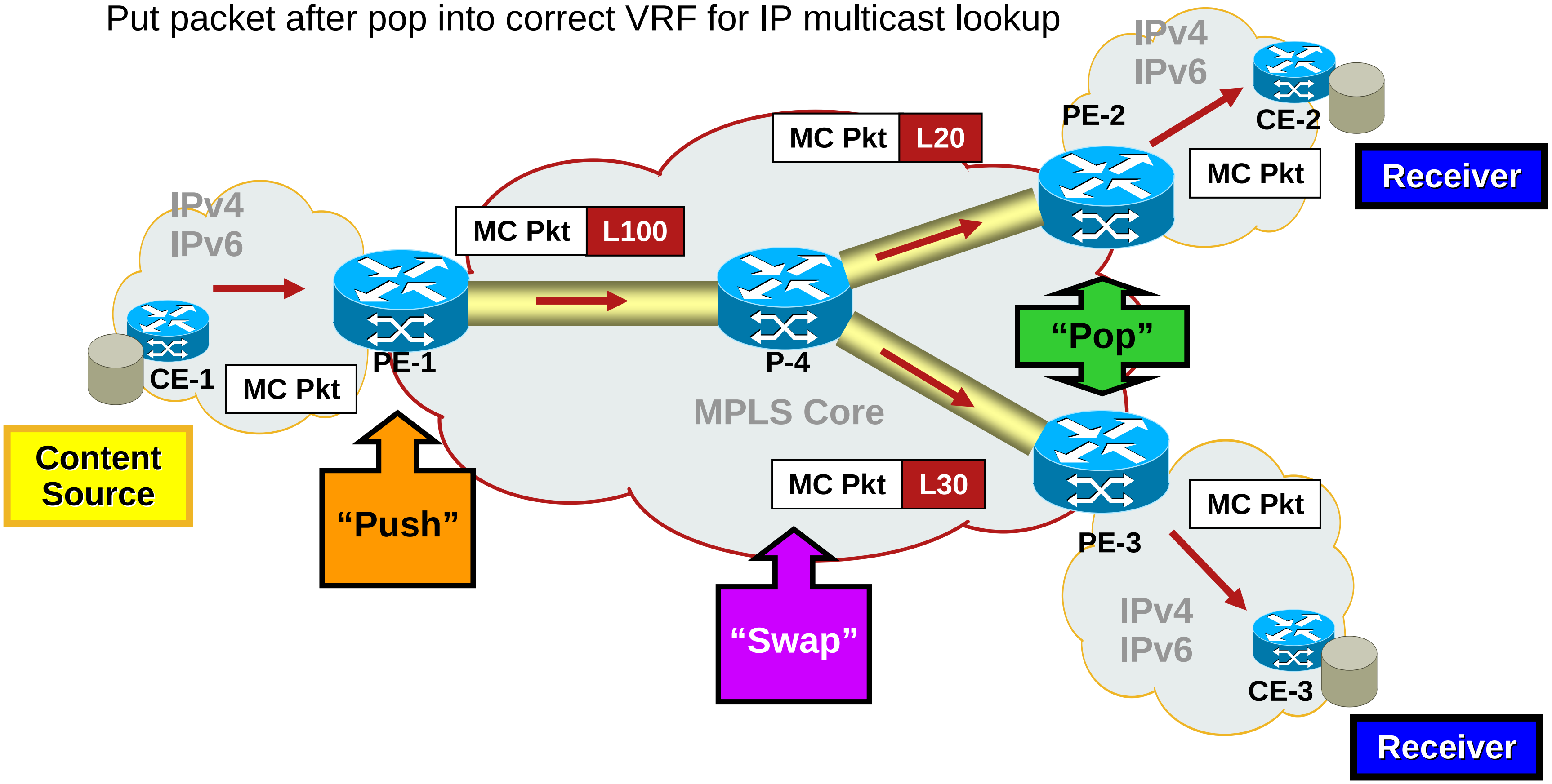


MVPN: Supporting Multiple Tree Types

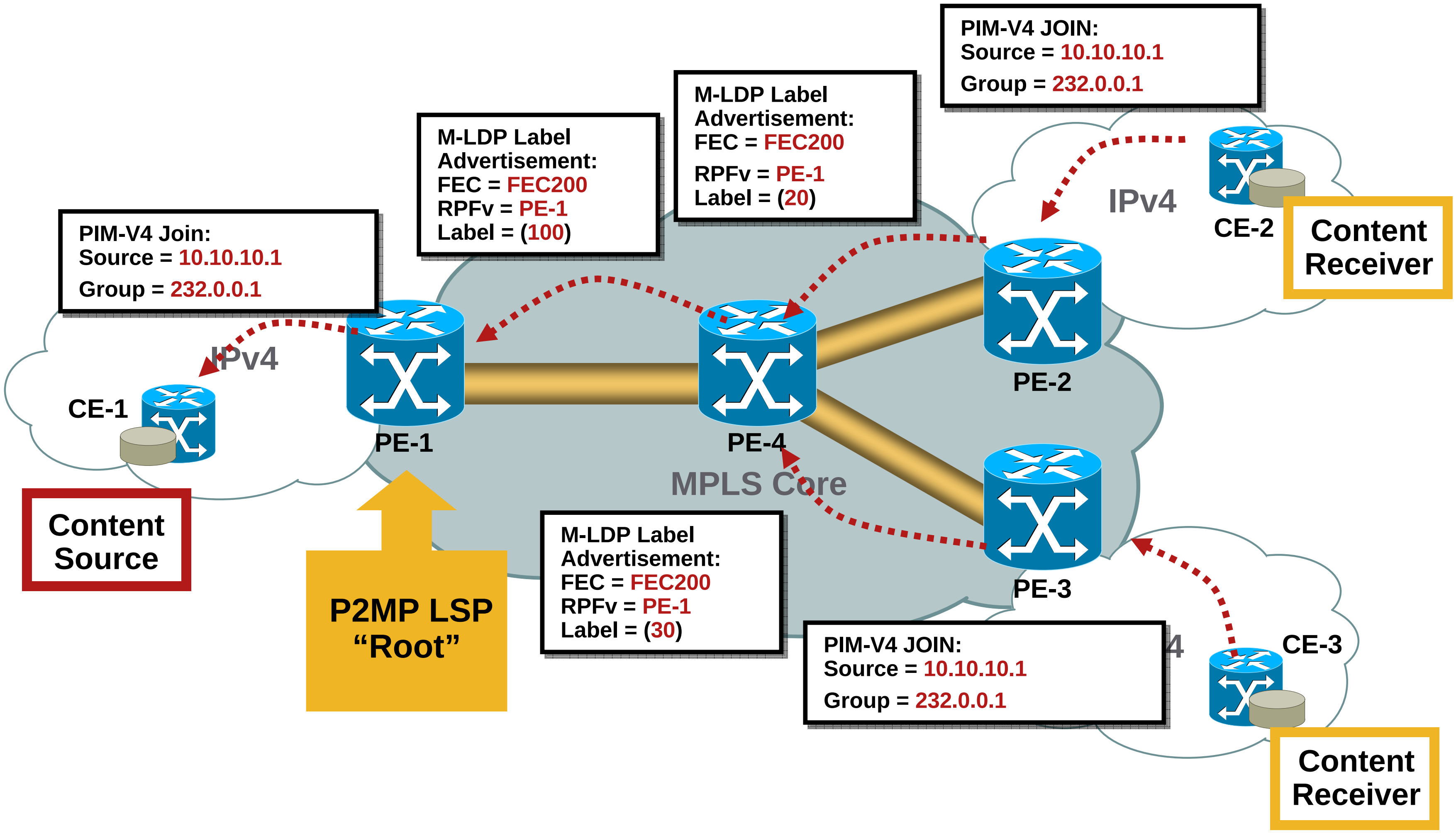
- Key Concept: Separation of a service (PMSI) from its instantiation (tunnels)
- Each PMSI is instantiated using a set of one or more tunnels
- Tunnels may be built by:
  - PIM (any flavor)
  - mLDP p2mp or mp2mp
  - RSVP-TE p2mp
  - Combining unicast tunnels with ingress PE replication
- Can map multiple PMSIs onto one tunnel (aggregation)
- Encaps a function of tunnel, not service
- Single provider can mix and match tunnel types

MPLS traffic forwarding

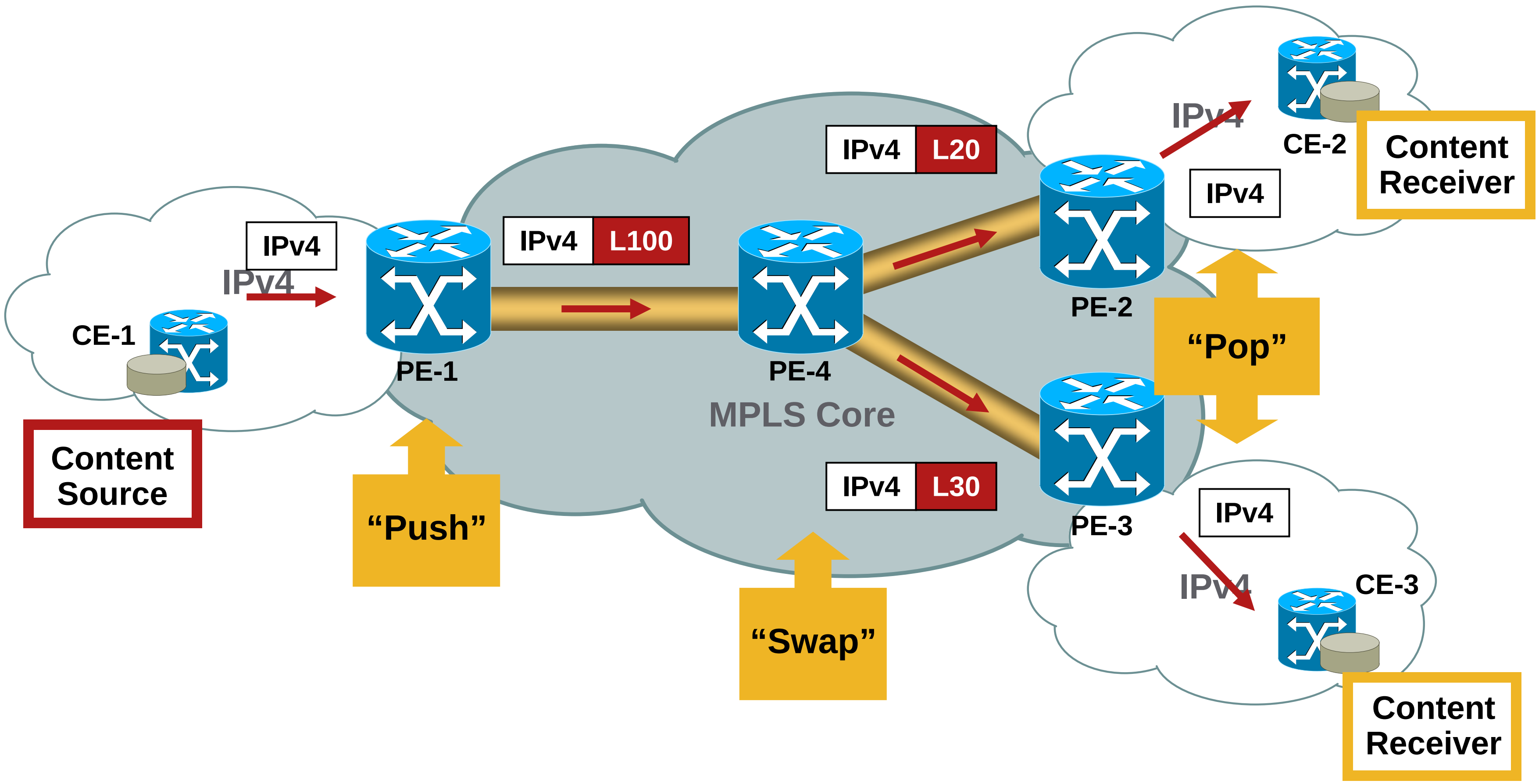
- Same forwarding (HW requirements) with mLDP / RSVP-TE
- Initial: “Single label tree” for both non-aggregated & aggregated
- No PHP: receive PE can identify tree
  - Put packet after pop into correct VRF for IP multicast lookup



MLDP: Transiting SSM (IPv4 Non-VPN)



mLDP: Transiting SSM (IPv4 Non-VPN)

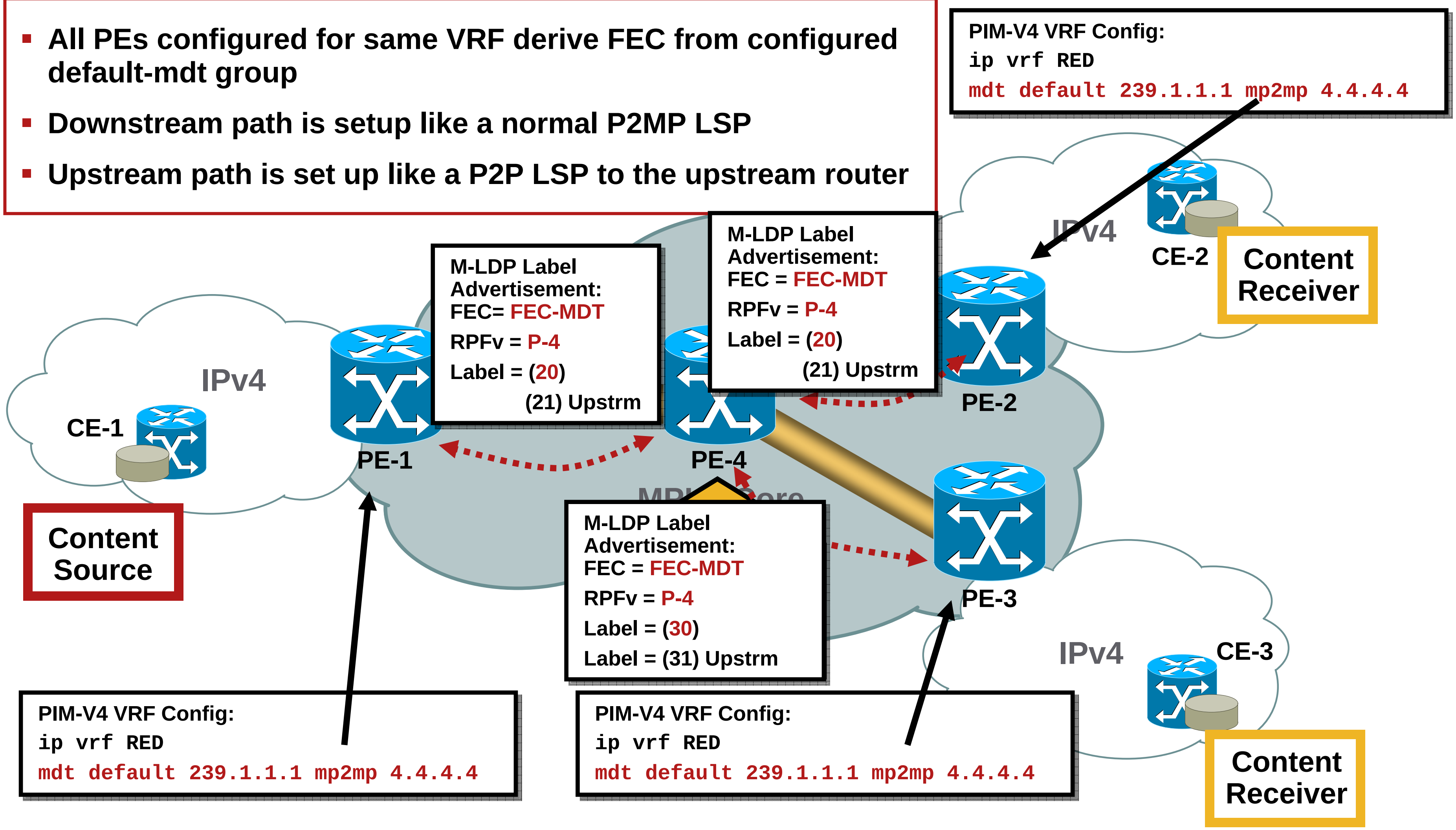


Multicast LDP-Based Multicast VPN

(Default-MDT)

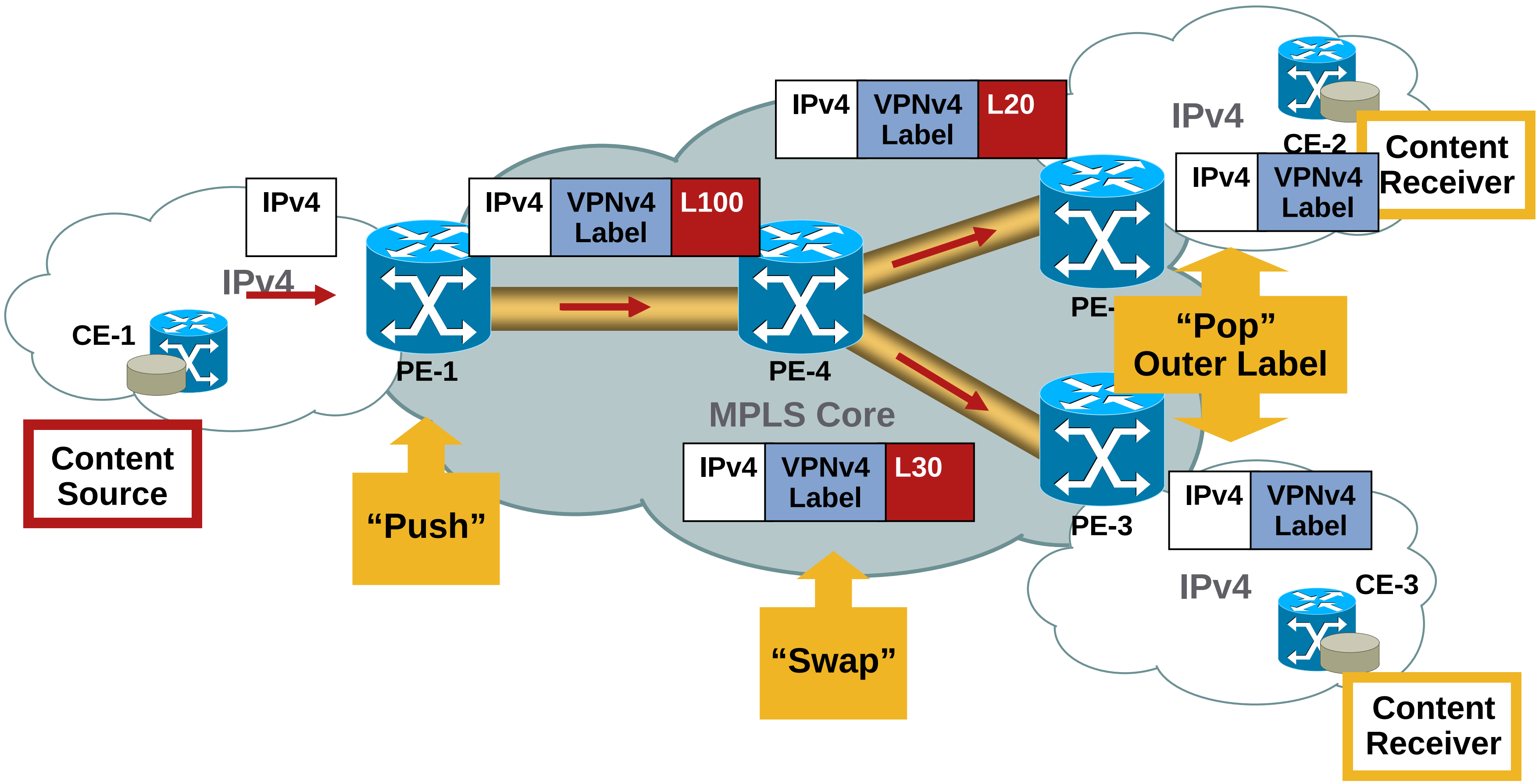
MP2MP Tree Setup Summary

- All PEs configured for same VRF derive FEC from configured default-mdt group
- Downstream path is setup like a normal P2MP LSP
- Upstream path is set up like a P2P LSP to the upstream router

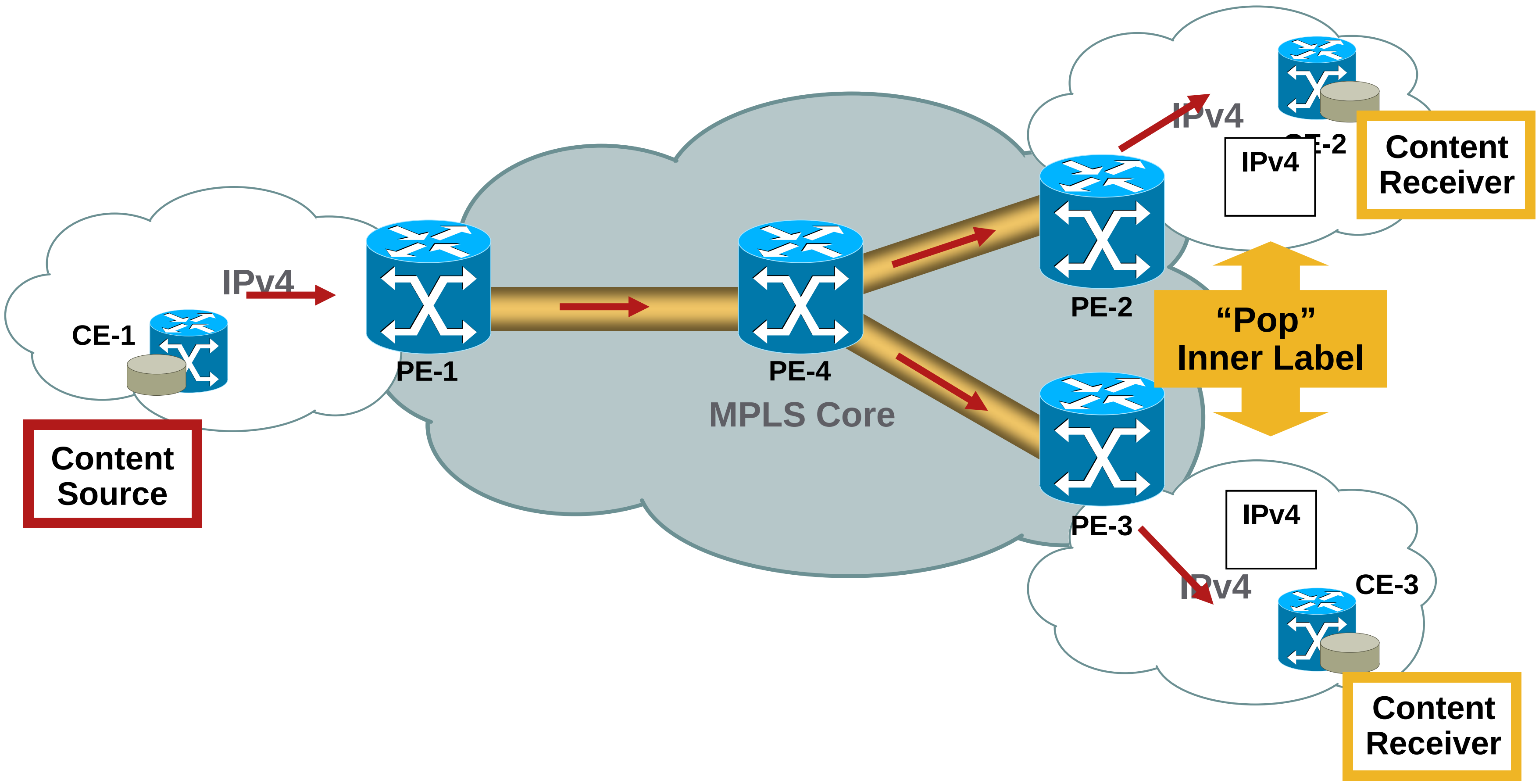


Multicast LDP-Based Multicast VPN

(Default-MDT)



Multicast LDP-Based Multicast VPN
(Default-MDT)



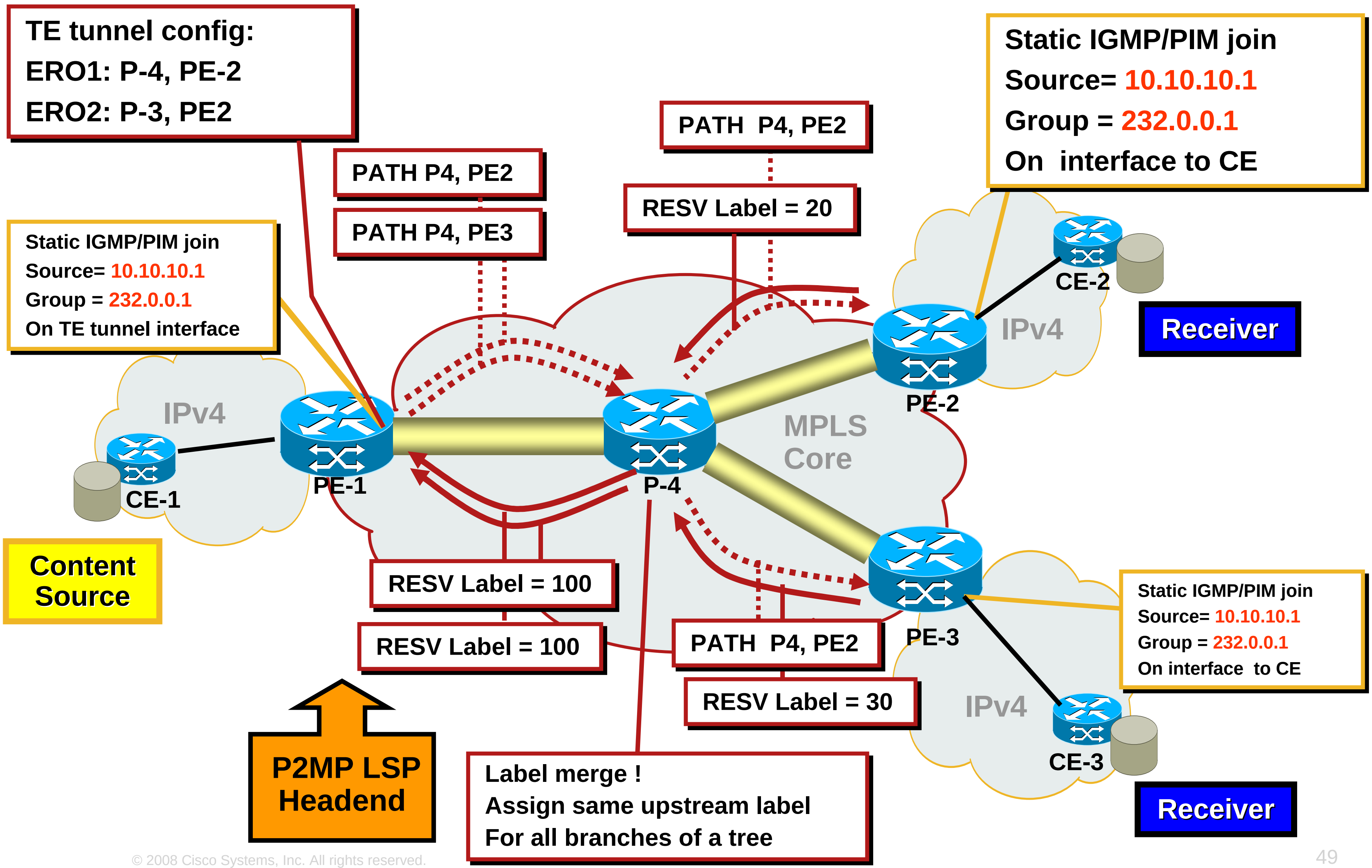
mLDP signaling
Summary

- **Best of PIM + MPLS**
  - Receiver side originated explicit joins – scalable trees
  - PIM-SSM = mLDP P2MP, Bidir-PIM ~= mLDP MP2MP
  - RPF-vector implicit (mLDP root)
- **Best of LDP**
  - Neighbor discovery, graceful restart, share unicast TCP session
  - No interaction for unicast label assignment (ships in the night)
- **Variable length FEC**
  - Allows overlay signaling free 1:1 tree building for ANY (vpn, v6,..) tree
- **All PIM complexity avoided**
  - No direct source/receiver support (DR) (just PE to PE)
  - No PIM-SM (need to emulate), No Bidir-PIM DF process
  - No hop-by-hop RP config (AutoRP, BSR, static) needed)
  - No asserts, other data-triggered events

Combinations with L3 on PE
with RSVP-TE P2MP

- RSVP-TE P2MP static / native
  - Core trees statically provisioned on Headend-PE:
    - Set of tailend-PE
    - All IP multicast traffic that need to be passed into the tree.
- RSVP-TE P2MP static in L3VPN context
  - TBD: Possible, some more per-VRF/VPN config
- RSVP-TE P2MP dynamic
  - TBD: MVPN or new PE-PE signaling (work in IETF, vendors)
  - Required / beneficial ?
    - Reason for RSVP-TE often explicit path definition
    - Not as easy predictable dynamic as static

RSVP-TE P2MP signaling
with static native IPv4 to customer



P2MP RSVP-TE Summary

- RSVP-TE P2P LSP
  - Path explicitly (hop-by-hop) built by headend LSR towards tailend LSR
  - RSVP PATH messages answered by RESV message
- P2MP RSVP-TE LSP
  - A P2MP LSP is built by building a P2P LSP for every tailend of P2MP LSP
  - Midpoint LSR performs “label merge” during RESVP:
    - Use same upstream label for all branches
- Almost all details shared with RSVP-TE P2P
  - All RSVP parameters (for bandwidth reservation)
  - ERO or CSPF, affinities
  - link protection
  - Node protection more difficult

PIM/mLDP benefits over RSVP-TE P2MP Examples

- Cost of trees (in node/network)

  N = # tailend LSR (#PE)

  PIM/mLDP P2MP: ~1, RSVP-TE P2MP: ~N

  Full mesh of RSVP-TE P2MP LSP: ~(N * N)

  Bidir-PIM/mLDP MP2MP: ~1

  Summary: No scaling impact of N for PIM/mLDP

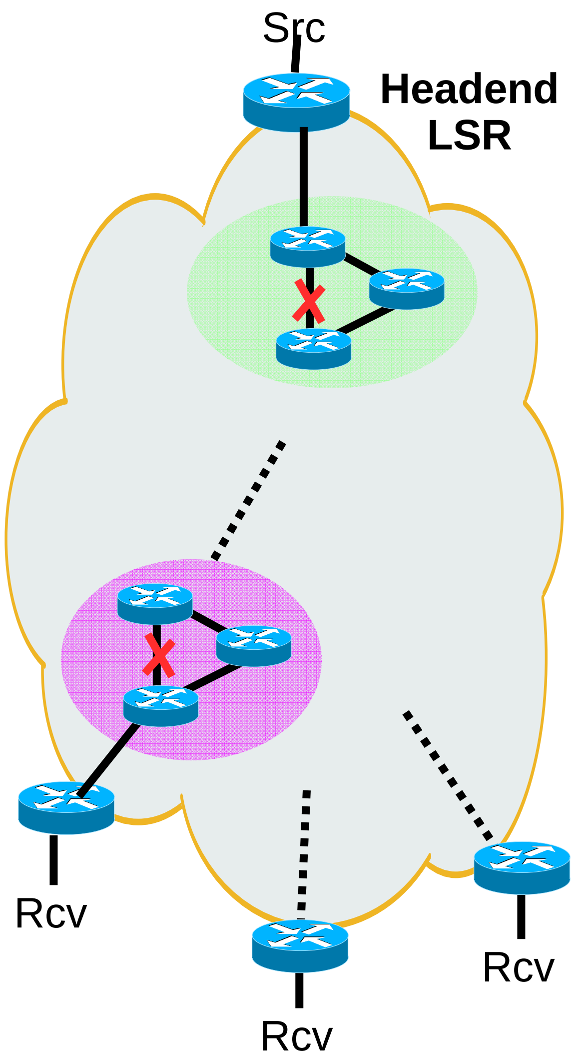
- Locality:

  Affects convergence/reoptimization speed:

  PIM/mLDP: Failure in network affects only router in region (eg: in pink region).

  RSVP: impact headend and all affected midpoint and tailends for RSVP-TE reoptimization.

  Join/leave of members affect only routers up to first router on the tree in mLDP/PIM. Will affect headend and all midpoints in RSVP-TE P2MP.



RSVP-TE P2MP benefits over PIM/mLDP Examples

- Sub 50 msec protection
- Load-split traffic across alternative paths (ECMP or not)

  PIM/mLDP tree follows shortest path, "dense" receiver population == dense use of links

  RSVP-TE P2MP ERO trees (**RED**/**PINK**) under control of headend LSR.

  CSPF load split based on available bandwidth.

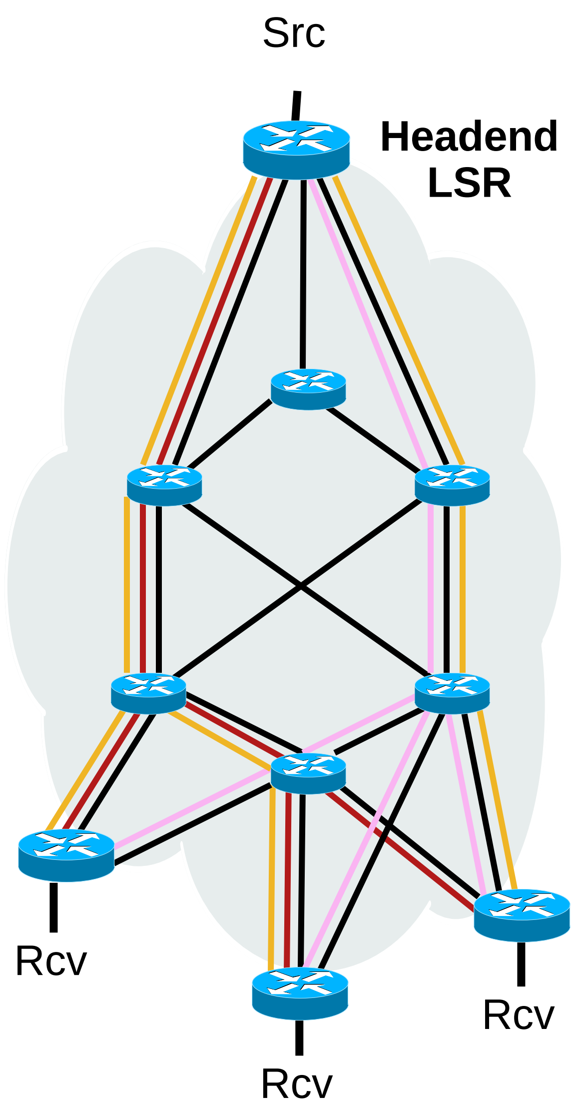
  "Steiner tree" CSPF modifications possible

- Block (stop) trees on redundancy loss

  Assume high-prio and low-prio trees.

  With full redundancy, enough bandwidth to carry all trees (with load-splitting)

  On link-loss, reconverge high-prio, block low-prio



Combining RSVP-TE P2MP and mLDP

- Rule of thumb:

  Think of mLDP and RSVP-TE P2MP as multicast versions of unicast counterparts (LDP, RSVP-TE)

  Use whenever unicast equivalent is used.

- Can run RSVP-TE P2MP and mLDP in parallel

  Each one running PE-PE – ships in the night !

- Can not combine RSVP-TE P2MP / mLDP along path !!!

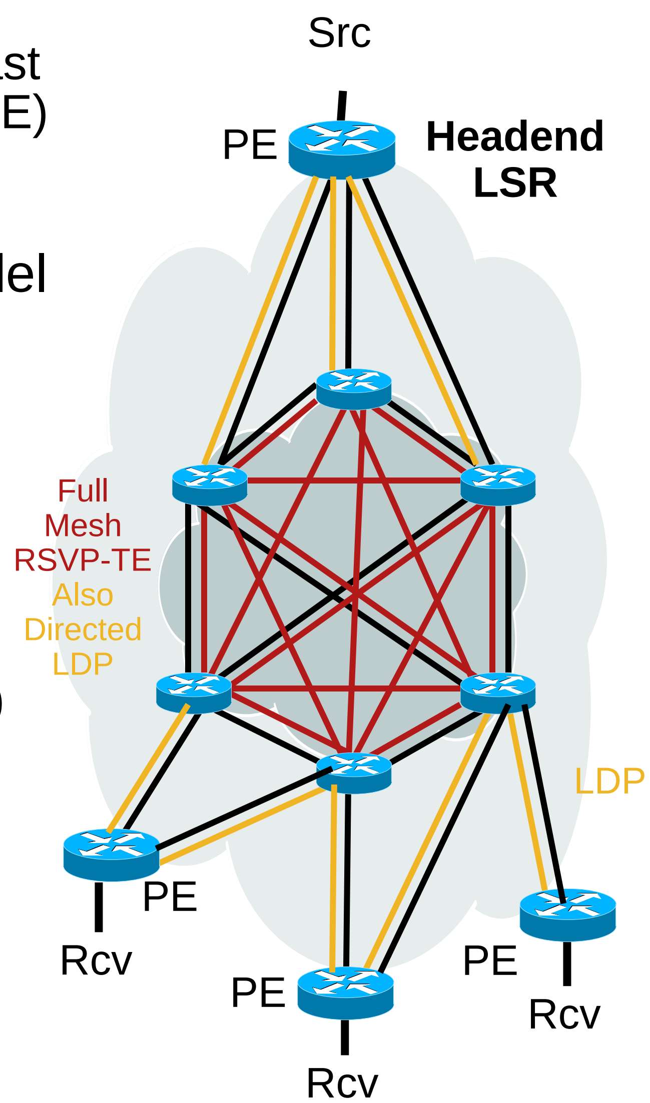
  Standard unicast design: full mesh RSVP-TE between P nodes, LDP on PE-P links.

  Limit size of full-mesh (RSVP-TE scalability)

  Multicast: to map mLDP tree onto RSVP-TE P2MP tree, P nodes would need to logically be 'PE' – running all PE-PE signaling (eg: P node running BGP-join extensions).

  NOT DESIGNED / SUPPORTED

  Static designs with PIM PE-P possible though (and RSVP-TE between P nodes)



L2VPN Considerations

- L2 preferred by non-IP ‘communities’
  - IP address transparency (unicast only issue)
  - PE “invisible” = customer free to choose protocols independent of provider
    - Not true if PE uses PIM/IGMP snooping!
- No (dynamic) P/PE L2 solution with P2MP trees
  - VPLS: full-mesh/hub&spoke P2P pseudowire only
  - Non P/PE models available: single-hop protected pseudowires
  - Recommended directions:
    - TBD: define how to use mLDP for L2VPN (VPLS)
    - Most simple: one mLDP MP2MP LSP per L2VPN (broadcast)
  - Recommend not to use IGMP/PIM snooping on L2VPN-PE!
    - Unless customer is provider (e.g., broadband-edge design)

Transit technologies for IPTV Summary / recommendations

- Native PIM-SSM + RPF-Vector

  Most simple, most widely deployed, resilient solution.

- PIM based MVPN

  Also many years deployed (IOS, JUNOS, TIMOS).

  Recommended for IPTV when VRF-isolation necessary

- mLDP

  Recommended Evolution for MPLS networks for all IP multicast transit:

  ‘Native’ (m4PE/m6PE)

  ‘Direct-MDT/MVPN-mLDP’ (IPv4/IPv6)

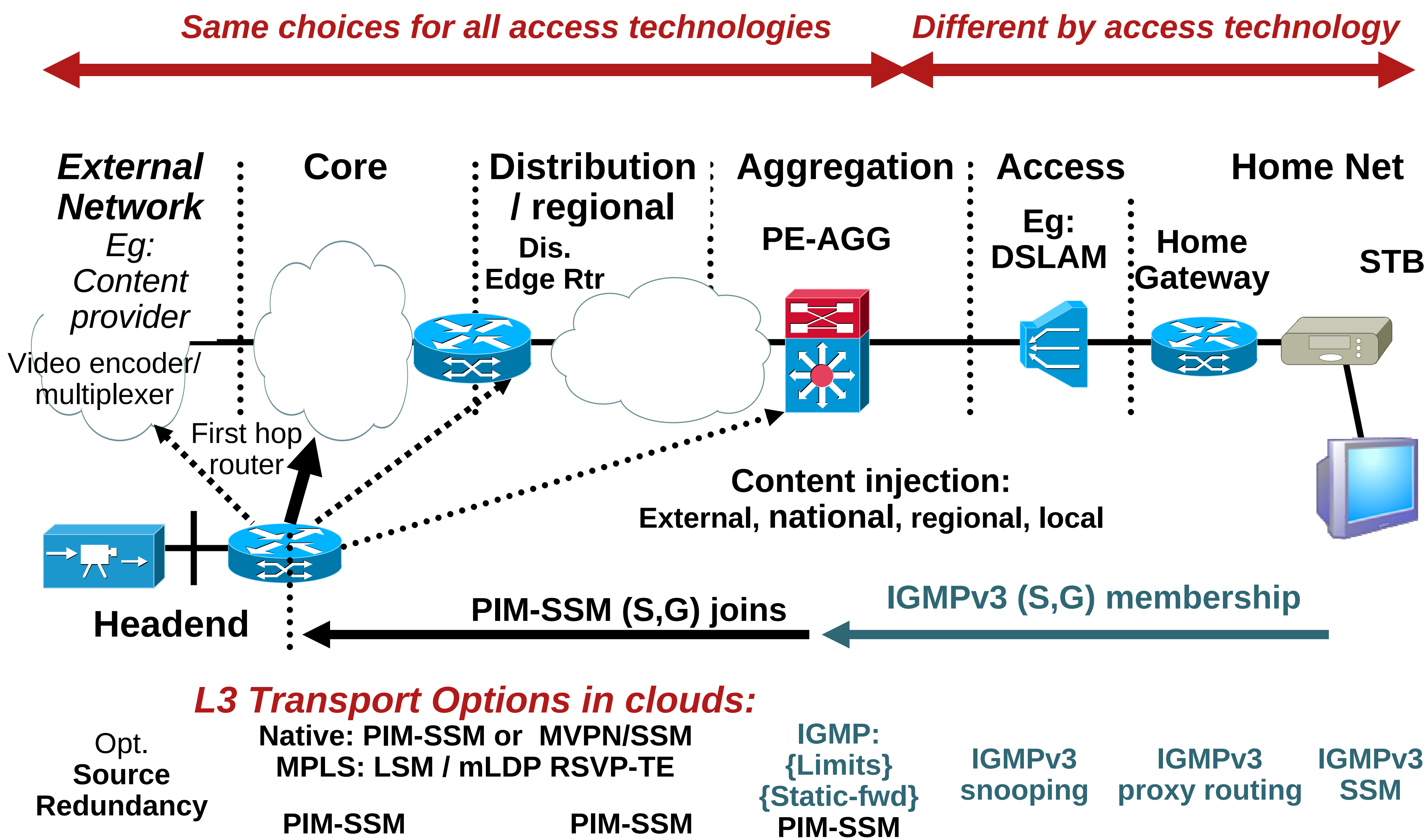
- RSVP-TE P2MP

  Strength in TE elements (ERO/CSPF + protection)

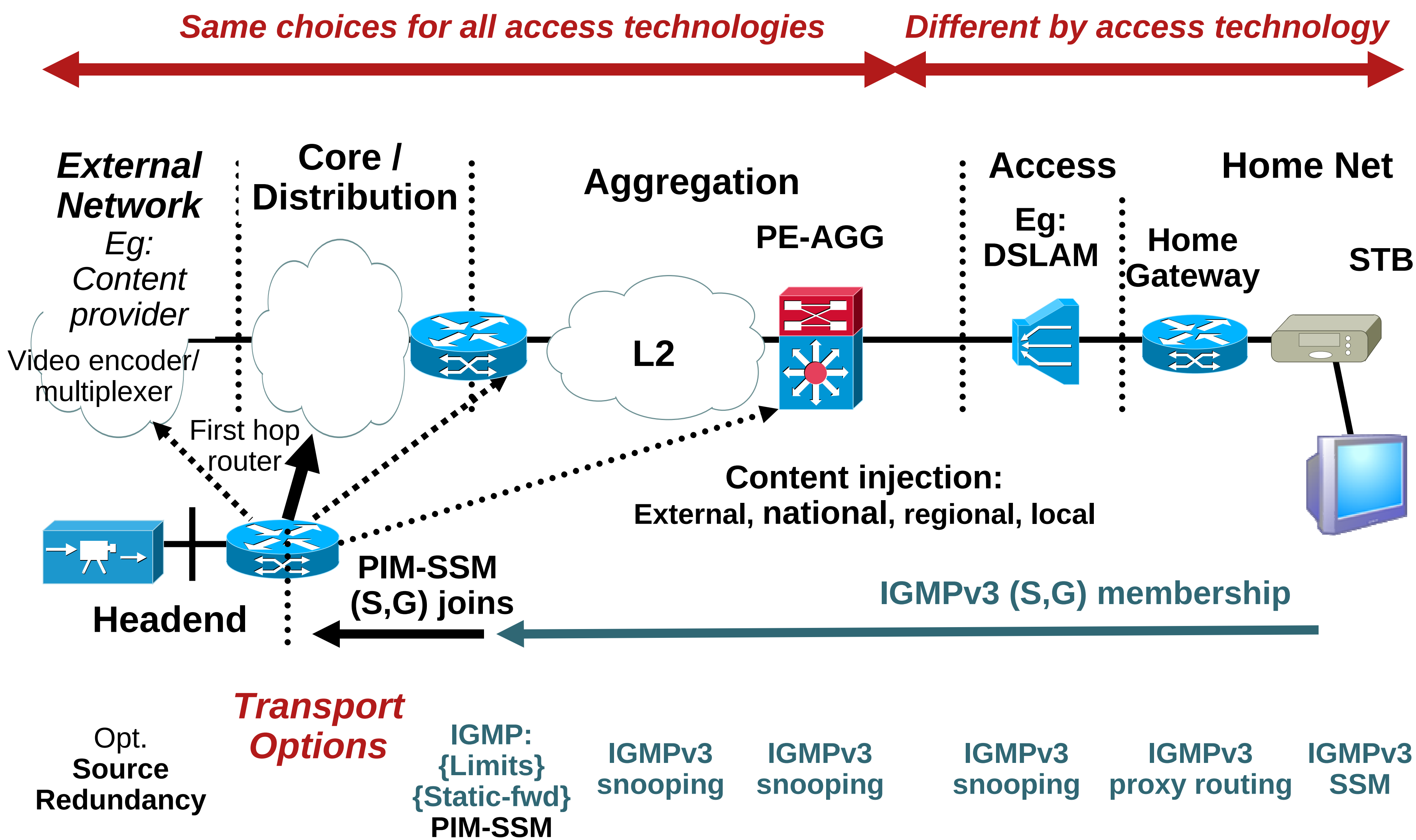
  Recommended for limited scale, explicit engineered designs, eg: IPTV contribution networks.

Broadband edge IP multicast

End-to-end protocol view
DSL, L3 aggregation



End-to-end protocol view
DSL, L2 aggregation



IGMP
snooping vs. proxy routing

- **IGMP snooping:**

  Performed by L2 switch. Intended to the transparent. Many vendor variations.

  IETF RFC 4541 – INFORMATIONAL ONLY

  Transparent: no snooping messages suppressed

  Report-suppression: guess which IGMP reports are redundant at router (can break explicit tracking, fast leaves).

  Proxy-reporting: fully emulate host.

  IGMPv3: Use source-IP address “0.0.0.0”

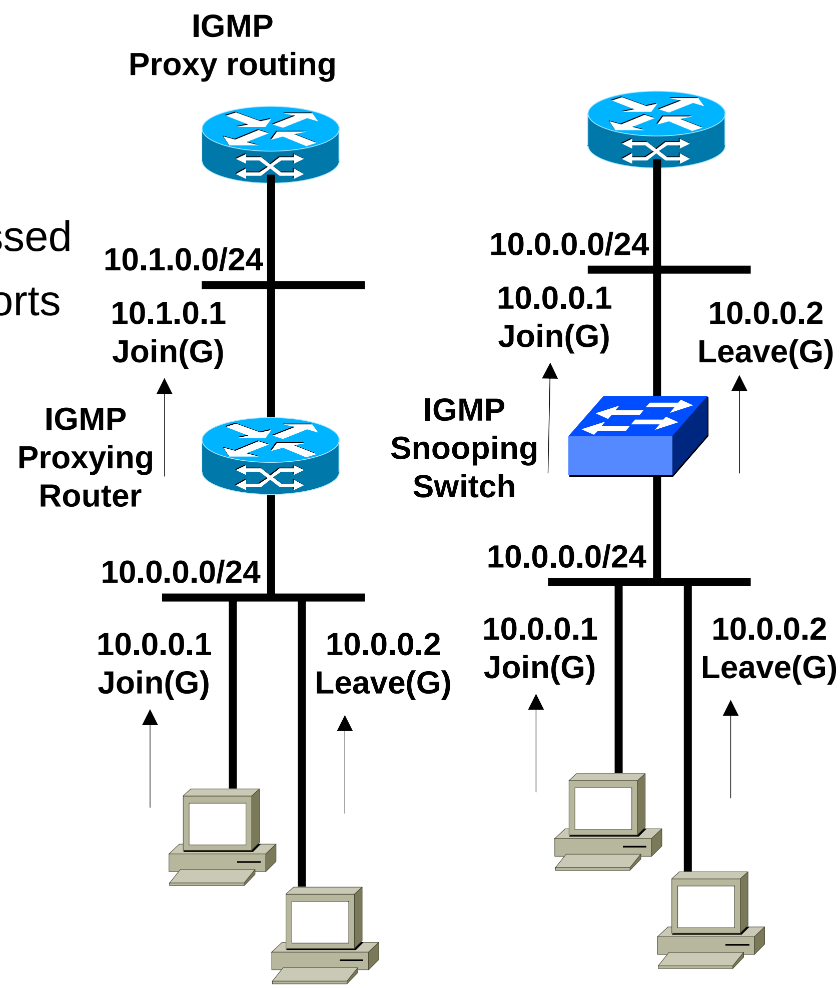
- **IGMP proxy-routing:**

  Performed by router:

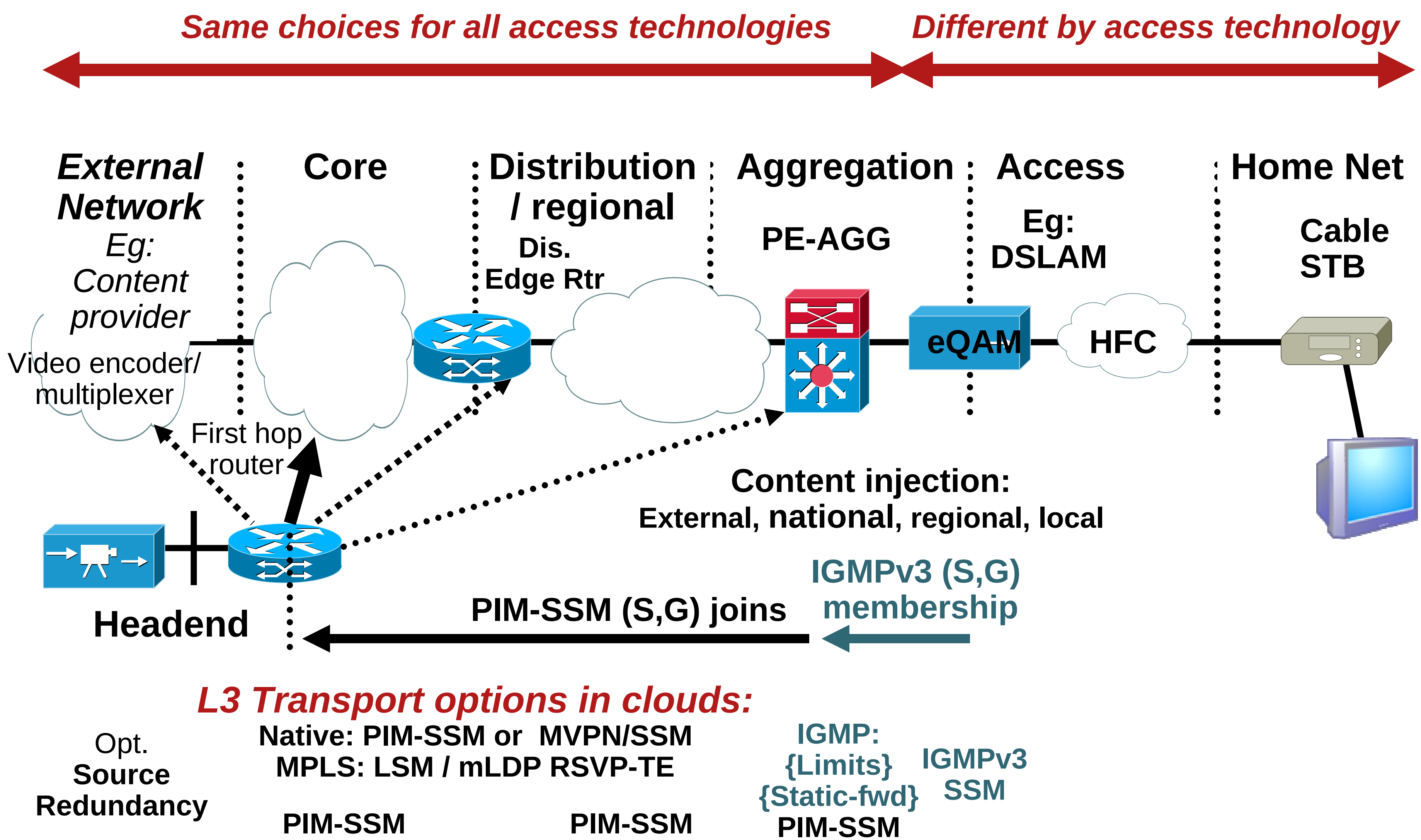
  IETF RFC4605 – STANDARDS TRACK

  IGMP proxy router need to act exactly like a single host on it’s upstream interface.

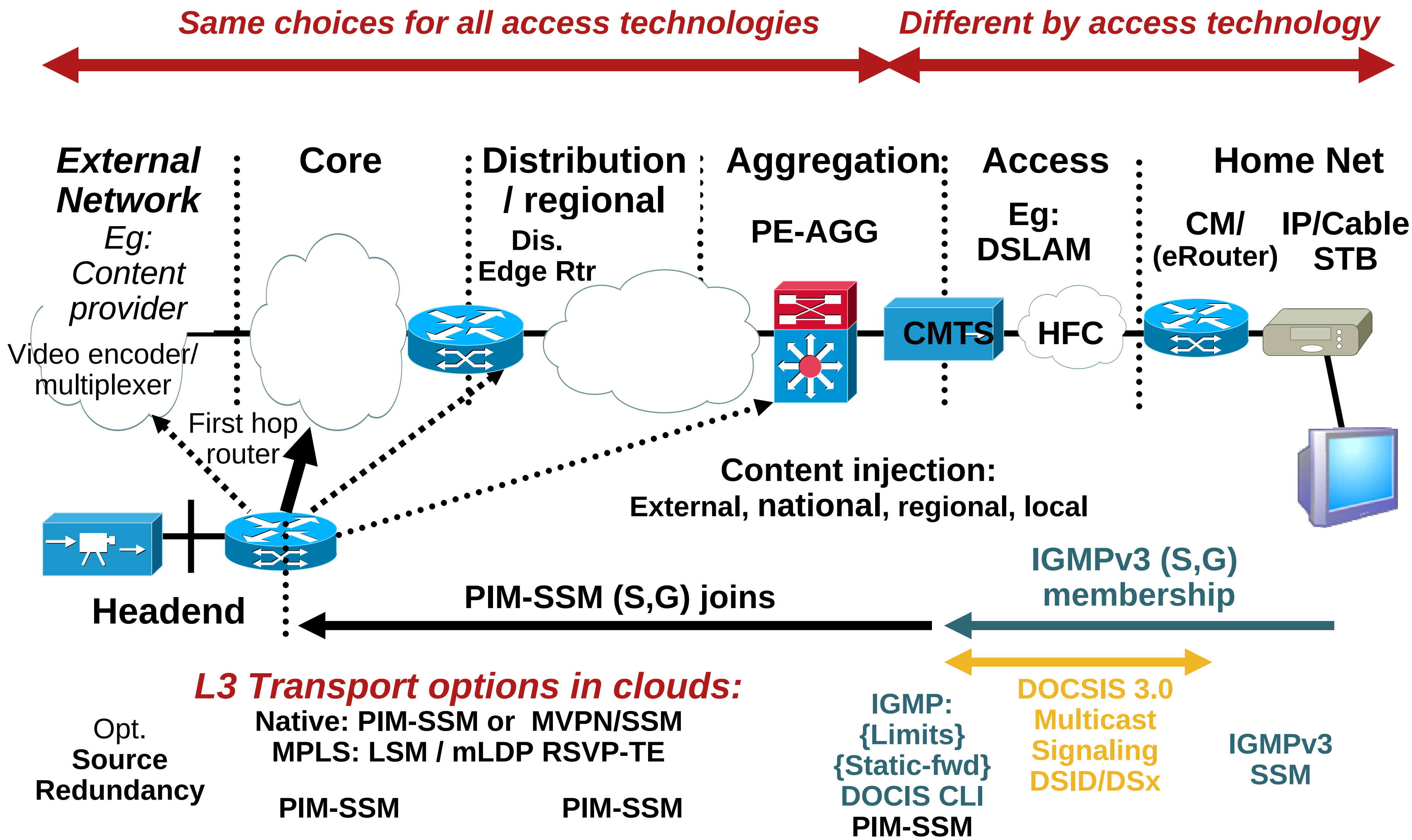
  Router can not transparently pass trough IGMP membership packets from downstream hosts: would have incorrect source-IP addresses.



End-to-end protocol view
digital cable (non DOCSIS)

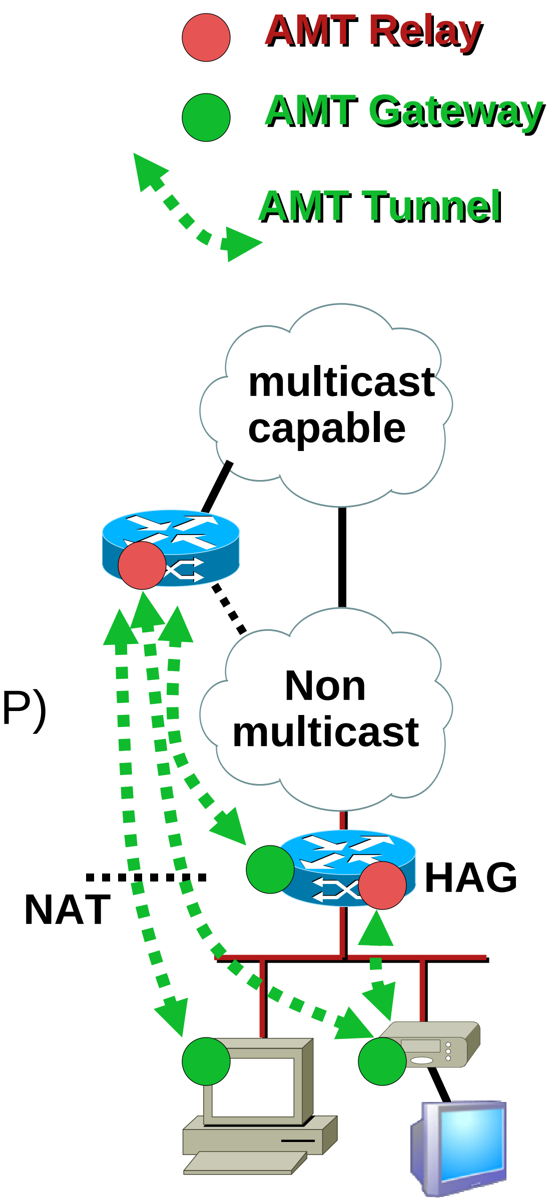


End-to-end protocol view
DOCSIS 3.0 cable

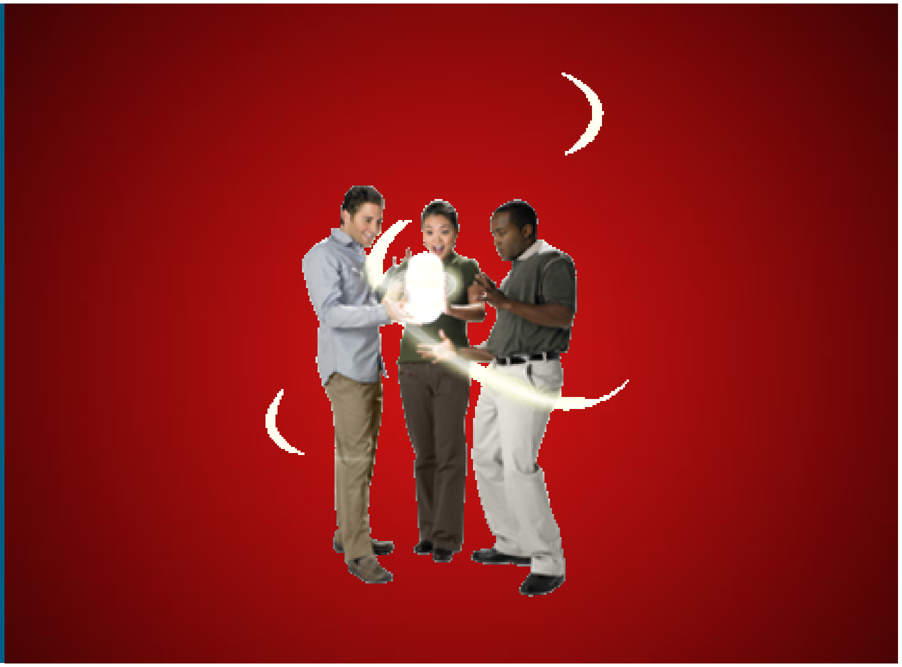


Auto Multicast Tunneling (AMT)

- Tunnel through non-multicast enabled network segment
  - Draft in IETF ; Primarily for SSM
  - GRE or UDP encap
  - Relay uses well known 'anycast' address
- Difference to IPsec, L2TPv3, MobileIP, …
  - Simple and targeted to problem
  - Consideration for NAT (UDP)
  - Ease implemented in applications (PC/STB) (UDP)
- Variety of target deployment cases
  - Relay in HAG – provide native multicast in home
  - Gateway in core-SP – non-multicast Access-SP
  - Access-SP to Home - non-multicast DSL
  - In-Home only – eg: multicast WLAN issues



Resiliency



Failure Impact Upon Viewer Experience

- Very hard to measure and quantify
- If I frames or frame-information is lost, impact will be for a whole GOP

  GOP can be 250 msec (MPEG2) .. 10 sec (WM9)

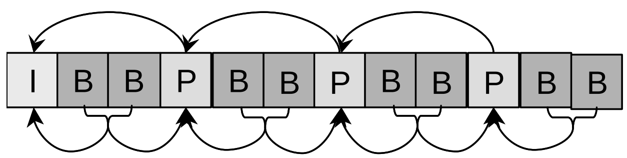
- Encoding and intelligence of decoder to “hide” loss impact quality as well
- IP/TV STB typically larger playout buffer than traditional non-IP STBs:

  Loss can cause catch-up: no black picture, but just a jump in the motion

- What loss is acceptable?

  Measured in number of phone calls from complaining customers?!

Impact of Packet Loss on MPEG Stream



- Compressed Digitized Video is sent as I, B, P Frames
- I-frames: Contain full picture information
  - Transmit I frames approximately every 15 frames (GOP interval)
- P-frames: Predicted from past I or P frames
- B-frames: Use past and future I or P frames

I-Frame Loss "Corrupts" P/B Frames for the Entire GOP

IP/TV Deployments Today

- Two schools of thought in deployments today:
  - I think I need 50ms convergence
  - IPMulticast is fast enough
- IPMulticast is UDP
  - The only acceptable loss is 0ms
  - How much is “reasonable”?
- 50ms “requirement” is not a video requirement
  - Legacy telco voice requirement
  - Efforts for 50ms only cover a limited portion network events
- Where to put the effort?
  - Make IPMulticast better?
  - Improve the transport?
  - Add layers of network complexity to improve core convergence?

Application Side Resiliency

- FEC: Forward Error Correction
  - Compensate for statistical packet loss
  - Use existing FEC, e.g. for MPEG transport to overcome N msec (>= 50 msec) failures?
  - Cover loss of N[t] introduces delay > N[t]!
- Retransmissions
  - Done e.g. with vendor IP/TV solutions—unicast retransmissions
    - Candidate large bursts of retransmissions!
    - Limit #retransmissions necessary
  - Multicast retransmissions (e.g. PGM ?)
    - No broadcast IP/TV solutions use this

Service Availability Overview

IP Host Components Redundancy

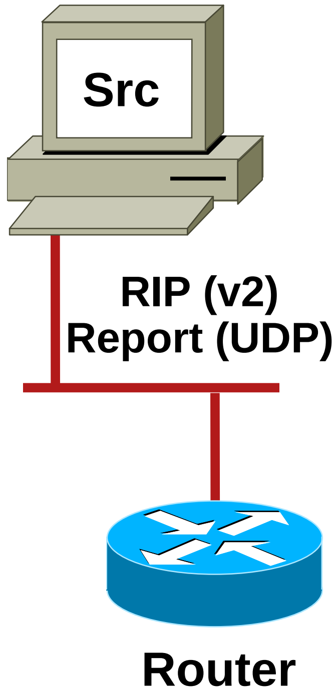
- Single transmission from Logical IP address
  - Anycast—Use closest instance
  - Prioritycast—Use best instance
    - Benefit over anycast: no synchronization of sources needed, operationally easier to predict which source is used
  - Signaling host to network for fast failover
    - RIPv2 as a simple signaling protocol
    - Normal Cisco IOS/IGP configuration used to inject these source server routes into the main IGP being used (OSPF/ISIS)
- Dual Transmission with Path separation

Video Source Redundancy: Two Approaches

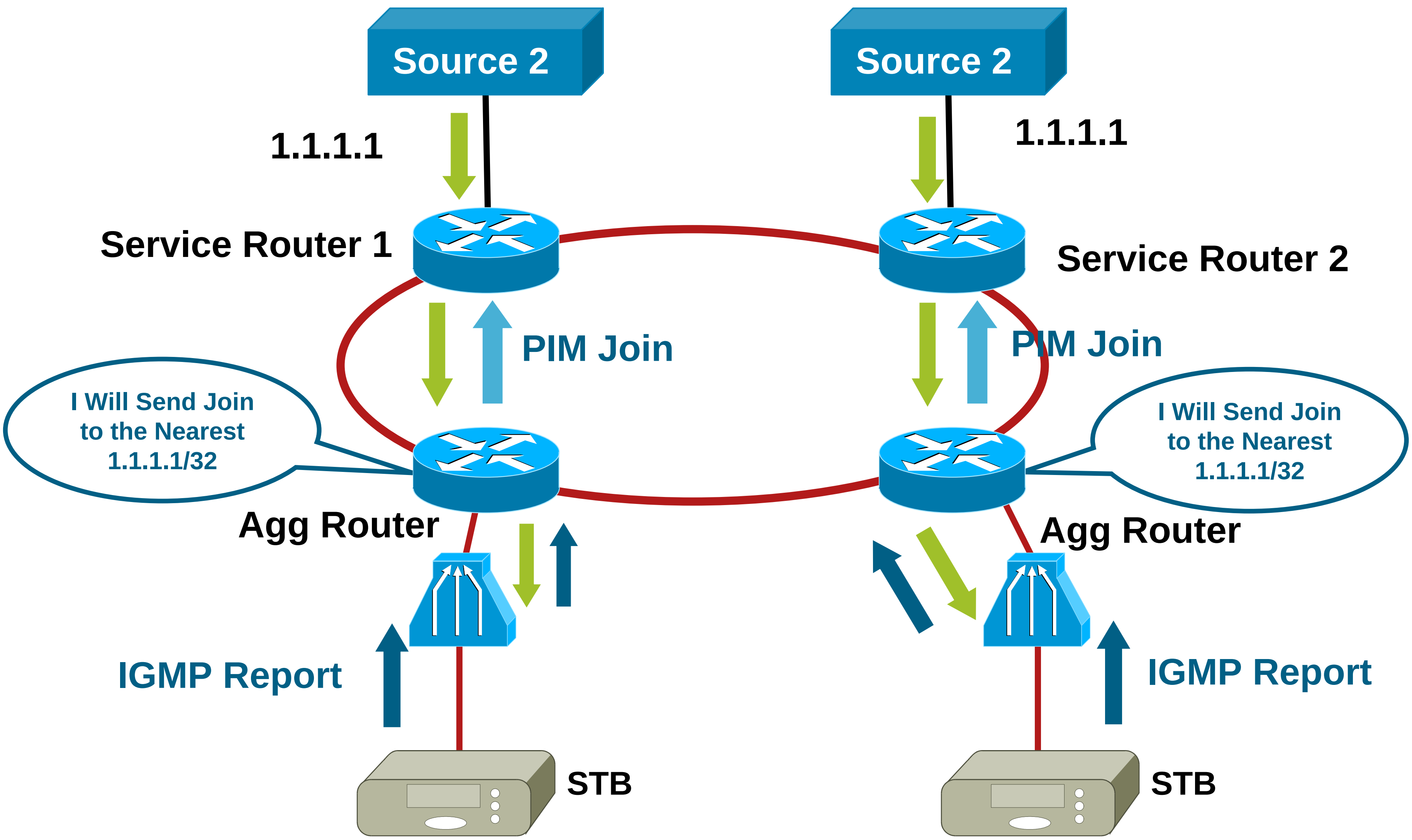
Primary Backup	Live-Live/Hot-Hot
▪ Two sources: one is active and src'ing content, second is in standby mode (not src'ing content) ▪ Heartbeat mechanism used to communicate with each other	▪ Two sources, **both** are active and src'ing multicast into the network ▪ No protocol between the two sources
▪ Only one copy is on the network at any instant ▪ Single multicast tree is built per the unicast routing table	▪ Two copies of the multicast packets will be in the network at any instant ▪ Two multicast trees on almost redundant infrastructure
▪ Uses required bandwidth	▪ Uses 2X network bandwidth
▪ Receiver's functionality simpler: Aware of only one src, failover logic handled between sources	▪ Receiver is smarter: Is aware/configured with two feeds (s1,g1), (s2,g2) / (*,g1), (*,g2) Joins both and receives both feeds
▪ This approach requires the network to have fast IGP and PIM convergence	▪ This approach does not require fast IGP and PIM convergence

Source Redundancy: Anycast/Prioritycast Signaling

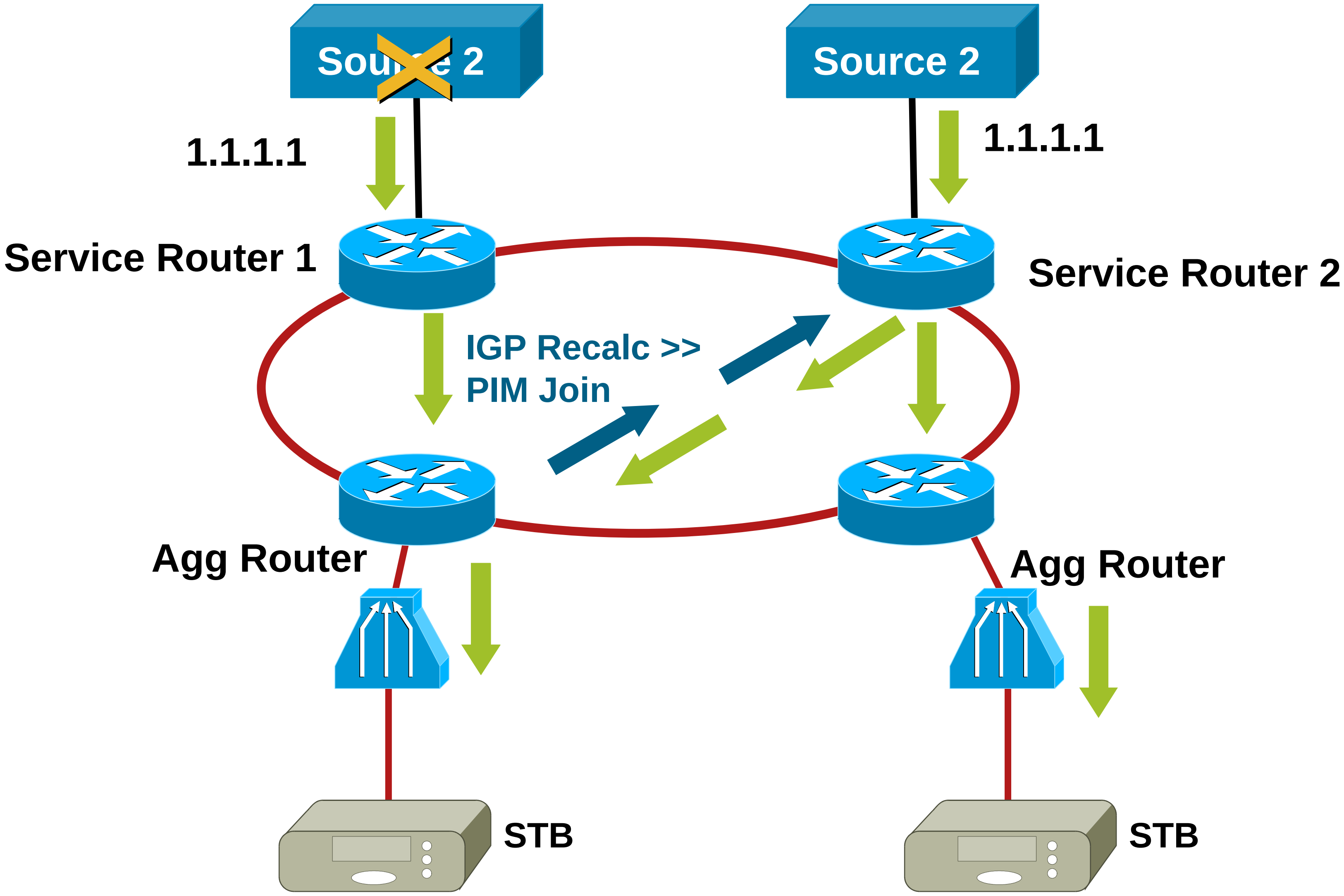
- Redundant sources or NMS announce Source Address via RIPv2
- Per stream source announcement
- Routers redistribute (with policy) into IGP
  - Easily done from IP/TV middleware (UDP)
  - No protocol machinery required—only periodic announce packets
  - Small periodicity for fast failure detection
  - All routers support RIPv2 (not deployed as IGP):
    - Allows secure constrained configuration on routers



Anycast-Based Load Balancing



Encoder Failover Using Anycast



Source Redundancy

Anycast/Prioritycast Policies

- Policies

    Anycast: Clients connect to the closest instance of redundant IP address

    Prioritycast: Clients connect to the highest-priority instance of the redundant IP address

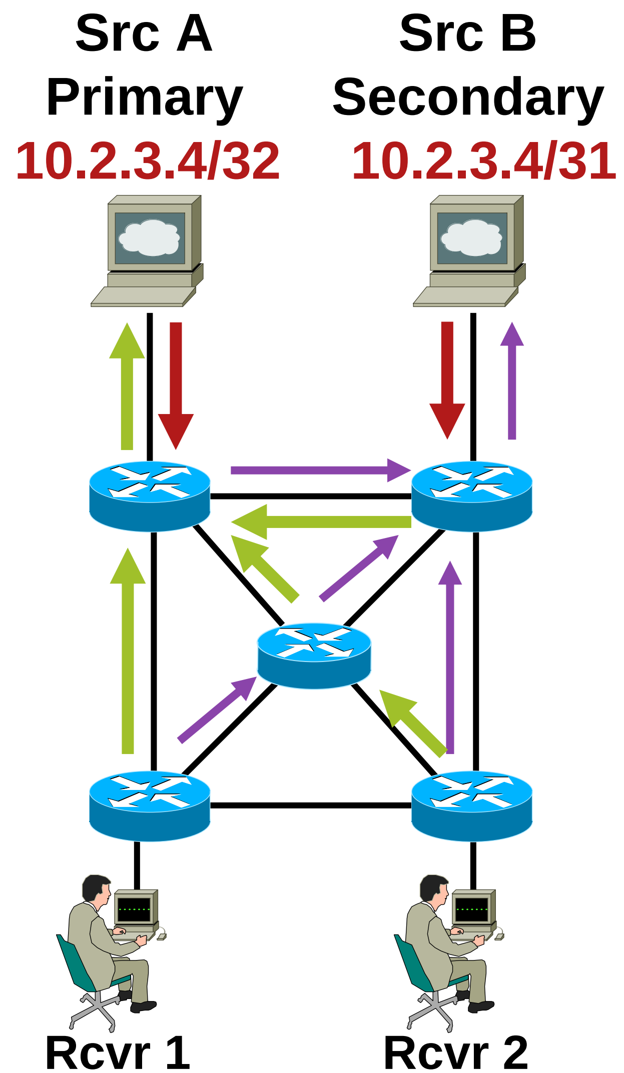
- Also used in other places

    e.g. PIM-SM and Bidir-PIM RP redundancy

- Policy simply determined by routing announcement and routing config

    Anycast well understood

    Prioritycast: Engineer metrics of announcements or use different prefix length



Example: Prioritycast with Prefixlength Announcement

Source Redundancy

Anycast/Prioritycast Benefits

- Sub-second failover possible
- Represent program channel as single (S,G)
  - SSM: single tree, no signaling; ASM: no RPT/SPT
- Move instances “freely” around the network
  - Most simply within IGP area
  - Regional to national encoder failover (BGP…)?
- No vendor proprietary source sync proto required
- Per program, not only per-source-device failover
  - Use different source address per program

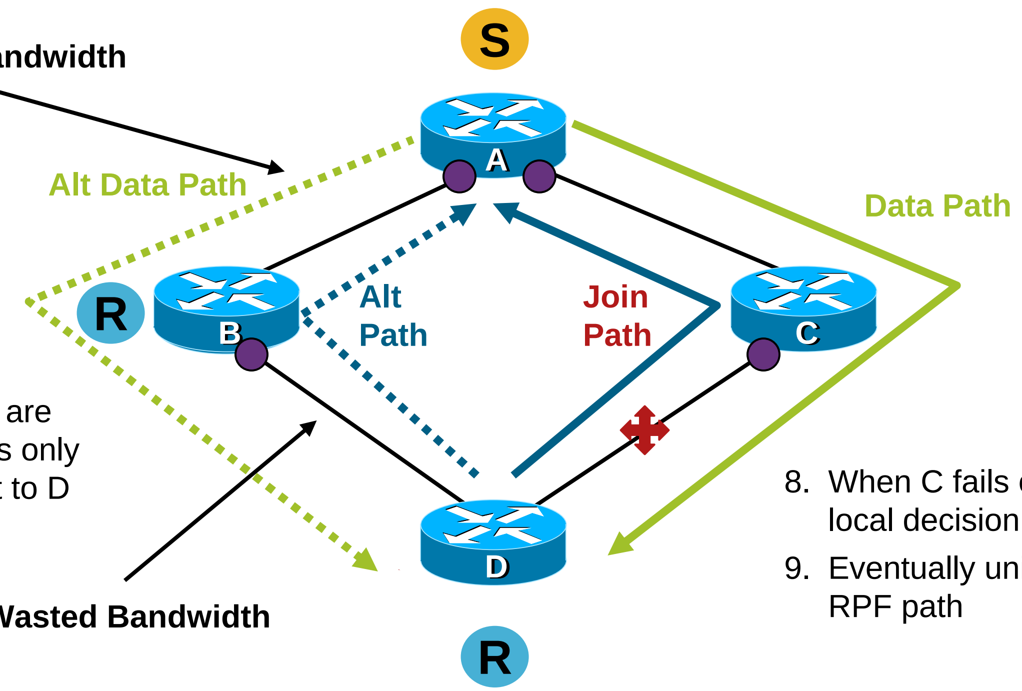
FRR for Native IP Multicast/mLDP

- Do not require RSVP-TE for general purpose multicast deployments
- Sub 50 msec FRR possible to implement for PIM or mLDP
  - Make-before-break during convergence
  - Use of link-protection tunnels
    - Initial: one-hop RSVP-TE P2P tunnels
    - Future: NotVia IPFRR tunnels (no TE needed then)

MoFRR

- It is make-before-break solution
- Multicast routing doesn’t have to wait for unicast routing to converge
- An alternative to source redundancy, but:
  - Don’t have to provision sources
  - Don’t have to sync data streams
  - No duplicate data to multicast receivers
- No repair tunnels
- No new setup protocols
- No forwarding/hardware changes

Concept Example



7. If upstream of D there are receivers, bandwidth is only wasted from that point to D

8. When C fails or DC link fails, D makes local decision to accept packets from B
9. Eventually unicast routing says B is new RPF path

1. D has ECMP path {BA, CA} to S
2. D sends join on RPF path through C
3. D can send alternate-join on BA path
4. A has 2 oifs leading to a single receiver
5. When RPF path is up, duplicates come to D
6. But D RPF fails on packets from B

Data Path
rpf Path (RPF Join)
Alt Join (Sent on Non-rpf)
Interface in oif-list
Link Down or RPF-Failed Packet Drop

Multicast Fast Convergence

- IP multicast

  All failures/topology changes are corrected by re-converging the trees

  Re-convergence time is sum of:

    Failure detection time (only for failure cases)

    Unicast routing re-convergence time

    ~ #Multicast-trees PIM re-convergence time

  Possible

    ~ minimum of 200 msec initial

    ~ 500 ... 4000 trees convergence/sec (perf)

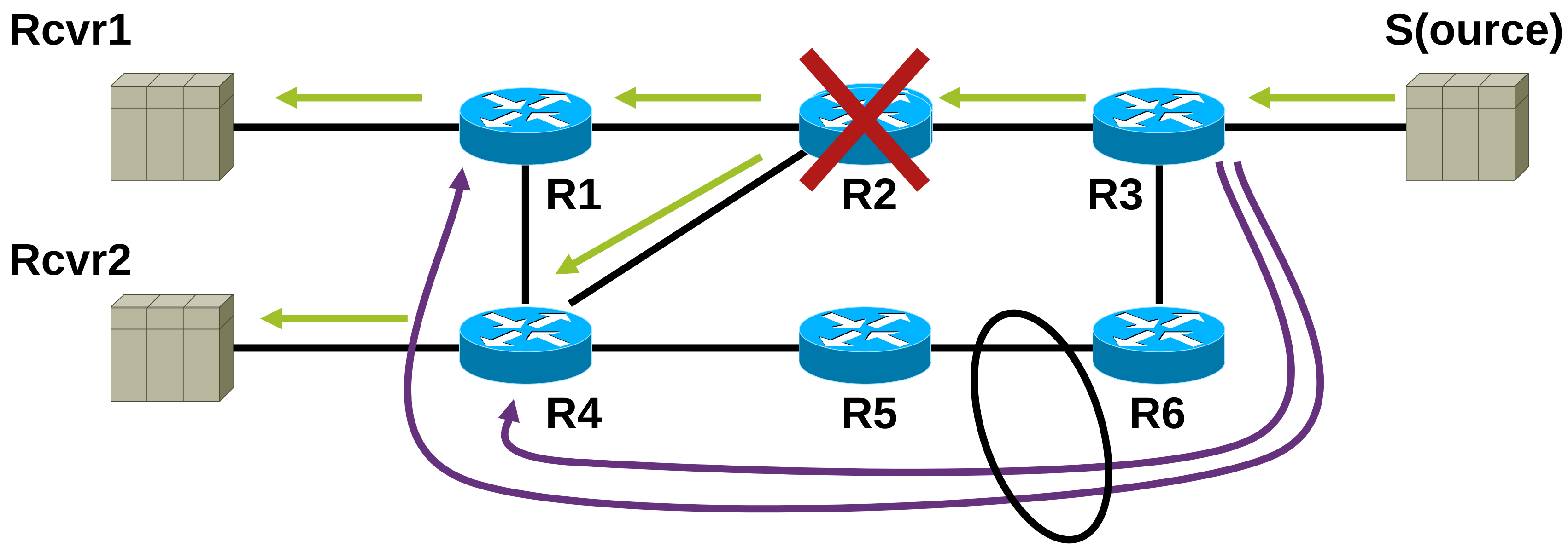
- Same behavior with mLDP

Multicast Node Protection with p2p Backup Tunnels

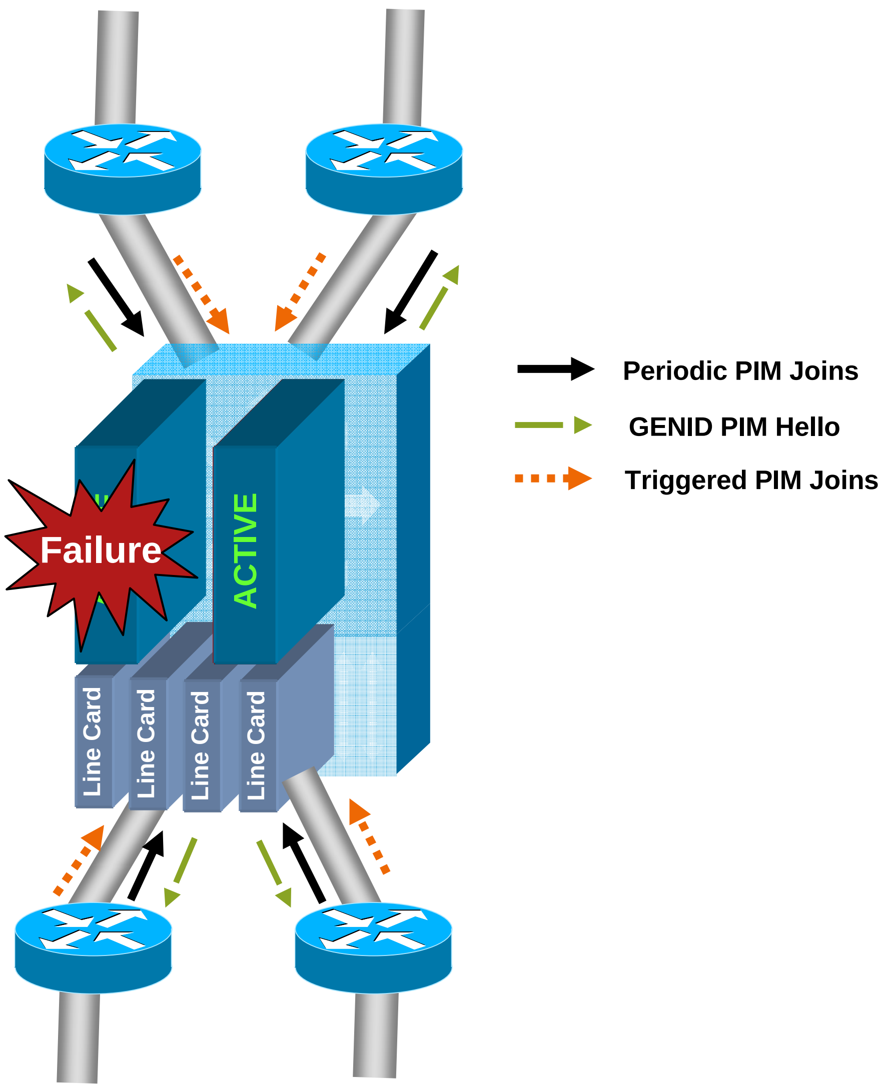
- If router with fan-out of N fails, N-times as much backup bandwidth as otherwise is needed

  Provisioning issue depending on topology!

- Some ideas to use multipoint backup to resolve this, but…
- Recommendation? Rely on Node HA instead!!



Multicast HA for SSM: Triggered PIM Join(s)



How Triggered PIM Join(s) Work When Active Route Processor Fails:

- Active Route Processor receives periodic PIM Joins in steady-state
- Active Route Processor fails
- Standby Route Processor takes over
- PIM Hello with GENID is sent out
- Triggers adjacent PIM neighbors to resend PIM Joins refreshing state of distribution tree(s) preventing them from timing out

Multi-Topology (MT)-Technology and IP Multicast

- … When not all traffic should flow on the same paths
- Interdomain: Incongruent routing
  - BGP SAFI2 (MBGP)
- Intradomain: Incongruent routing workarounds
  - Static mroutes
  - Multiple IGP processes (tricky)
- Intradomain: Multi-Topology-Routing
  - Multicast and Unicast solution; multiple topologies for unicast and multicast
- Intradomain: MT-technology for multicast
  - Subset of MTR: Only the routing component, sufficient for incongruent routing for IP multicast

MBGP Overview

MBGP: Multiprotocol BGP

- Defined in RFC-2283 (extensions to BGP)
- Can carry different types of routes
  - IPv4/v6 Unicast/Multicast
- May be carried in same BGP session
- Does not propagate multicast state information
  - Still need PIM to build Distribution Trees
- Same path selection and validation rules
  - AS-Path, LocalPref, MED, …

MBGP Update Message

- Address Family Information (AFI)
  - Identifies Address Type (see RFC-1700)
    - AFI = 1 (IPv4)
    - AFI = 2 (IPv6)
- Sub-Address Family Information (Sub-AFI)
  - Sub-category for AFI Field
  - Address Family Information (AFI) = 1 (IPv4)
    - Sub-AFI = 1 (NLRI is used for unicast)
    - Sub-AFI = 2 (NLRI is used for multicast RPF check)

MBGP: NLRI Information

BGP Update from Peer

MP_REACH_NLRI: 192.192.2/24
AFI: 1, Sub-AFI: 1 (unicast)
AS_PATH: 300 200
MED:
Next-Hop: 192.168.200.2

Unicast BGP Table

Network	Next-Hop	Path
*>i160.10.1.0/24	192.20.2.2	i
*>i160.10.3.0/24	192.20.2.2	i
*>i192.192.2.0/24	192.168.200.2	300 200 i

Multicast BGP Table

Network	Next-Hop	Path
*>i160.10.1.0/24	192.20.2.2	i
*>i160.10.3.0/24	192.20.2.2	i

Storage of arriving NLRI information depends on AFI/SAFI fields in the Update message

- **Unicast BGP Table only (AFI=1/SAFI=1 or old style NLRI)**

MBGP: NLRI Information

BGP Update from Peer

MP_REACH_NLRI: 192.192.2/24
AFI: 1, Sub-AFI: 2 (multicast)
AS_PATH: 300 200
MED:
Next-Hop: 192.168.200.2

Unicast BGP Table

Network	Next-Hop	Path
*>i160.10.1.0/24	192.20.2.2	i
*>i160.10.3.0/24	192.20.2.2	i

Multicast BGP Table

Network	Next-Hop	Path
*>i160.10.1.0/24	192.20.2.2	i
*>i160.10.3.0/24	192.20.2.2	i
*>i192.192.2.0/24	192.168.200.2	300 200 i

Storage of arriving NLRI information depends on AFI/SAFI fields in the Update message

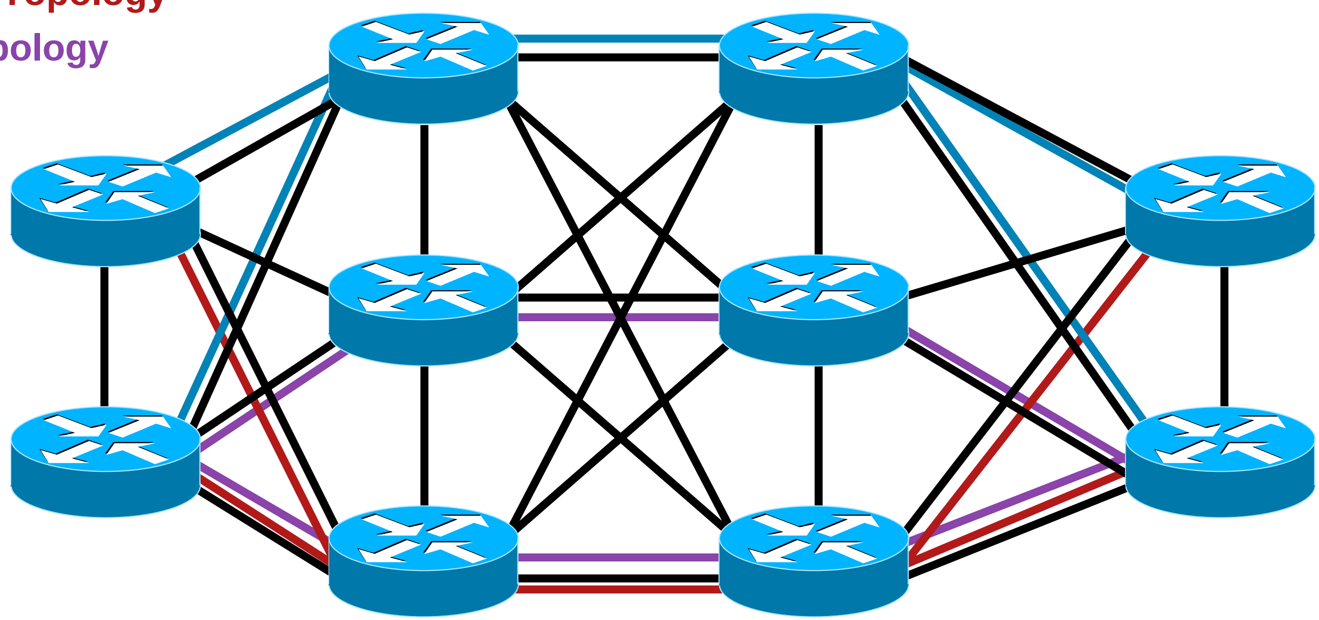
- Unicast BGP Table only (AFI=1/SAFI=1 or old style NLRI)
- **Multicast BGP Table only (AFI=1/SAFI=2)**

Multi-Topology Routing (MTR)

Full Solution with Both MT-Technology Routing and Forwarding

Base Topology
Voice Topology
Multicast Topology
Video Topology

Start with a Base Topology Includes All Routers and All Links



- Define traffic-class specific topologies across a contiguous subsection of the network
- Individual links can belong to multiple topologies

Applications for Multiple Topologies for IP Multicast

- Original MTR Reasons
  - Different QoS through choice of different paths:
    - Well applicable to multicast:
      - Low-latency and low-loss: hoot&holler/IPICs multicast
      - Low-latency: finance market-data (stream redundancy against loss)
      - High-bandwidth: ACNS content provisioning network
      - Low-loss: video
    - Not too critical:
      - Most networks today only run one type of business critical multicast apps (about to change?!)
- Live-Live with Path Diversity
  - Also called stream redundancy with path separation
  - Examples shown in various stages of deployment with other approaches or workarounds to multi-topology multicast
    - But multicast with multiple topology considered most easy/flexible approach to problem

Live-Live

- Live-Live—Spatial Separation

  Two separate paths through network; can engineer manually (or with RSVP-TE P2MP)

  Use of two topologies (MTR)

  “Naturally” diverse/split networks work well (SP cores, likely access networks too), especially with ECMP

  Target to provide “zero loss” by merging copies based on sequence number

- Live-Live—Temporal Separation

  In application device—delay one copy—need to know maximum network outage

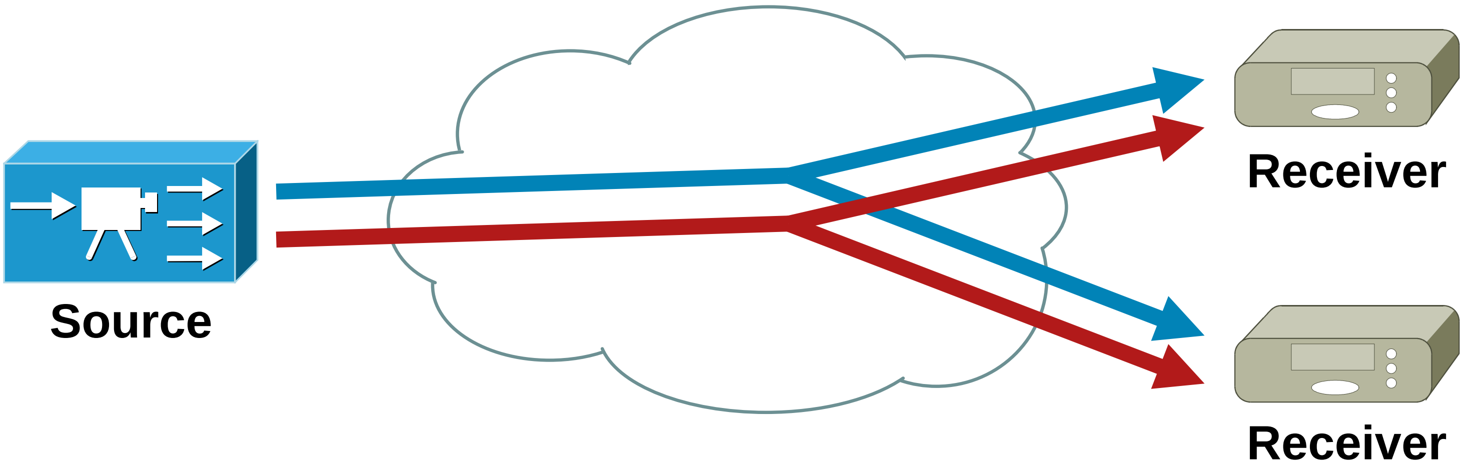
What Is Live-Live (with Path Diversity)?

- Transport same traffic twice across the network…

  Receivers can merge traffic by sequence-number

- … On diverse paths to achieve the Live-Live promise:

  Every single failure in the network will only affect one copy of the traffic



What Is Live-live (with Path Diversity)?

- Why bother?

  Only resiliency solution in the network that that can be driven to provide zero packet loss under any single failure in the network—without introducing more than path propagation delay (latency)!

- Much more interesting for multicast than unicast

  Individual unicast packet flow typically for just one receiver

  Individual multicast flow (superbowl) for N(large) receivers!

- Path diversity in the network

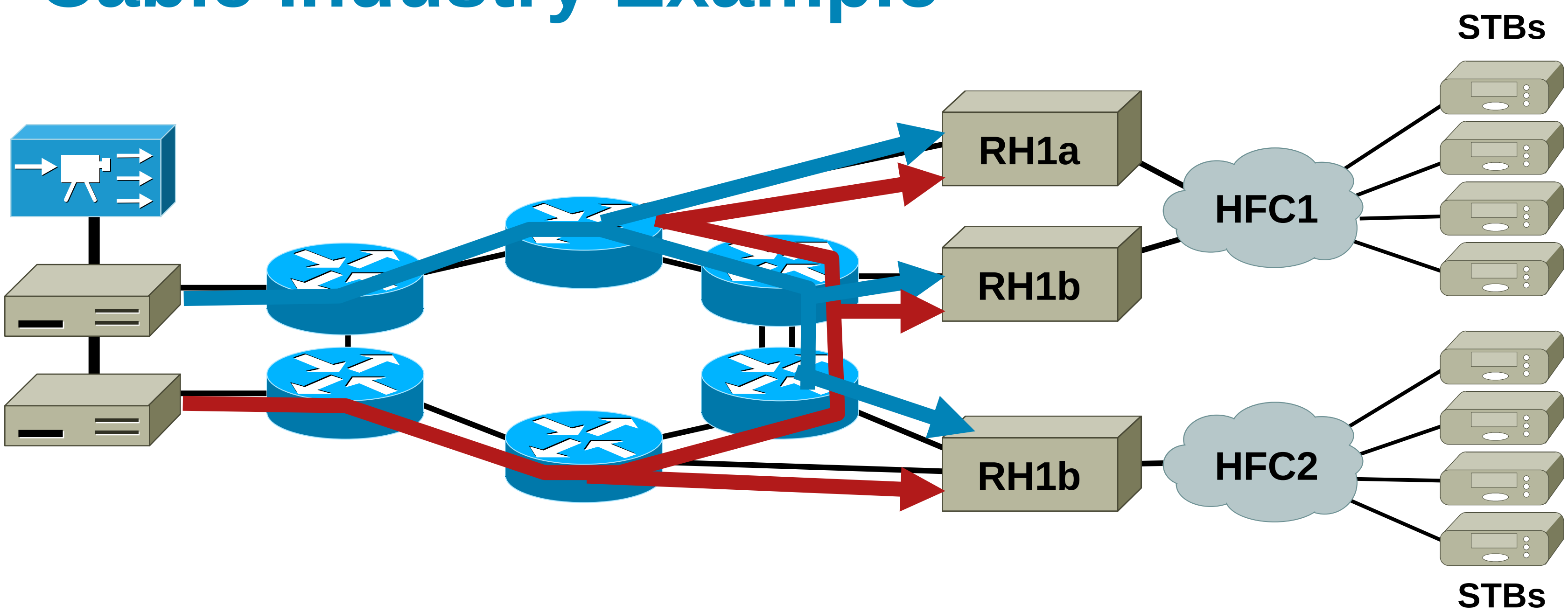
  Lots of alternatives: VRF-lite, routing tricks, RSVP-TE, L2 VLAN

  Multi-topology routing considered most simple/flexible approach!

- Standard solution in finance market data networks

  Legacy: Path diversity through use of two networks!

Cable Industry Example



- Path separation does not necessarily mean separate parts of network!

  Carrying copies counterclockwise in rings allows single ring redundancy to provide live-live guarantee; less expensive network

- Target in cable industry (previously used non-IP SONET rings!)

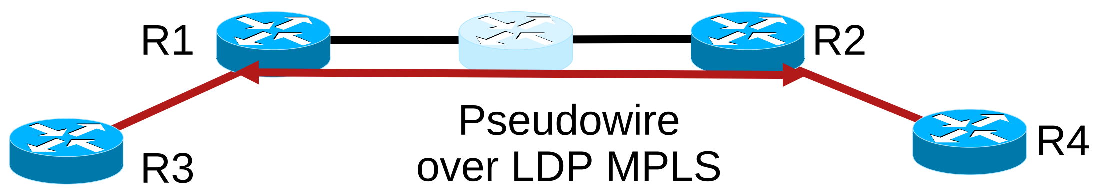
  IP live-live not necessarily end-to-end (STB), but towards Edge-QAM (RH*)—merging traffic for non-IP delivery over digital cable

  With path separation in IP network and per-packet merge in those devices solution can target zero packet loss instead of just sub 50msec

Protected pseudowires

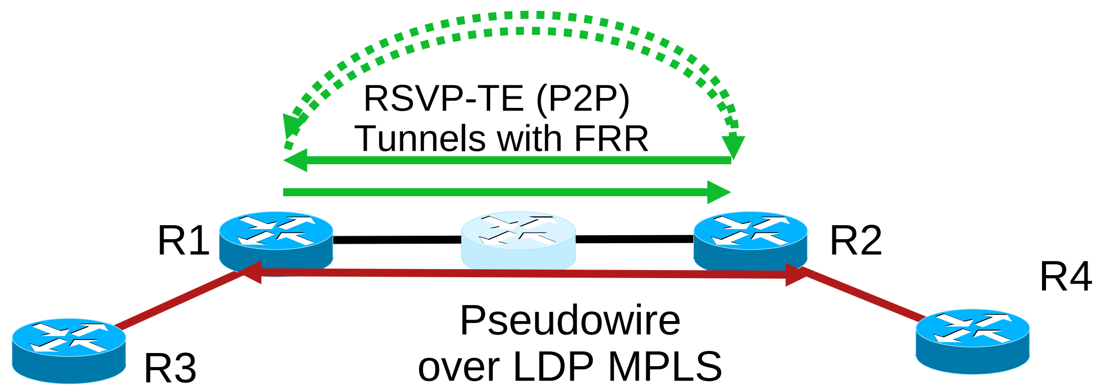
Classic pseudowire

- R1/R2 provide pseudowire for R3/R4 accepting/delivering packets from/to physical interface.



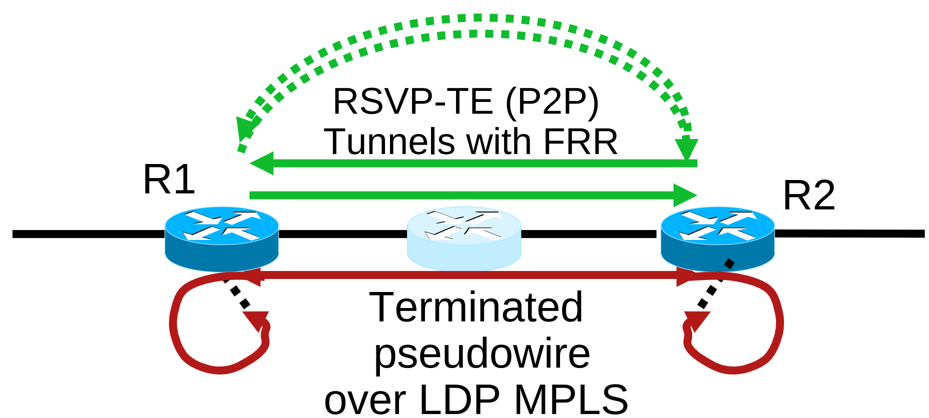
Protected pseudowire

- Provide sub 50msec link protection for packets of pseudowire (or any other MPLS packets) by configuring RSVP-TE LSP with FRR backup tunnel

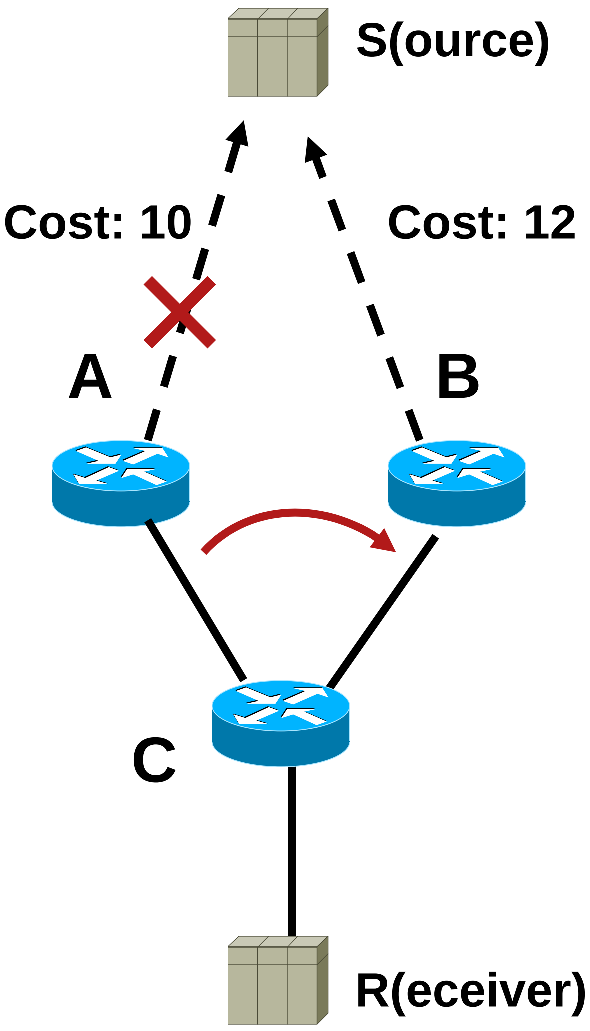


Terminated pseudowire

- R1/R2 terminate pseudowire on internal port instead of physical interface. Can bridge (VLAN) or route from/to port.



cFRR
PIM/mLDP Break before Make



RPF change on C from A to C:

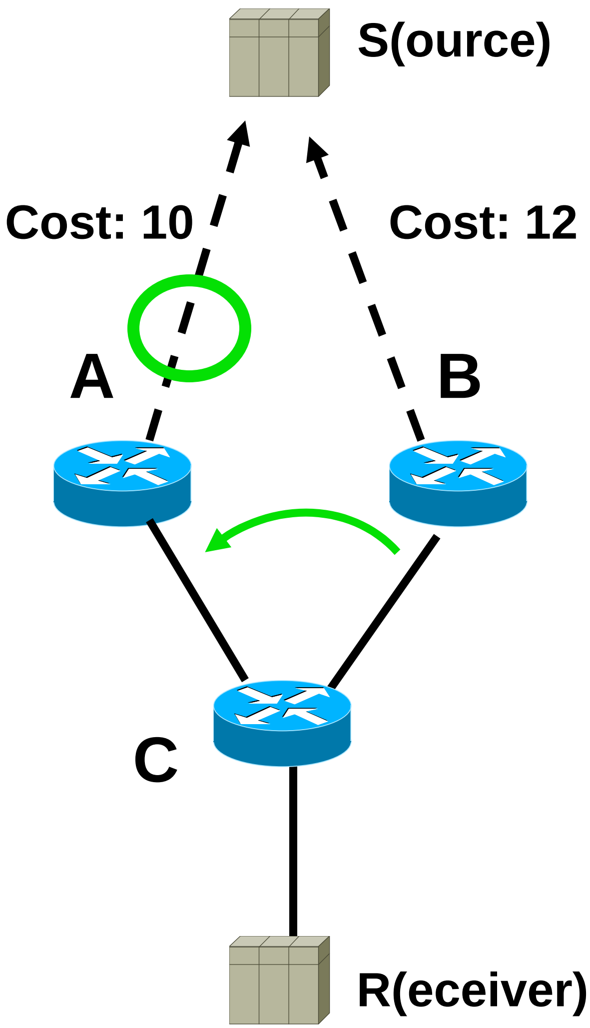
1. Receive RPF change from IGP
2. Send prunes to A
3. Change RPF to B
4. Send joins to B

Same methodology, different terminology in mLDP

RPF == ingres label binding

Some more details (not discussed)

cFRR
PIM/mLDP Make before Break

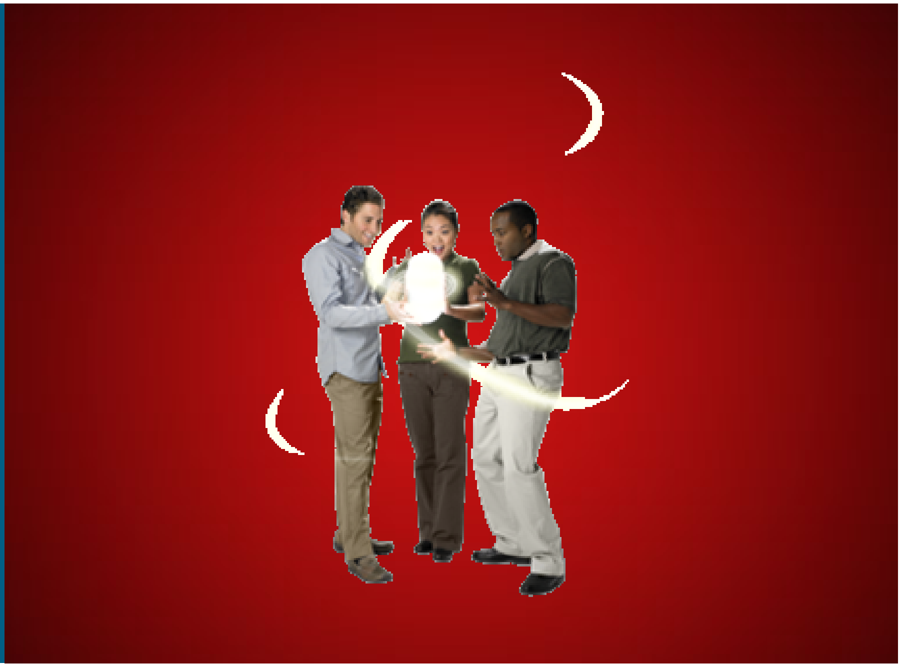


1. Receive RPF change from unicast
2. Send joins to A
3. **Wait for right time to go to 4.**
   Until upstream is forwarding traffic
4. Change RPF to A
5. Send prunes to B

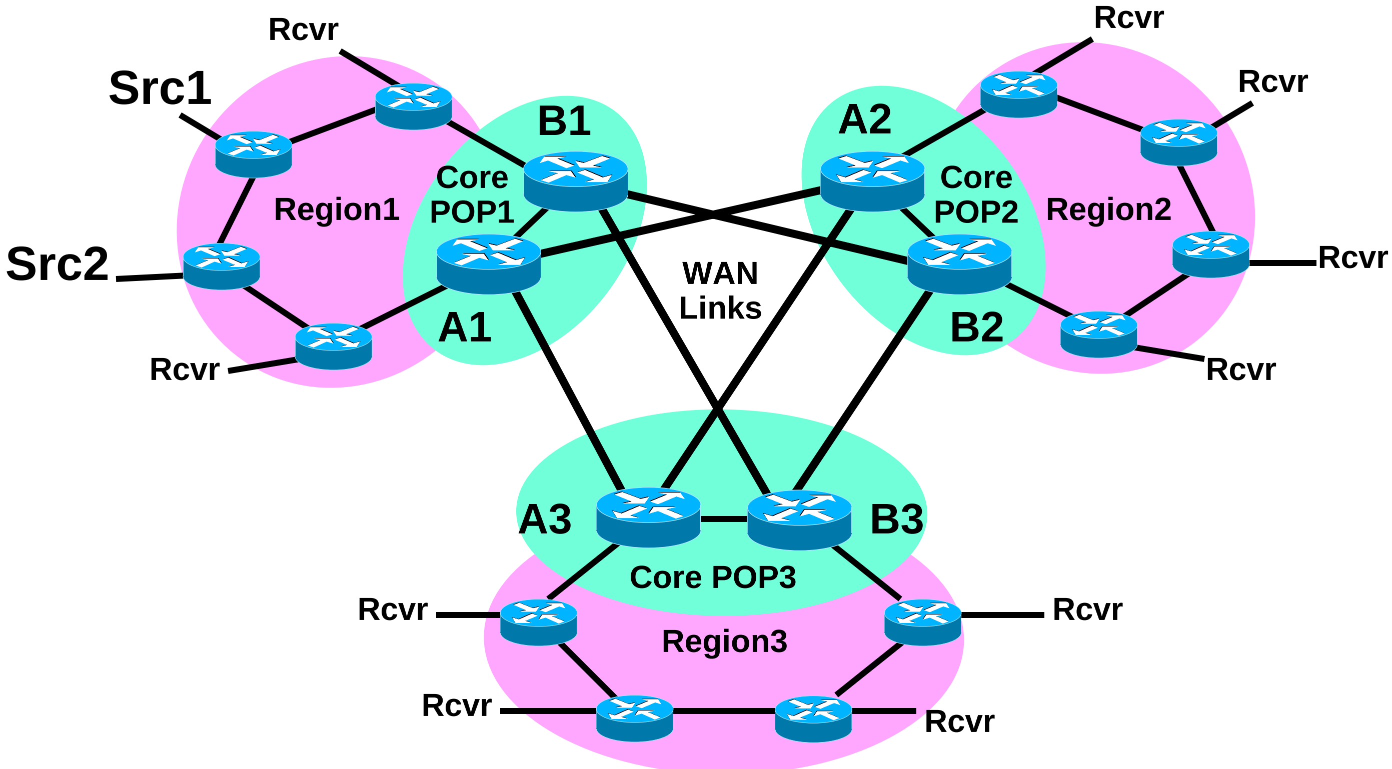
Should only do Make-before-Break when old path (B) is known to still forward traffic after 1.

- Path via B failed but protected
- Path to A better, recovered
- Not: path via B fails, unprotected
  - Make before Break could cause more interruption than Break before Make !

Path selection

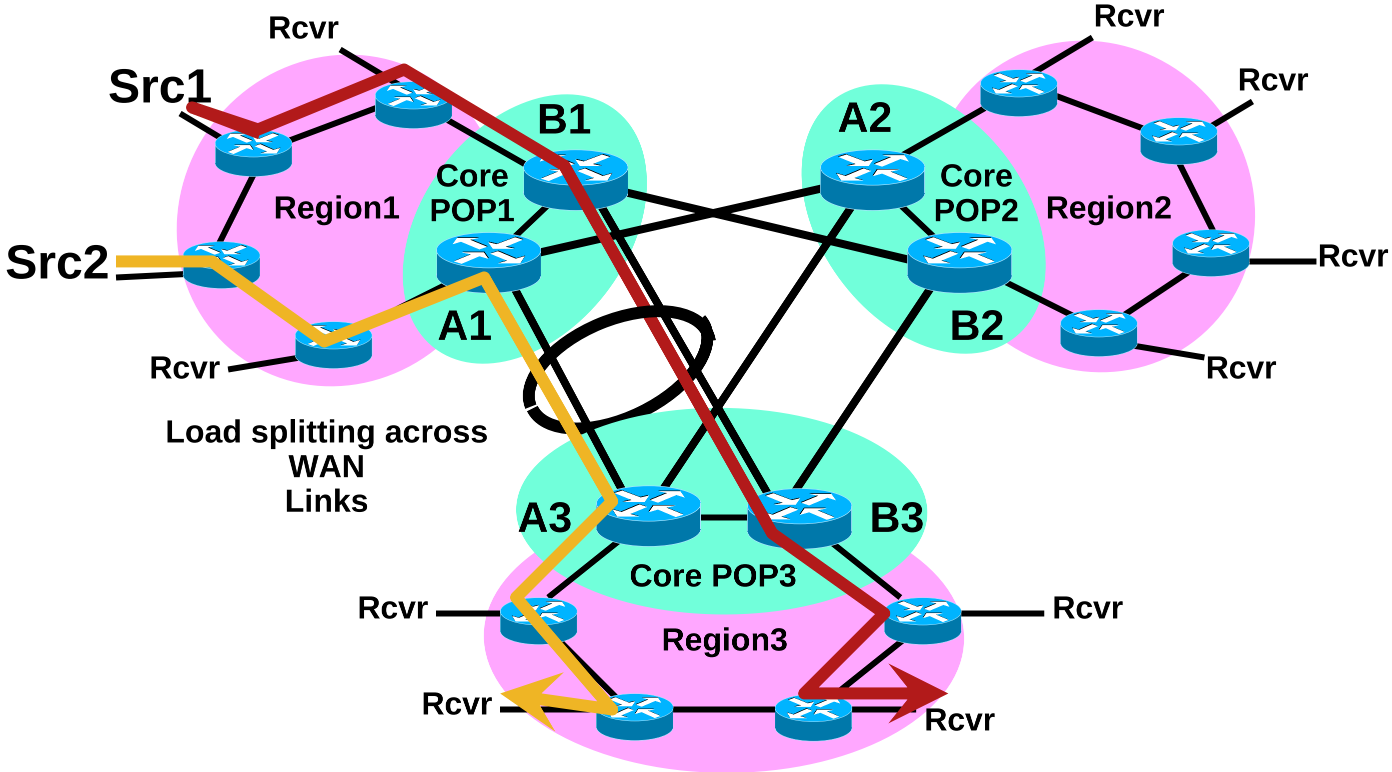


MT-IGP
Cost optimization



- Consider simplified example core/distribution network toplogy
- Core pops have redundant core routers, connectivity via (10Gbps) WAN links, redundant. Simple setup: A/B core routers, A/B links
- Regions use ring(s) for redundant connectivity

MT-IGP
Cost optimization

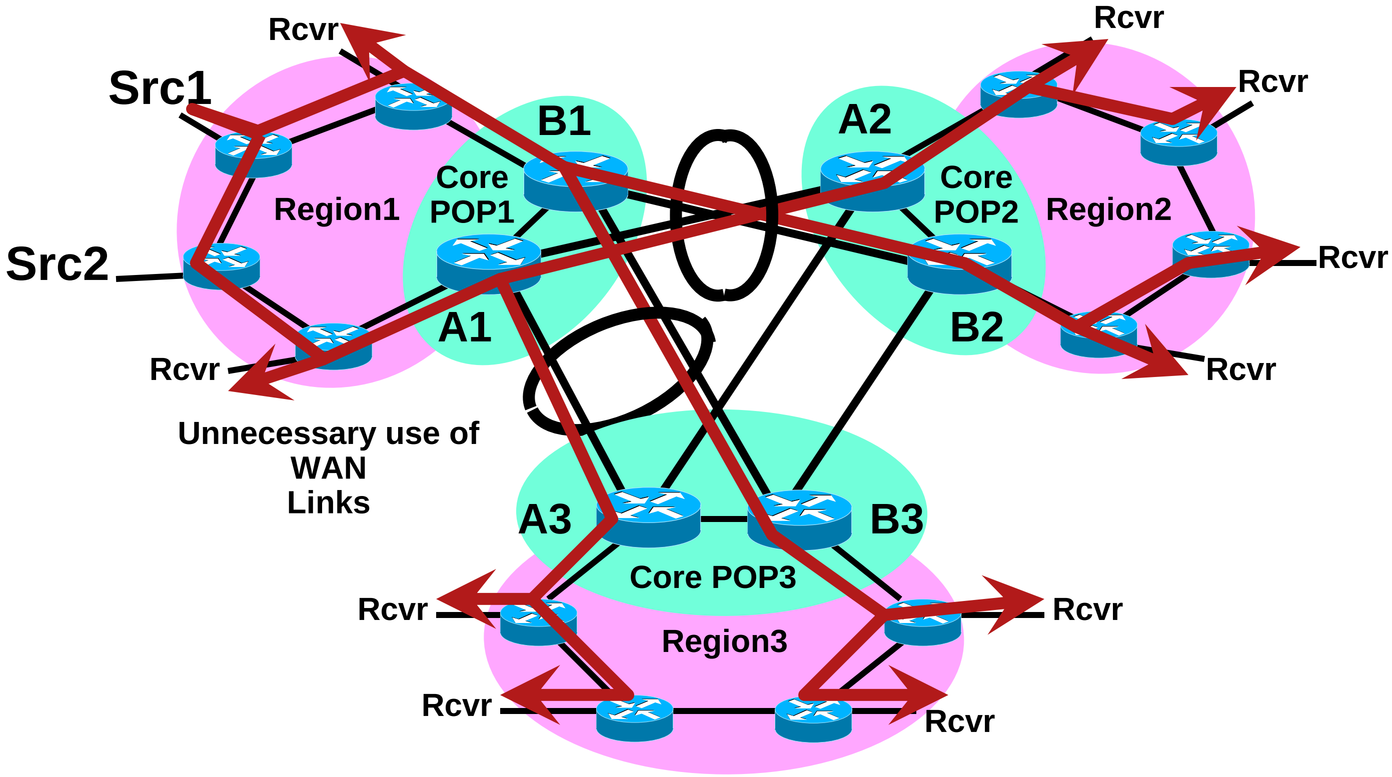


- IGP metric for load distribution across redundant core.

  Manual IGP metric setting and/or tools (Cariden)

- Result: Unicast traffic is load split across redundant core links

MT-IGP
Cost optimization

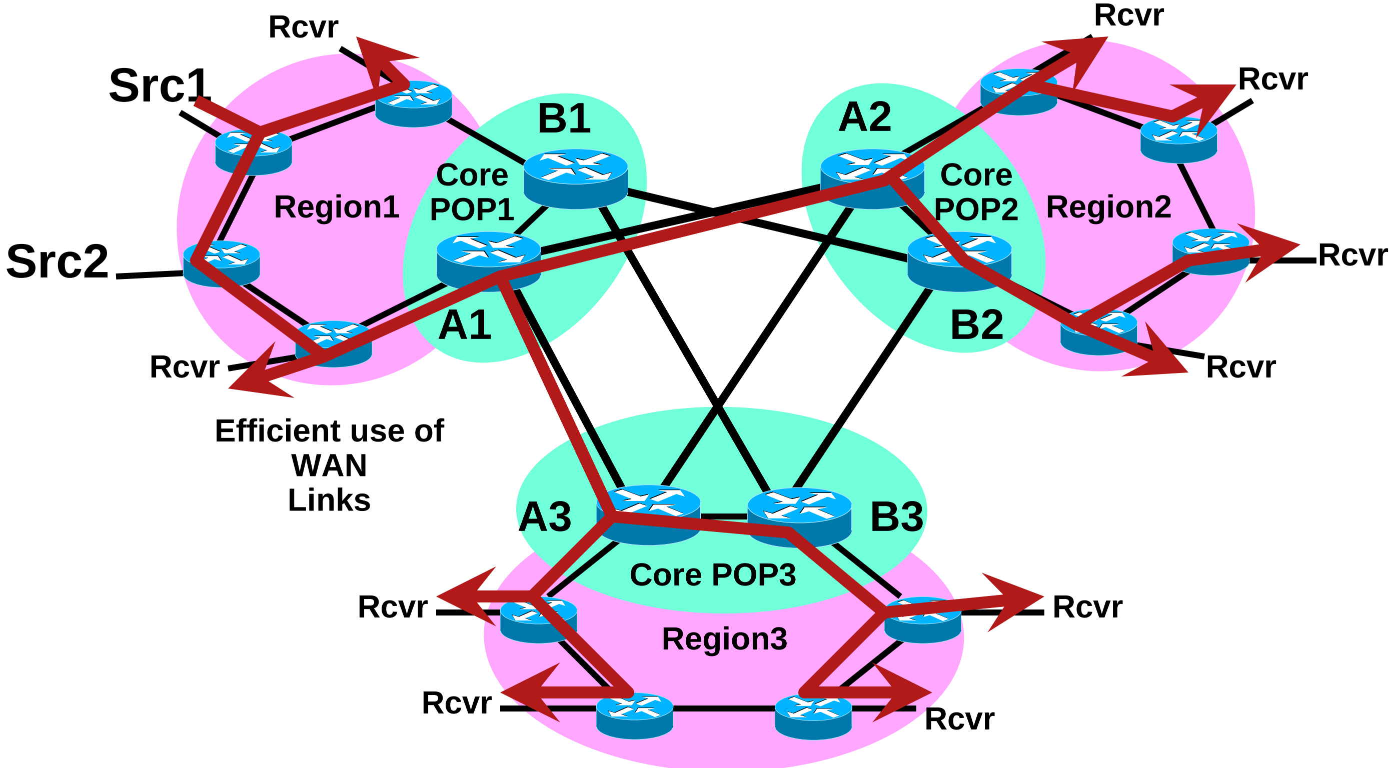


- The same metric good for unicast load splitting cause multicast traffic to go unnecessarily across both the A and B WAN links.

  10 Gbps WAN links, 1..2 Gbs multicast => 10..20% WAN waste (cost factor)

- Can not resolve problem well without multicast specific topology

MT-IGP
Cost optimization



- Simple? to minimize tree costs with a multicast specific topology

  Manual or tool based

  Example toplogy: make B links very expensive for multicast (cost 100), so they are only us as last resort (loss of A connectivity)

IP multicast (and mLDP) ECMP
non-polarizing means non-predictable

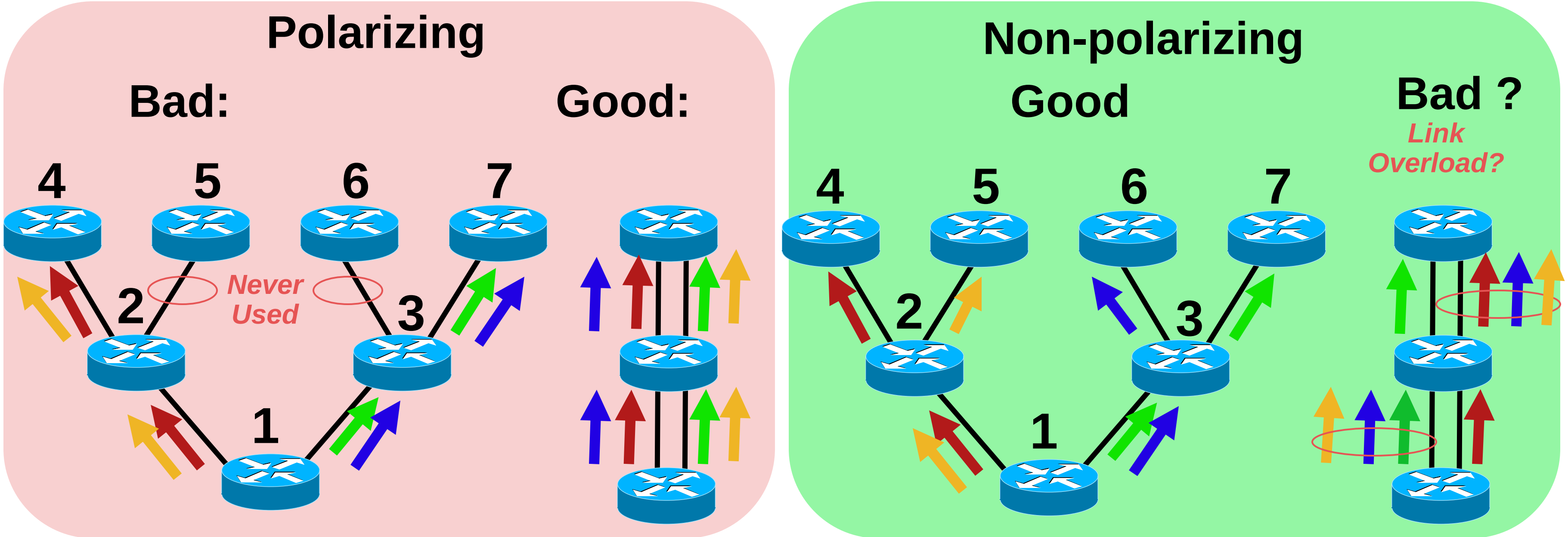
- **Polarization:**

  All routers along network path choose same relative interface for a multicast tree.

- **Predictability**:

  With algorithm known, group addresses G of (S,G) can be assigned by operator such that traffic is well split across multiple hops (link bundles)

  Workaround, not recommended – for highly utilized links (> 85% ?)



IP multicast (and mLDP) ECMP
stability, consistency

- Multicast ECMP different from unicast:

    Unicast ECMP non-polarizing, but also non-stable, non-consistent.

    Not a problem for unicast, but multicast:

- Stability

    If path fails only trees on that path will need to reconverge. If path recovers, only trees that will use the new path will reconverge

    Polarizing multicast algoritm is NOT stable!

- Consistency

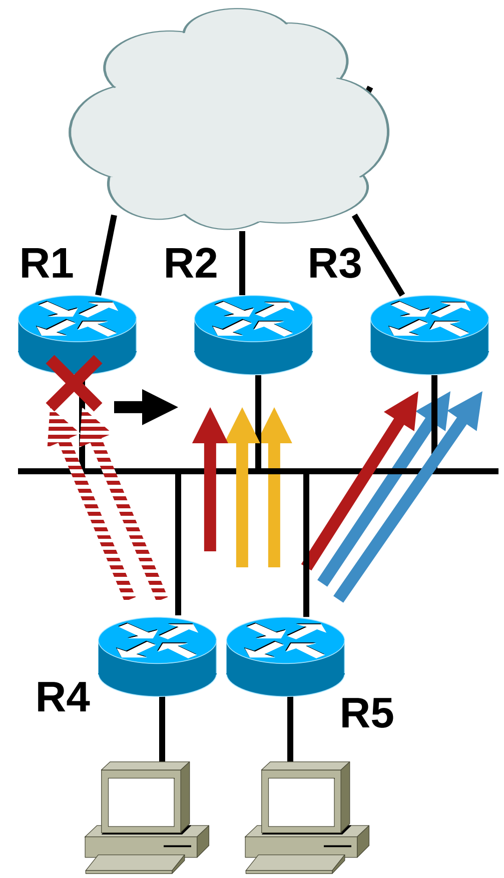
    Multiple downstream router on same LAN (R4, R5) will select same upstream router.

    Avoids “assert” problem in PIM-SSM

    Polarizing multicast ECMP also consistent.

- mLDP targeting same algorithms

    No Assert problems though…



Path selection review
RSVP-TE/P2MP

- CSPF/ERO "Traffic Engineering"
  (bandwidth, priority and affinity based path selection)
- Very powerful "can do everything we can think of"
- "Offline" management (ERO) most common

  Network provider incooperates "off-network" information about necessary multipoint trees

- "Online" / CSPF based path selection

  Ideal for single headends.

  How much better than SPF without coordinated CSPF for multiple headends ?

  Network wide coordinated CSPF calculation TBD

Path selection review
PIM (native multicast) / mLDP

- Can not load split across non-equal-cost paths
- Path engineering with topologies and ECMP:
- ECMP

  best when multipoint traffic << link bandwidth (30%?)

  Higher utilization deployments – special considerations
  (due to statistical chance of congestion)

- Topologies

  Single incongruent topology – cost opt / route around obstacles.

  Two topologies for path separation (live-live)

  Could use more topologies for more functionality – eg: non-equal-cost load-splitting – but maintaining many topologies likely not less complex than RSVP-TE

  Note: MT-technology for multicast
  only happens in control plane. No forwarding plane impact

Admission control

Static vs. dynamic trees

1. "Broadcast Video"

    Dynamic IGMP forward up to DSLAM

        DSL link can only carry required program!

    static forwarding into DSLAM

    Fear of join latency

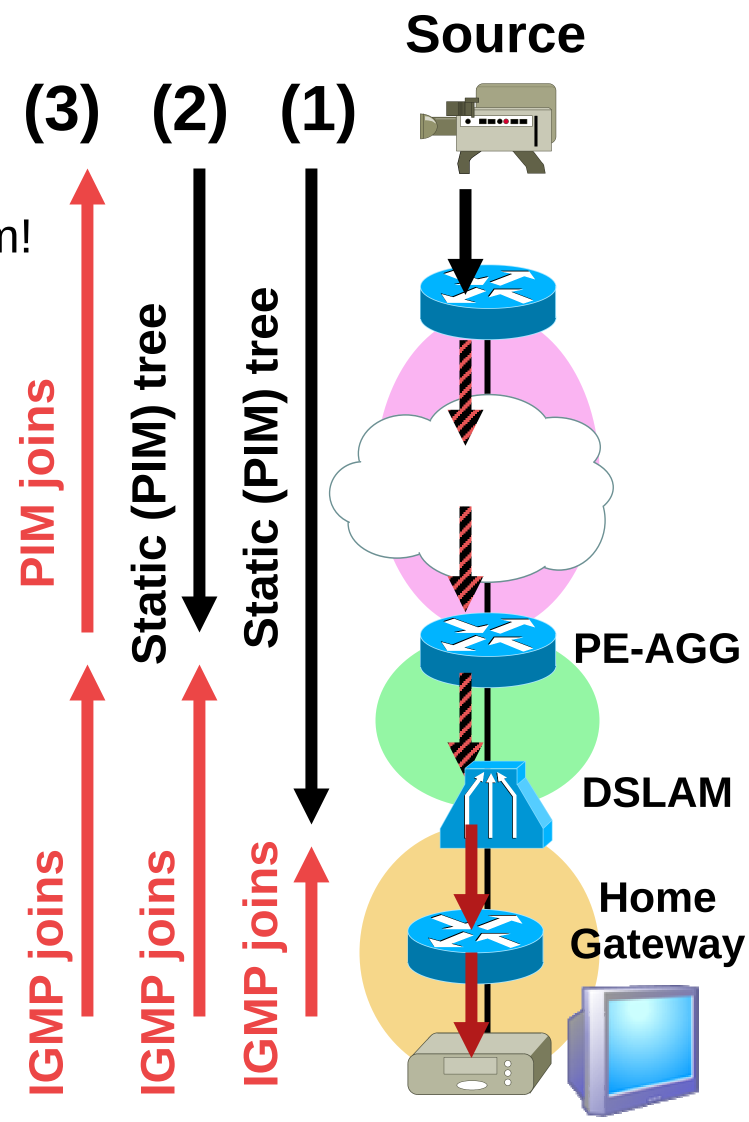
    History (ATM-DSLAM)

2. "Switched Digital Video"

    Allow oversubscription of PE-AGG/DSLAM link

3. "Real Multicast"

    dynamic tree building full path



Switched Digital Video

Why oversubscription of access links makes sense

- **Switched Digital Video**
  - Consider 500…1000 users on DSLAM
  - Consider 300 available TV programs
  - Monitor customer behavior – what is being watched ?
    - Example (derived from actual MSO measurements)
    - Some 50 TV programs almost always watched (big channels)
    - Out of remaining 220 TV programs never than ¼ watched
    - Never need more bandwidth than ~ 125 channels!
- **Dynamic joining towards core ?**
  - Todays offered content << #users aggregated -> worst case traffic will always flow.
  - More a provisioning issue – and when content expands well beyond current cable-TV models

Admission control

- **Congestion must be avoided**

  Inelastic: TV traffic can not throttle upon congestion

  One flow too many disturbs all flows

  Need to do per TV-flow admission control

- **Router-links**

  Router local CLI solution

  Strategic solution: RSVP

    Already used for unicast VoD

    Can only share bandwidth between unicast and multicast with RSVP

- **Broadband access (DSL link, Cable)**

  Issues with L2 equipment (eg: DSLAM)

Multicast Call Admission Control

Example CAC use:

1. Three CPs

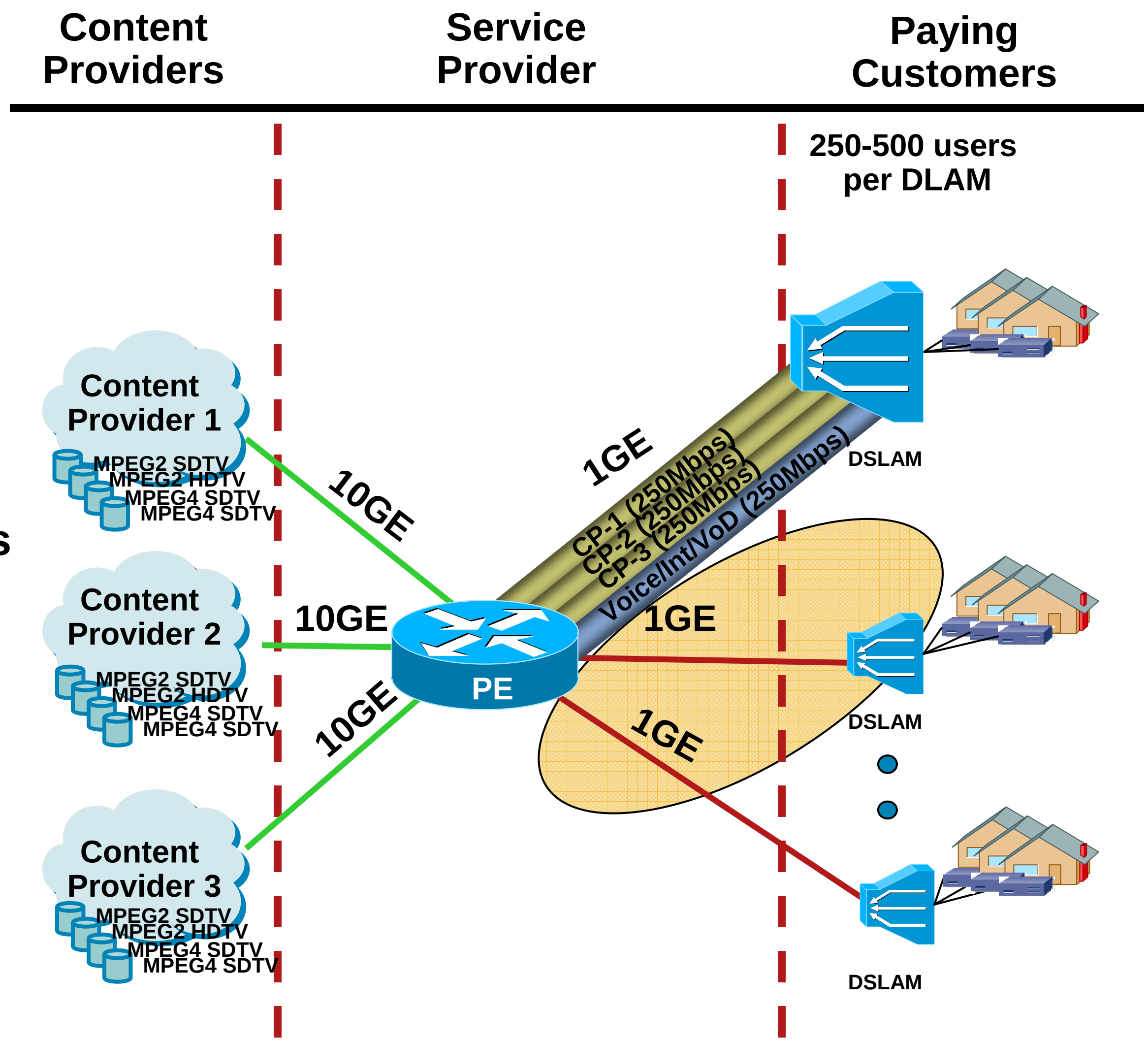
2. Different BW:

   - MPEG2 SDTV: 4 Mbps
   - MPEG2 HDTV: 18 Mbps
   - MPEG4 SDTV: 1.6 Mbps
   - MPEG4 HDTV: 6 Mbps

3. Fair sharing of bandwidth

4. 250 Mbps for each CP
   250 Mbps Internet/etc

5. Simply add global costs



Broadband link access, admission control

- **No IGMP snooping (replication) on DSLAM**

  PE-AGG access/admission control on PE-AGG link affects only single subscriber == equivalent to do access/admission control on DSL link.

  Or BRAS (if traffic not native but via PPPoE tunnel

- **IGMP snooping on DSLAM**

  PE-AGG stopping multicast traffic on PE-AGG link will affect all subscriber. Only DSLAM can control DSL link multicast traffic

- **IP Multicast extensions to ANCP (Access Node Control Protocol)**

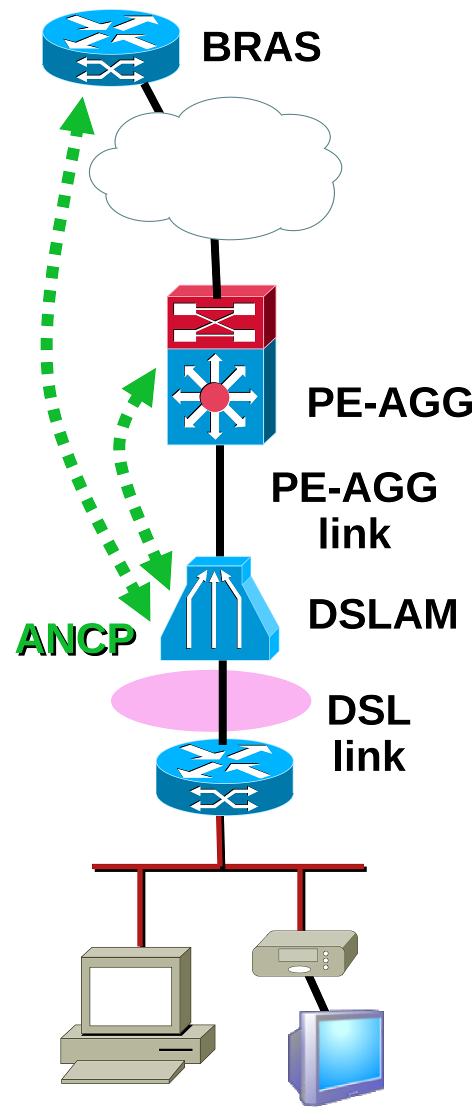
  Work in IETF

  In IGMP snooping on DSLAM, before forwarding, request authorization from ANCP server.

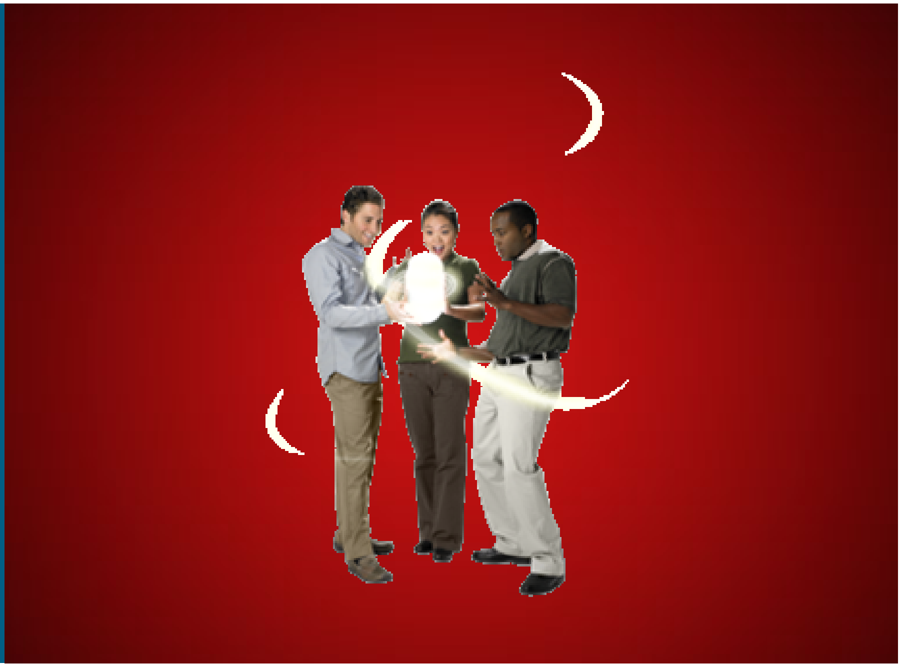
  Allow ANCP server to download access control list to DSLAM.

- **Similar model as defined in DOCSIS 3.0**

  CMTS controls CM



Channel changing



Join Latency

- Static forwarding (to PE-AGG, or DSLAM)

  To avoid join latency

  Sometimes other reasons too (policy, …)

- Bogus ?

  Hop-by-hop Join latency (PIM/IGMP) very low, eg: individual < 100 msec …

  Joins stop at first router/switch in tree that already forwards tree

  Probability for joins to go beyond PE-AGG very low !

  If you zap to a channel and it takes ¼ sec more: You are the first guy watching this channel in a vicinity of eg: 50,000 people. Are you sure you want to watch this lame program ?

- Important

  Total channel zapping performance of system – Primetime TV full hour or (often synchronized) commercial breaks.

  Join latency during bursts might be worse than on average. (DSLAM performance)

IGMPv2 leave latency
Obsolete problem

- Congesting issues due to IGMPv2 leave latency when only admission control mechanism is:

  DSL link fits only N TV programs …
  and subscriber can only have N STB.

- Example:

  4Mbps DSL link, 3.5 Mbps MPEG2

  Can only receive one TV channel at a time

  Leave latency on channel change complex (triggers IGMP queries from router/DSLAM) and long (spec default: 2 seconds)

- Resolved with IGMPv3/MLDv2

  Ability for explicit tracking (vendor specific)

  Can immediately stop forwarding upon leaves

Channel Changing
GOP size and channel changing

- GOP size of N seconds causes channel change latency *USER EXPERIENCE* >= N seconds

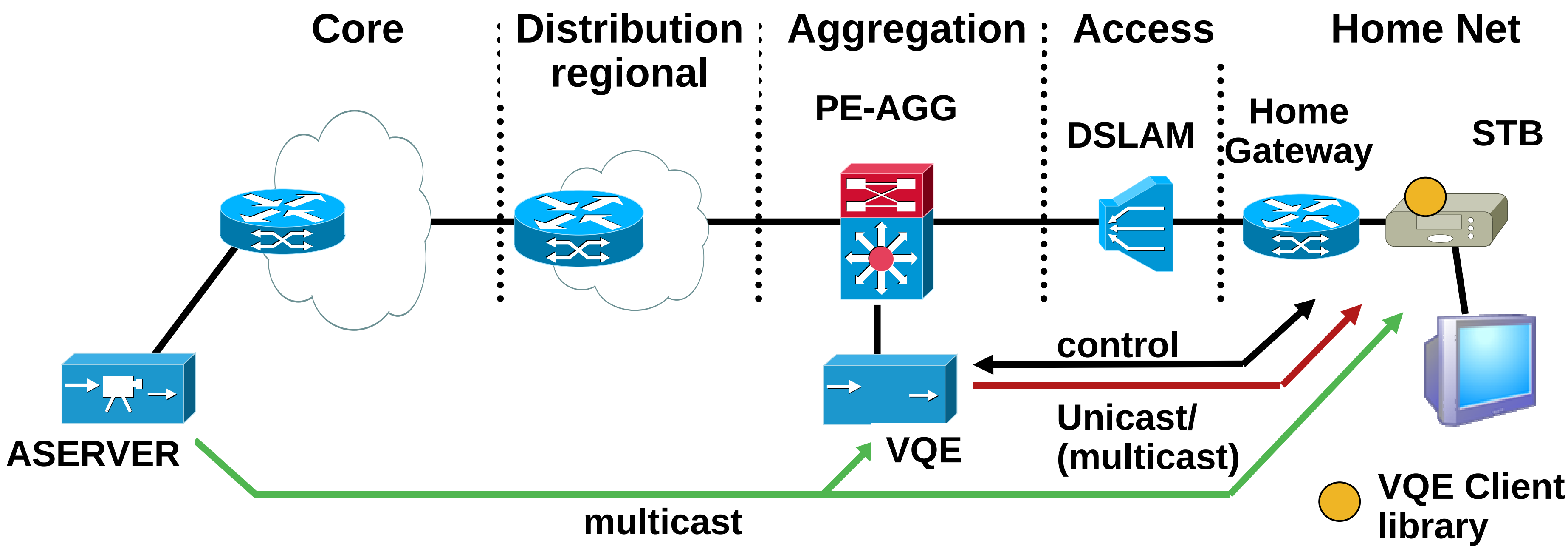
  Can not start decoding before next I-frame

- Need/should-have channel change acceleration for GOP sizes > 0.5 sec ?

- Many codec dependencies:

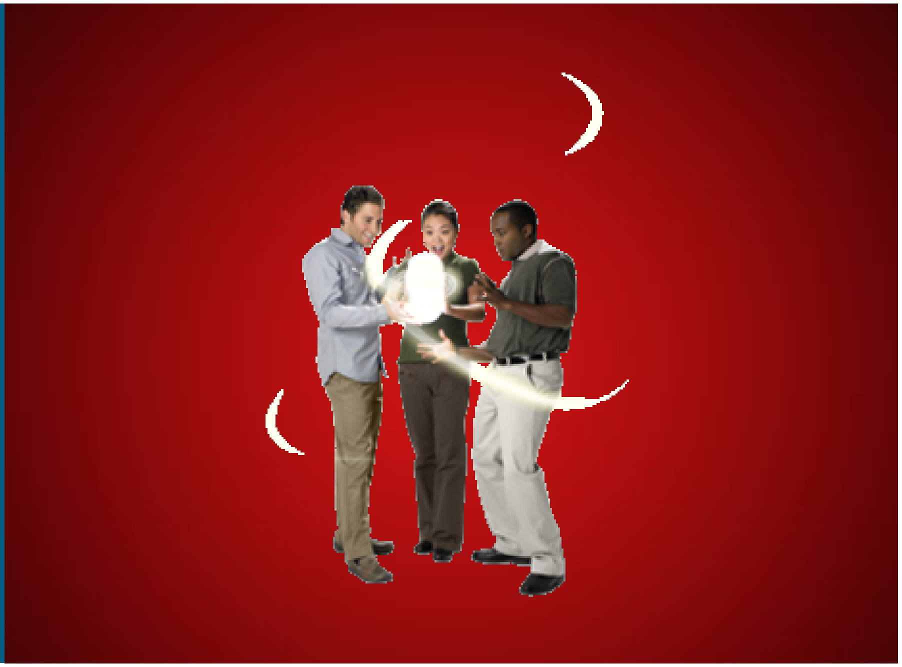
  How much bandwidth is saved in different codecs by raising GOP size but keep the quality.

Video Quality Experience



- Three functions (currently): Video Quality monitoring, FEC/ARQ support for DSL links, Fast Channel change
- Uses standards RTP/RTCP, FEC extensions.
- Fast channel channel by RTCP “retransmission” triggered resend of missing GOP packets from VQE (cached on VQE).

Summary



Multicast and IPTV Summary

- Design IP multicast WITH SSM as generic infrastructure service – for IPTV and beyond
- Select transport design
  - Native IP multicast or mLDP (MPLS core) for most networks
  - RSVP-TE P2MP for eg: contribution network
- Understand your L2 broadband edge specifics
  - IGMPv3 snooping and SSM + lots of options
- Determine appropriate resilience support
- Path selection
  - ECMP and multicast or multiple topologies
- Admission control
  - Router local and broadband specific
- Channel changing
  - GOP size, total performance

Q and A