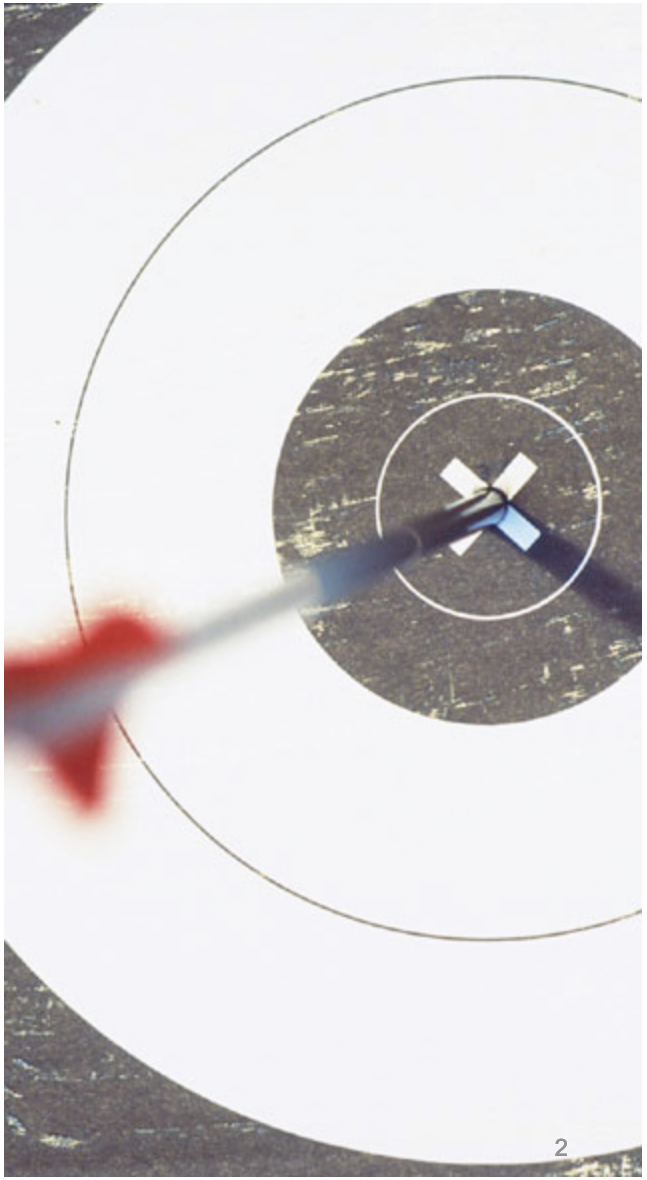
SANOG 14:
Introduction to Multicast

Srini Irigi, SPG TME, Cisco Systems, CCIE 6147

Session Goal

To provide you with a thorough understanding of the concepts, mechanics and protocols used to build IP Multicast networks

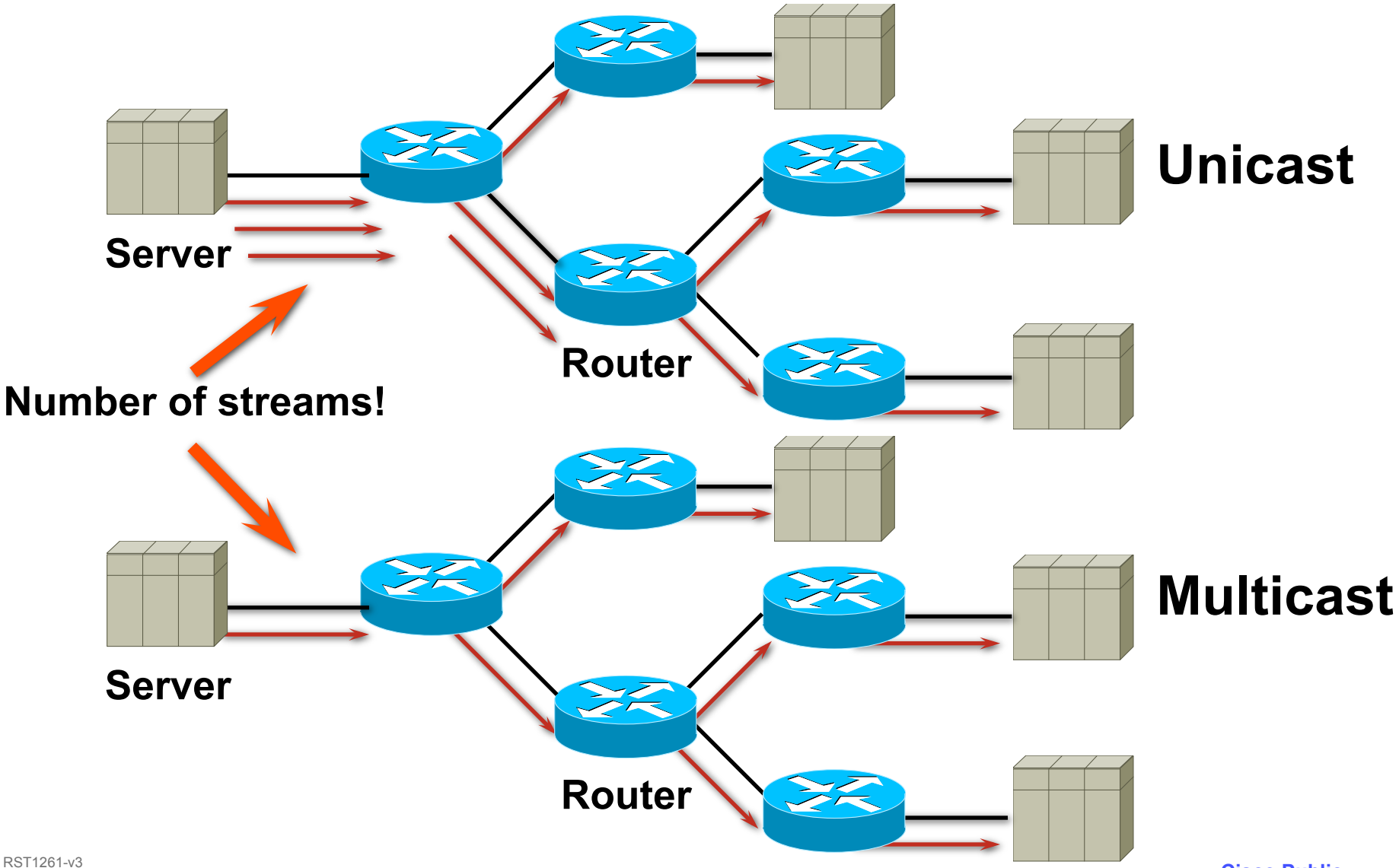


Agenda

- Why Multicast?
- Multicast Fundamentals
- PIM Protocols
- RP choices
- Multicast at Layer 2
- Interdomain IP Multicast
- Some Latest Additions

Why Multicast?

Unicast vs. Multicast



Multicast Uses

- Any Applications with multiple receivers
  - 1-to-many or many-to-many
- Live Video distribution
- Collaborative groupware
- Periodic Data Delivery - "Push" technology
  - stock quotes, sports scores, magazines, newspapers, adverts
- Server/Web-site replication
- Reducing Network/Resource Overhead
  - more than multiple point-to-point flows
- Resource Discovery
- Distributed Interactive Simulation (DIS)
  - wargames
  - virtual reality

Unicast vs. Multicast

- TCP Unicast but NOT Multicast
  - TCP is connection oriented protocol
  - Requires 3 way Handshake
  - Reliable due to sequence numbers + Ack
  - Flow control
- UDP Unicast and Multicast
  - Connectionless
  - Unreliable (application layer awareness)

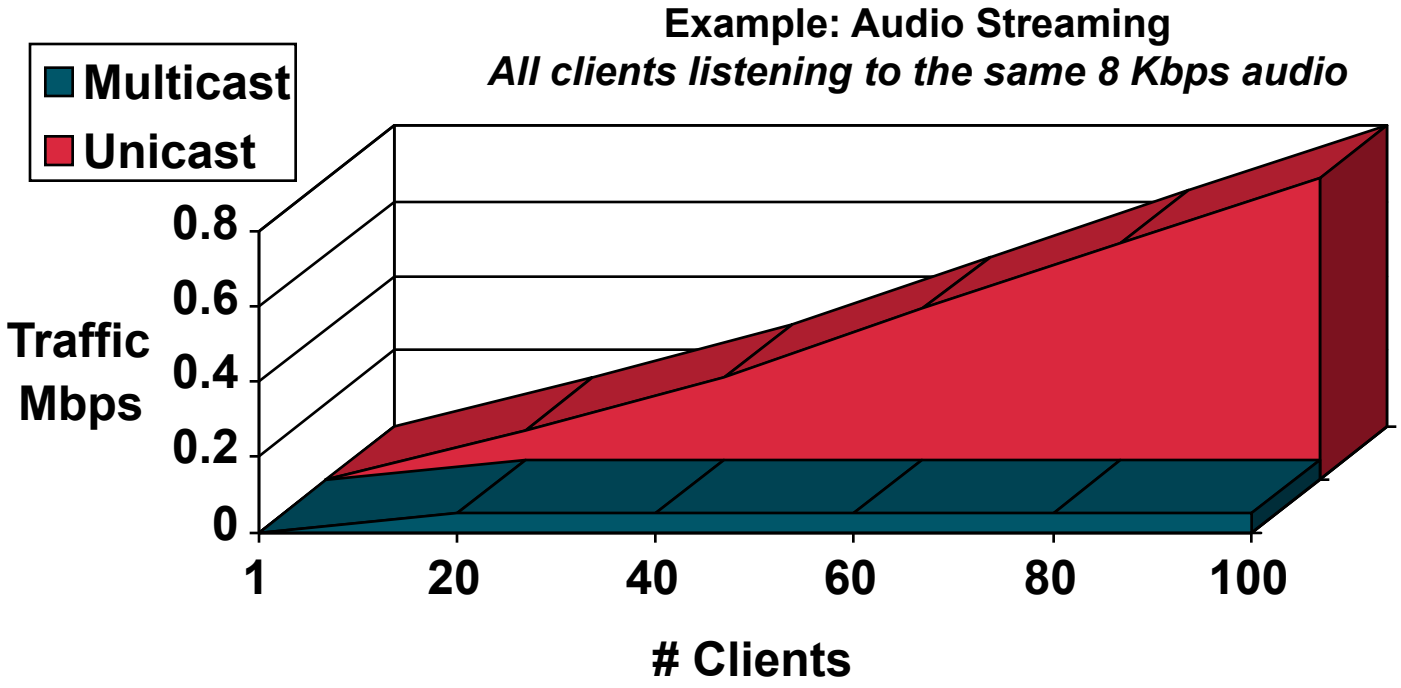
Multicast Disadvantages

Multicast Is UDP Based!!!

- ***Best Effort Delivery*: Drops are to be expected. Multicast applications should not expect reliable delivery of data and should be designed accordingly. Reliable Multicast is still an area for much research. Expect to see more developments in this area. PGM, FEC, QoS**
- ***No Congestion Avoidance*: Lack of TCP windowing and "slow-start" mechanisms can result in network congestion. If possible, Multicast applications should attempt to detect and avoid congestion conditions.**
- ***Duplicates*: Some multicast protocol mechanisms (e.g. Asserts, Registers and SPT Transitions) result in the occasional generation of duplicate packets. Multicast applications should be designed to expect occasional duplicate packets.**
- ***Out of Order Delivery* : Some protocol mechanisms may also result in out of order delivery of packets.**

Multicast Advantages

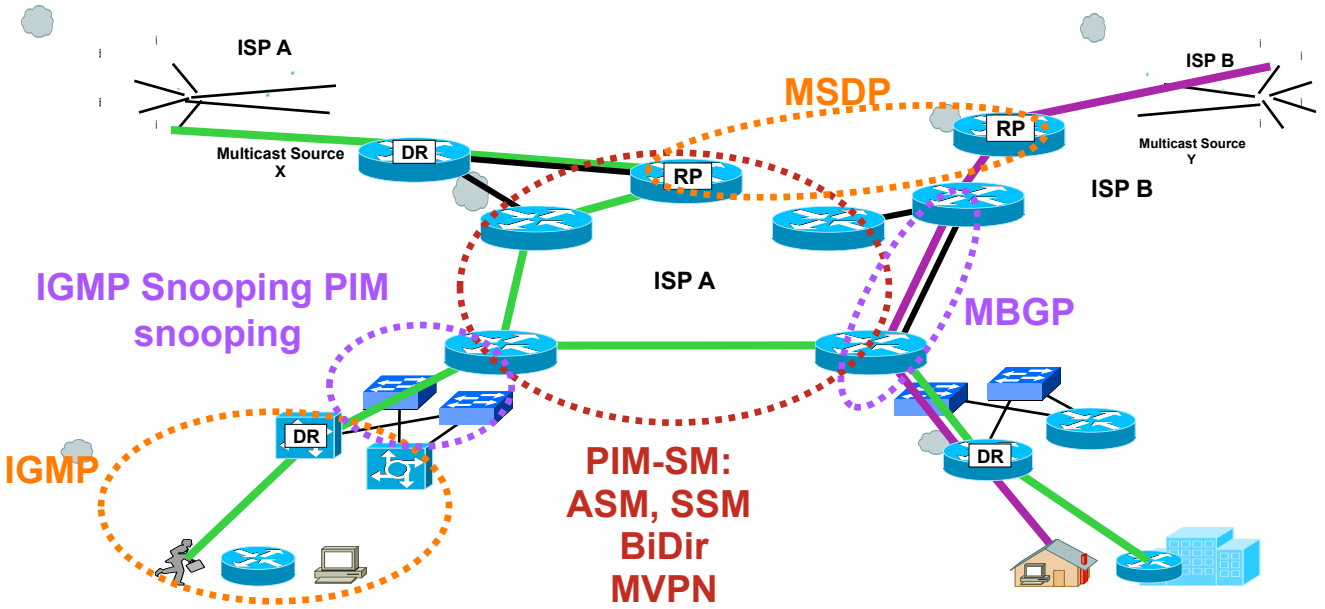
- ***Enhanced Efficiency*: Controls network traffic and reduces server and CPU loads**
- ***Optimized Performance*: Eliminates traffic redundancy**
- ***Distributed Applications*: Makes multipoint applications possible**



Multicast Fundamentals

Multicast Components

Cisco End-to-End Architecture



- End Stations (hosts-to-routers):
  - IGMP
- Multicast routing across domains
  - MBGP

Campus Multicast

Switches (Layer 2 Optimization):
IGMP Snooping PIM snooping

Routers (Multicast Forwarding Protocol):
PIM Sparse Mode or Bidirectional PIM

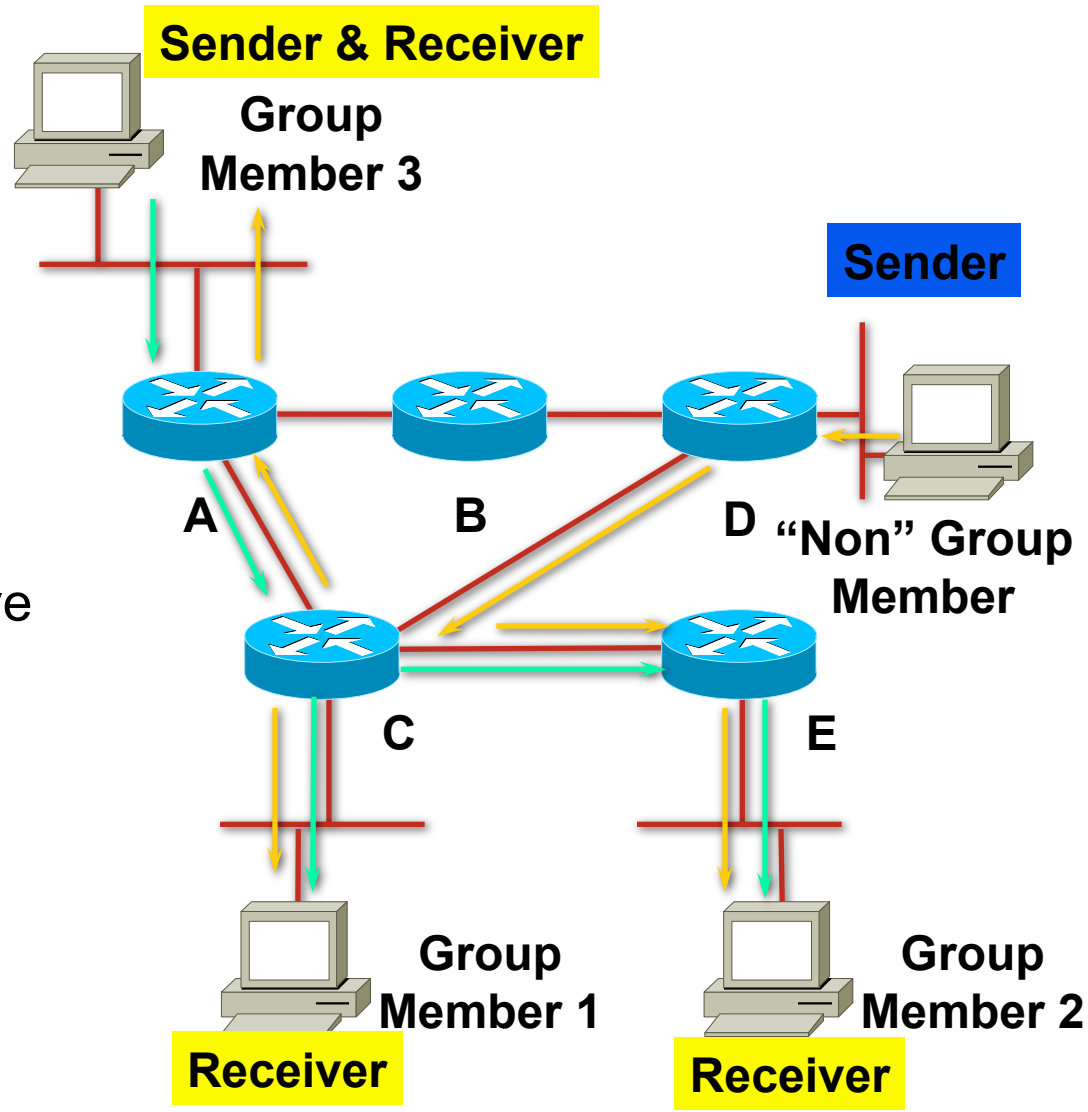
Interdomain Multicast

Multicast Source Discovery
MSDP with ASM

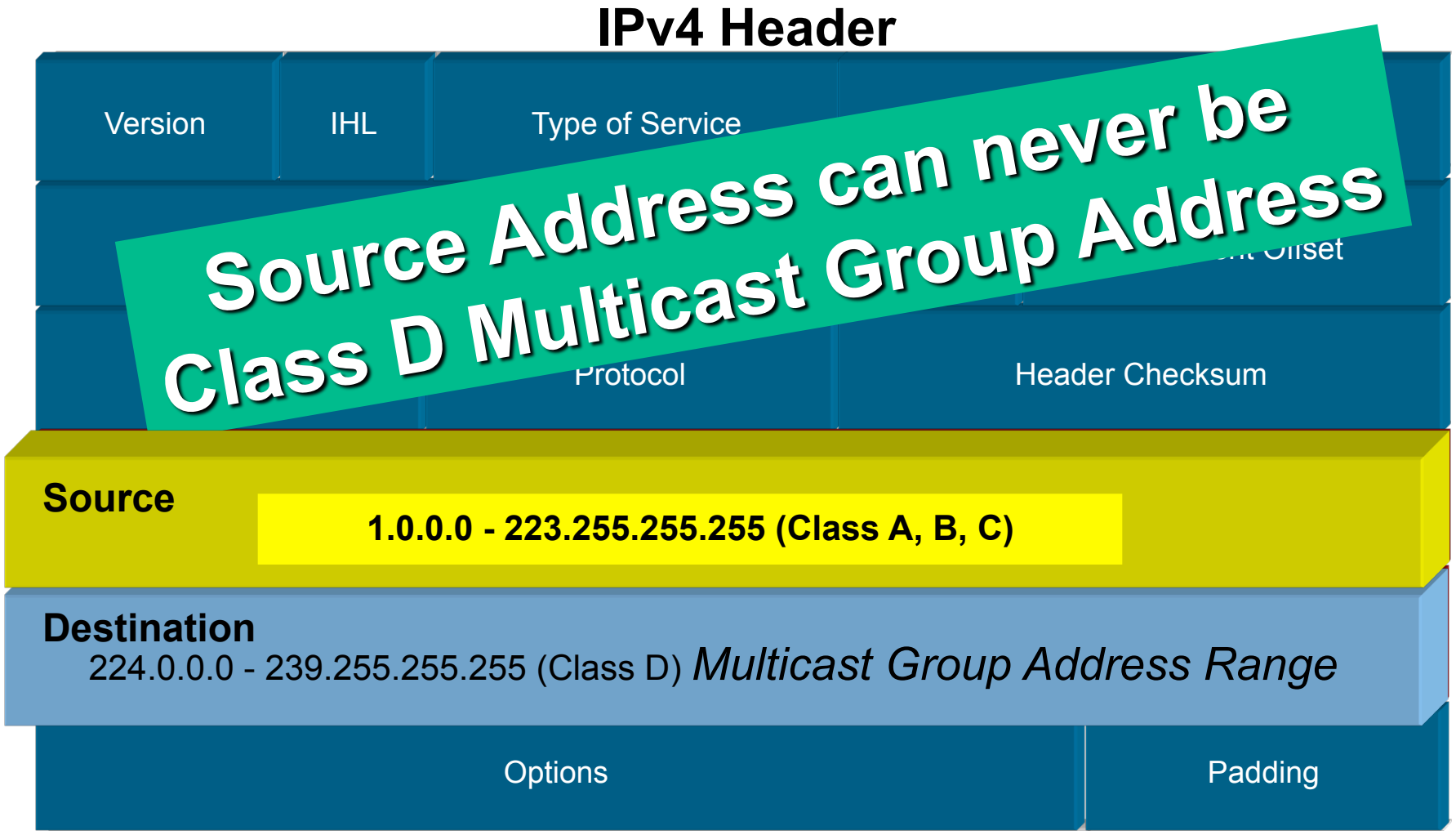
Source Specific Multicast
SSM

IP Multicast Group Concept

1. You must be a "member" of a group to receive its data
2. If you send to a group address, all members receive it
3. You do not have to be a member of a group to send to a group



Multicast Addressing



Multicast Addressing - 224/4

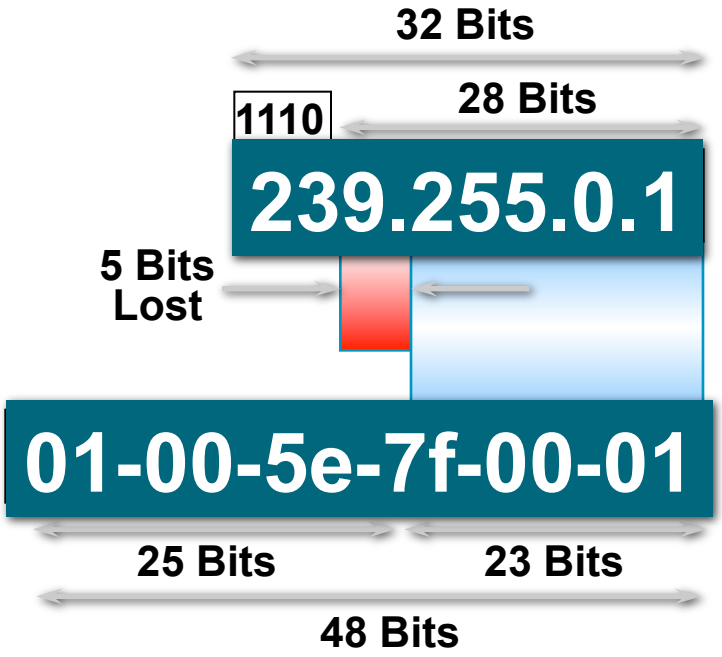
- Reserved Link-Local Addresses
  - 224.0.0.0 – 224.0.0.255
  - Transmitted with TTL = 1
  - Examples:
    - 224.0.0.1 All systems on this subnet
    - 224.0.0.2 All routers on this subnet
    - 224.0.0.5 OSPF routers
    - 224.0.0.13 PIMv2 Routers
    - 224.0.0.22 IGMPv3
- Other Reserved Addresses
  - 224.0.1.0 – 224.0.1.255
  - Not local in scope (Transmitted with TTL > 1)
  - Examples:
    - 224.0.1.1 NTP Network Time Protocol
    - 224.0.1.32 Mtrace routers
    - 224.0.1.78 Tibco Multicast1

Multicast Addressing - 224/4

- Administratively Scoped Addresses
  - 239.0.0.0 – 239.255.255.255
  - Private address space
    - Similar to RFC1918 unicast addresses
    - Not used for global Internet traffic - scoped traffic
- SSM (Source Specific Multicast) Range
  - 232.0.0.0 – 232.255.255.255
  - Primarily targeted for Internet style Broadcast
- GLOP (honest, it's not an acronym)
  - 233.0.0.0 - 233.255.255.255
  - Provides /24 group prefix per ASN

Multicast Addressing

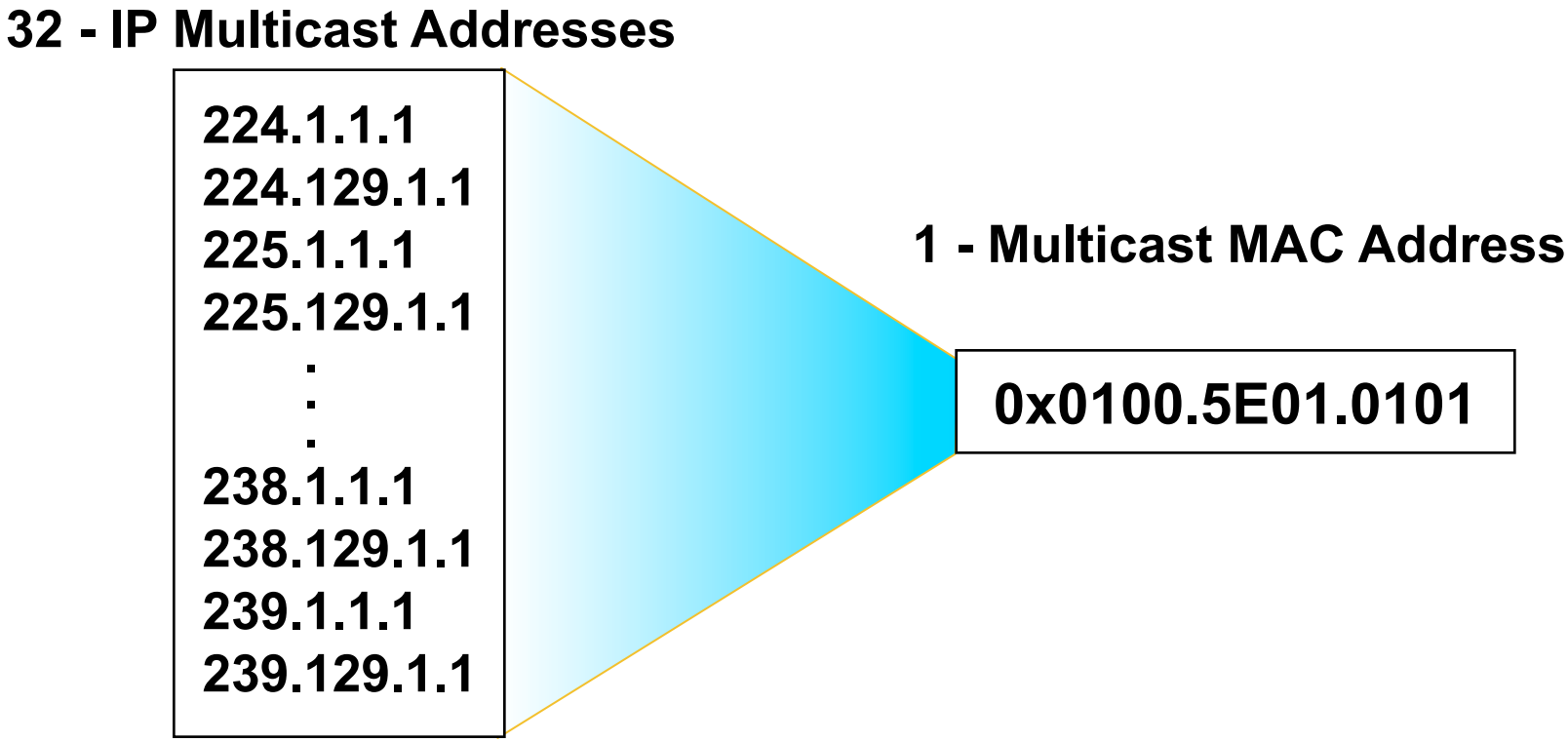
IP Multicast MAC Address Mapping



Multicast Addressing

IP Multicast MAC Address Mapping

Be Aware of the 32:1 Address Overlap



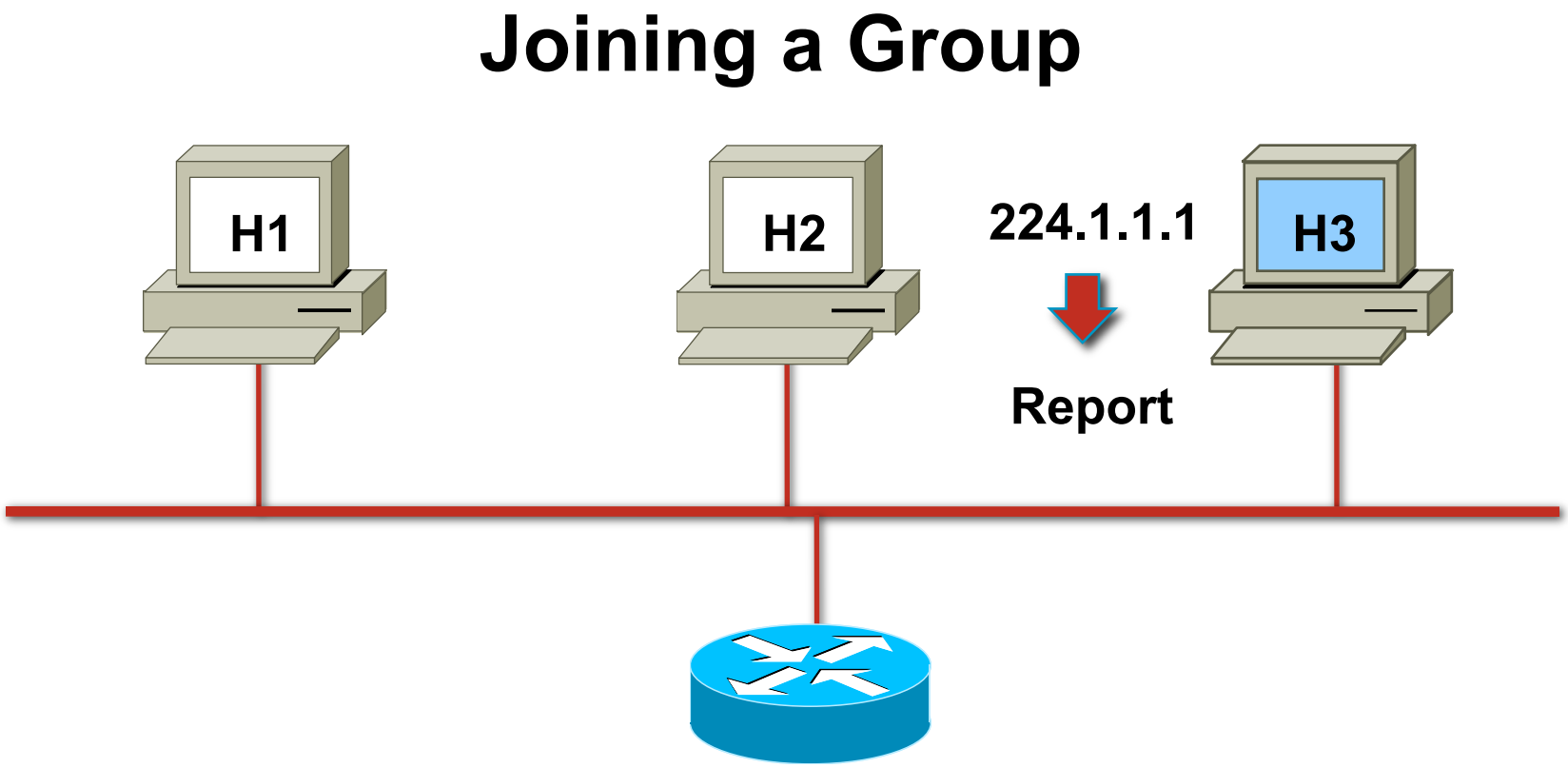
How are Multicast Addresses Assigned?

- Static Global Group Address Assignment
  - Temporary method to meet immediate needs
  - Group range: 233.0.0.0 – 233.255.255.255
    - Your AS number is inserted in middle two octets
    - Remaining low-order octet used for group assignment
  - Defined in RFC 2770
    - “GLOP Addressing in 233/8”
- Manual Address Allocation by the Admin !!
  - Is still the most common practice

Host-Router Signaling: IGMP

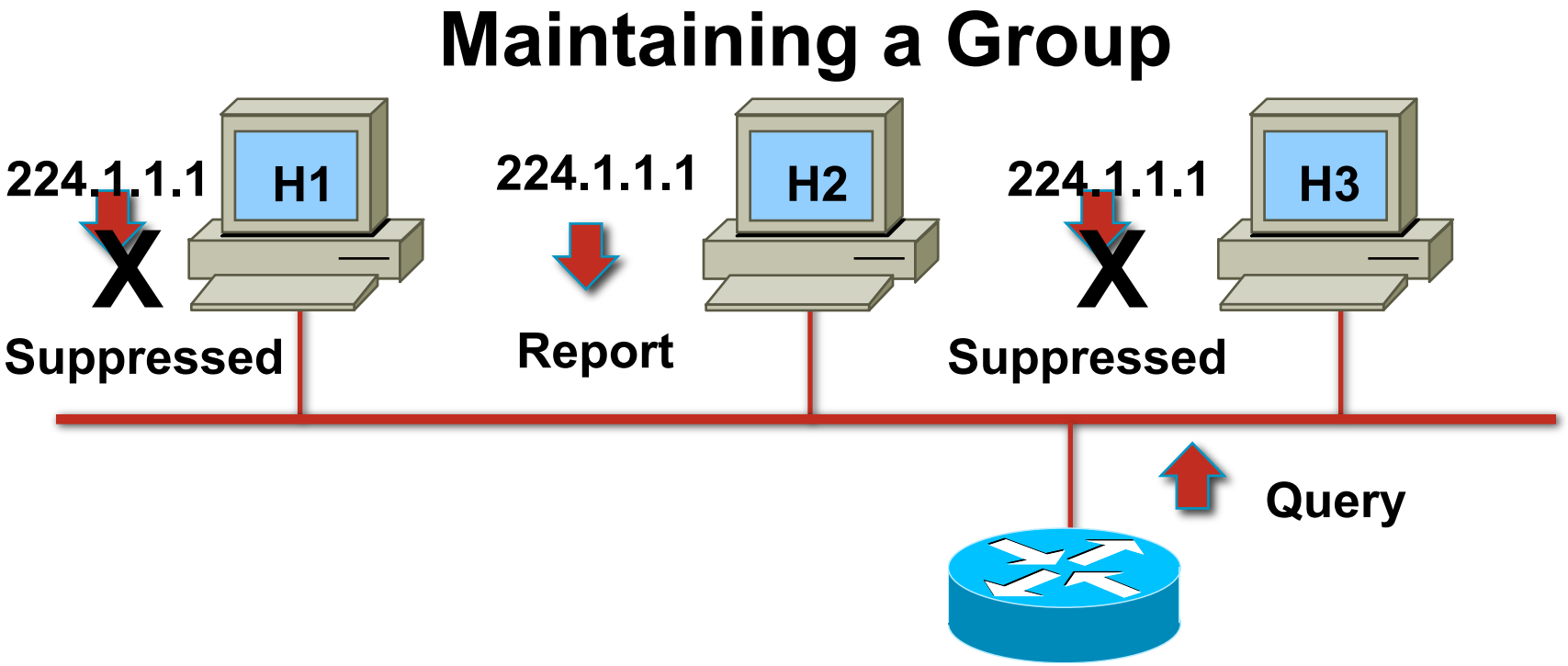
- How hosts tell routers about group membership
- Routers solicit group membership from directly connected hosts
- RFC 1112 specifies version 1 of IGMP
  - Supported on Windows 95
- RFC 2236 specifies version 2 of IGMP
  - Supported on latest service pack for Windows and most UNIX systems
- RFC 3376 specifies version 3 of IGMP
  - Supported in Window XP and various UNIX systems

Host-Router Signaling: IGMP



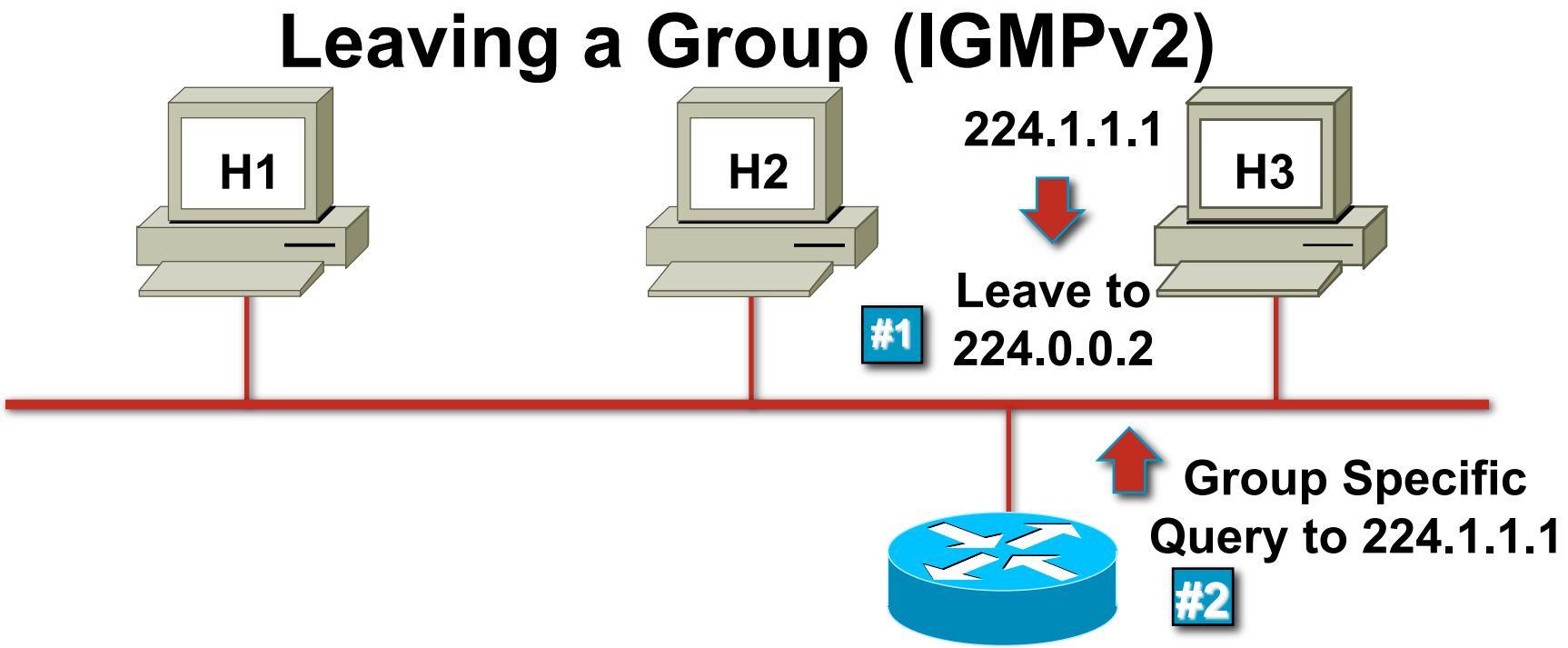
- Host sends IGMP Report to join group

Host-Router Signaling: IGMP



- Router sends periodic Queries to 224.0.0.1
- One member per group per subnet reports
- Other members suppress reports

Host-Router Signaling: IGMP



- Host sends Leave message to 224.0.0.2
- Router sends Group specific query to 224.1.1.1
- No IGMP Report is received within ~3 seconds
- Group 224.1.1.1 times out

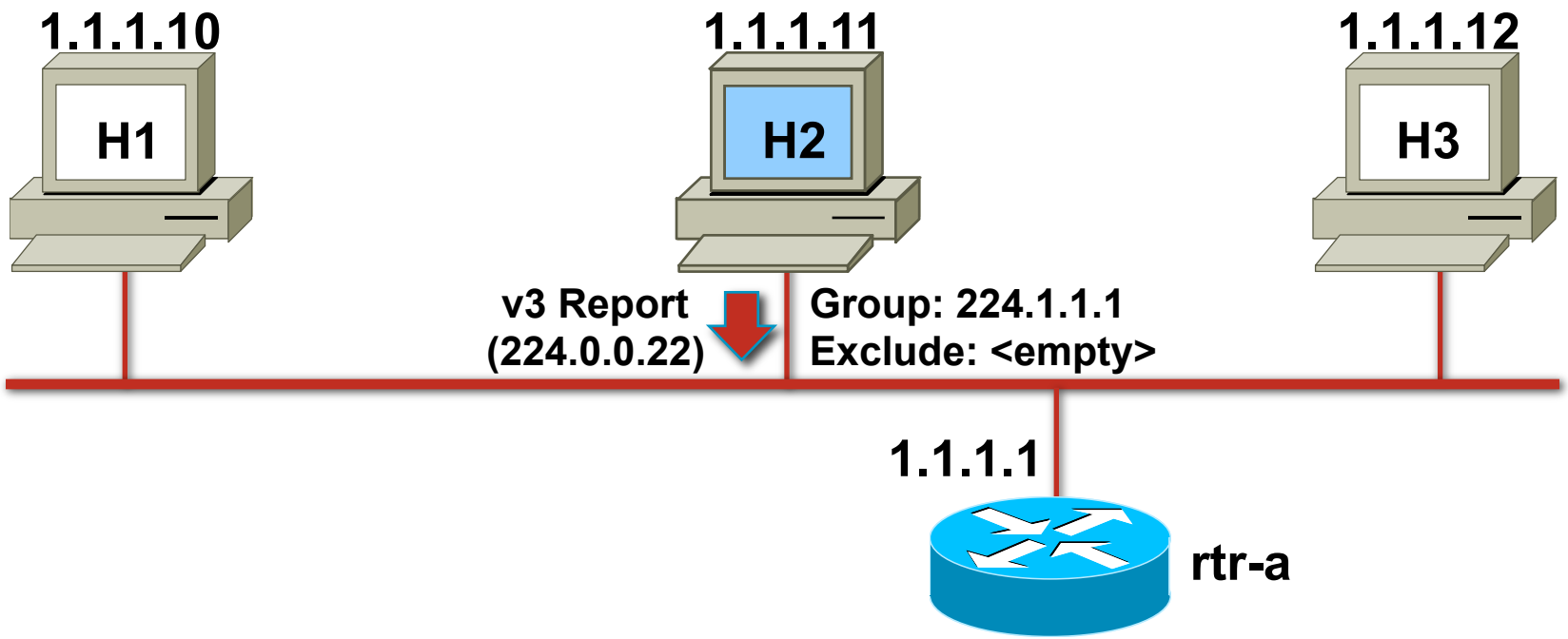
Host-Router Signaling: IGMPv3

- RFC 3376
  - Adds Include/Exclude Source Lists
  - Enables hosts to listen only to a specified subset of the hosts sending to the group
  - Requires new 'IPMulticastListen' API
    - New IGMPv3 stack required in the O/S.
  - Apps must be rewritten to use IGMPv3 Include/ Exclude features

Host-Router Signaling: IGMPv3

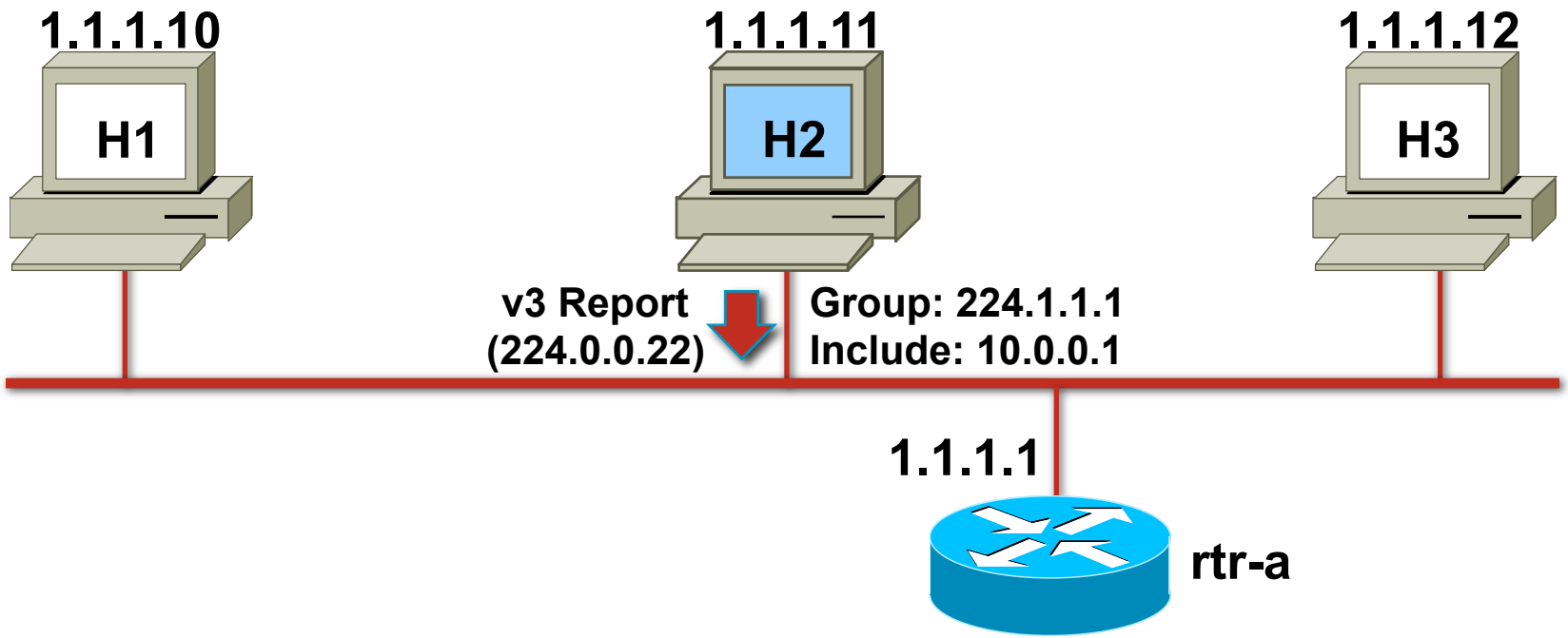
- New Membership Report address
  - 224.0.0.22 (IGMPv3 Routers)
    - All IGMPv3 Hosts send reports to this address
      - Instead of the target group address as in IGMPv1/v2
    - All IGMPv3 Routers listen to this address
    - Hosts do not listen or respond to this address
  - No Report Suppression
    - All Hosts on wire respond to Queries
      - Host's complete IGMP state sent in single response
    - Response Interval may be tuned over broad range
      - Useful when large numbers of hosts reside on subnet

IGMPv3—Joining a Group



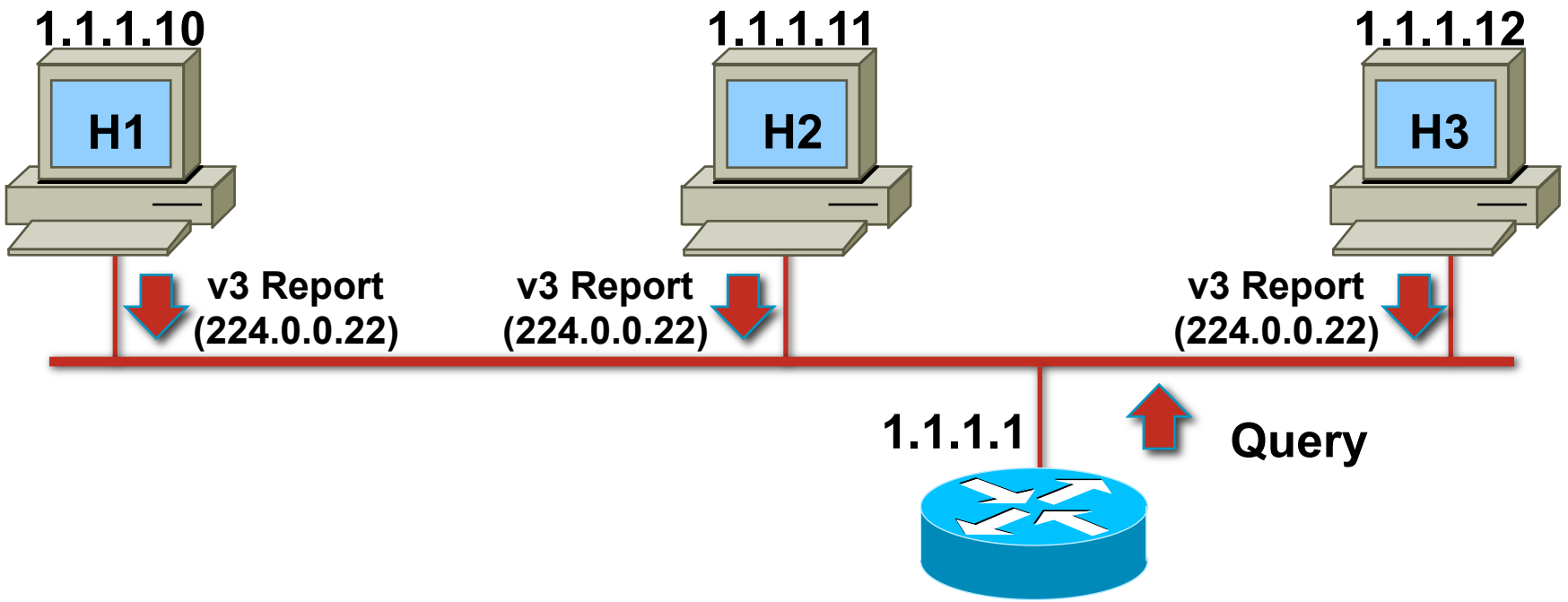
- Joining member sends IGMPv3 Report to 224.0.0.22 immediately upon joining

IGMPv3—Joining specific Source(s)



- IGMPv3 Report contains desired source(s) in the Include list.
- Only “Included” source(s) are joined.

IGMPv3—Maintaining State



- Router sends periodic queries
- All IGMPv3 members respond
- Reports contain multiple Group state records

Multicast L3 Forwarding

- Multicast Routing is backwards from Unicast Routing
  - Unicast Routing is concerned about where the packet is going.
  - Multicast Routing is concerned about where the packet came from.

Unicast vs. Multicast Forwarding

- Unicast Forwarding
  - Destination IP address directly indicates where to forward packet.
  - Forwarding is hop-by-hop.
    - Unicast routing table determines interface and next-hop router to forward packet.

Unicast vs. Multicast Forwarding

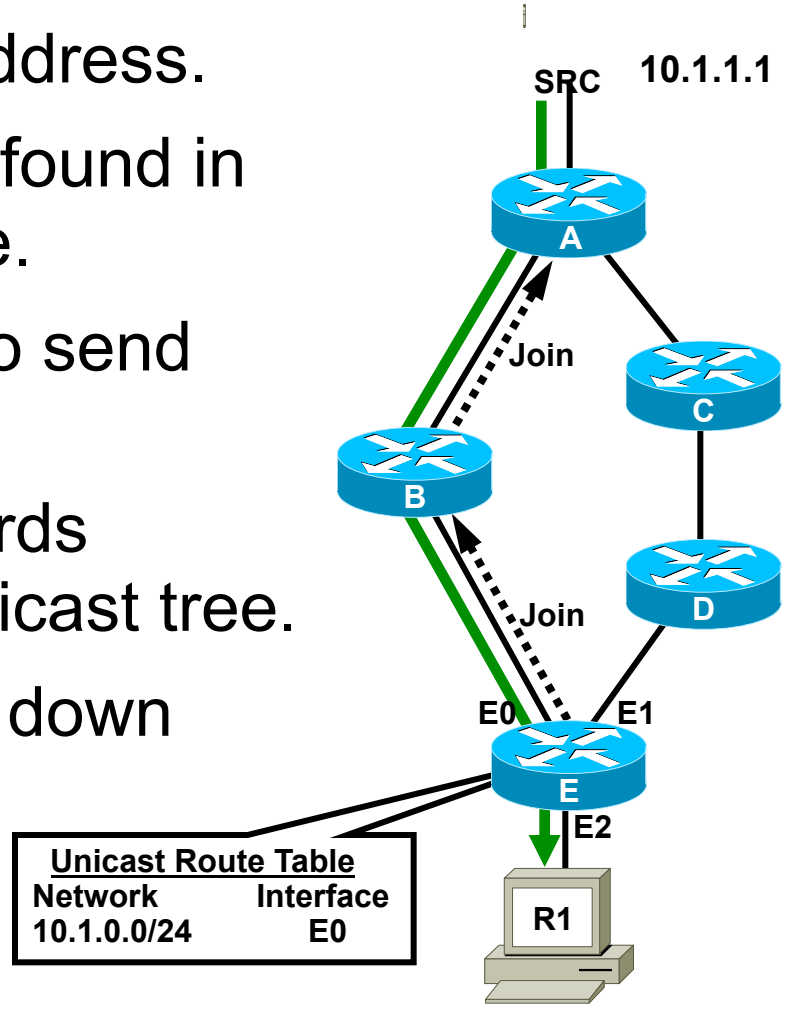
- Multicast Forwarding
  - Destination IP address (group) doesn't directly indicate where to forward packet.
  - Forwarding is connection-oriented.
    - Receivers must first be "connected" to the tree before traffic begins to flow.
      - Connection messages (PIM Joins) follow unicast routing table toward multicast source.
      - Build Multicast Distribution Trees that determine where to forward packets.
      - Distribution Trees rebuilt dynamically in case of network topology changes.

Reverse Path Forwarding (RPF)

- The RPF Calculation
  - The multicast source address is checked against the unicast routing table.
  - This determines the interface and upstream router in the direction of the source to which PIM Joins are sent.
  - This interface becomes the “Incoming” or RPF interface.
    - A router forwards a multicast datagram only if received on the RPF interface.

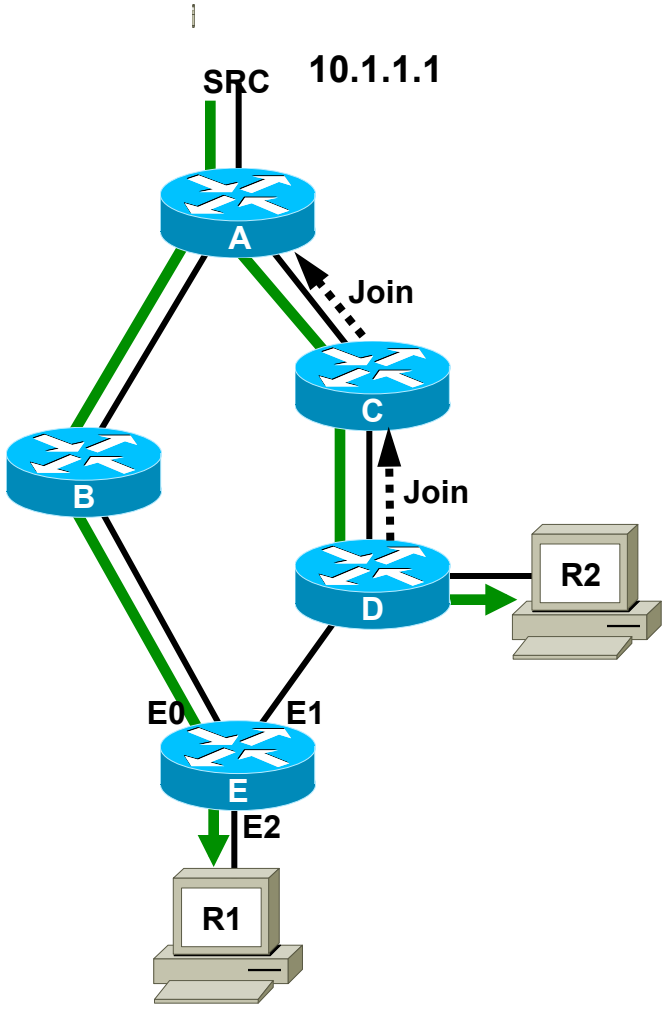
Reverse Path Forwarding (RPF)

- RPF Calculation
  - Based on Source Address.
  - Best path to source found in Unicast Route Table.
  - Determines where to send Join.
  - Joins continue towards Source to build multicast tree.
  - Multicast data flows down tree.



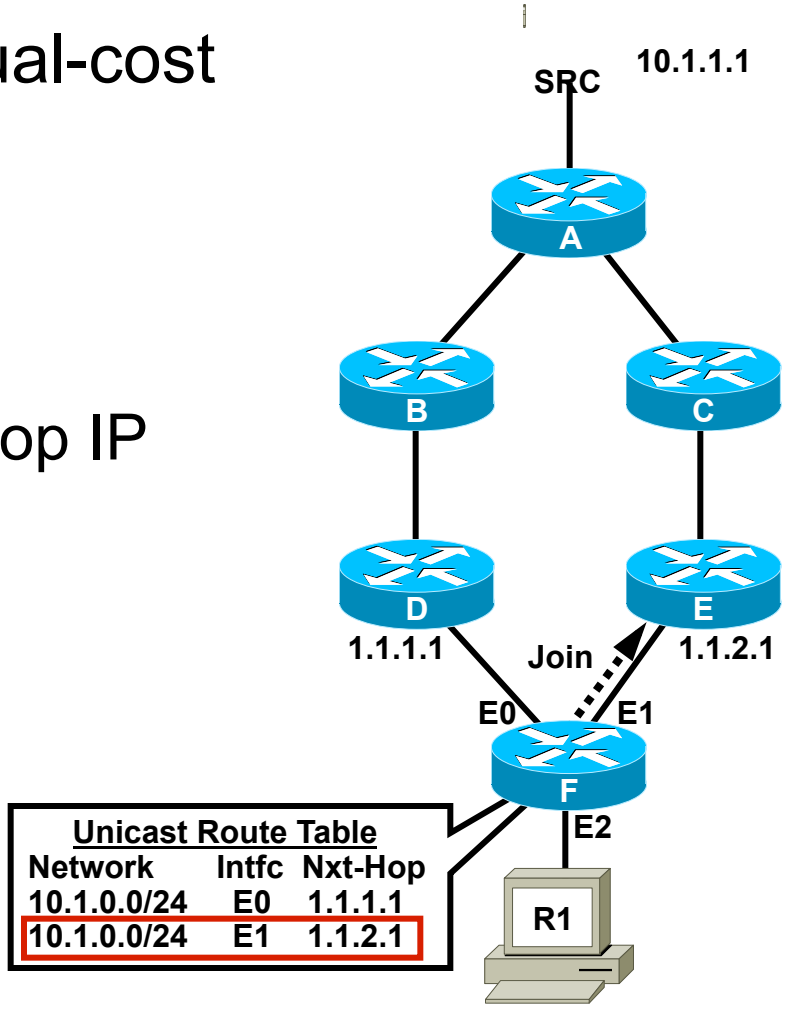
Reverse Path Forwarding (RPF)

- RPF Calculation
  - Based on Source Address.
  - Best path to source found in Unicast Route Table.
  - Determines where to send Join.
  - Joins continue towards Source to build multicast tree.
  - Multicast data flows down tree.
  - Repeat for other receivers.



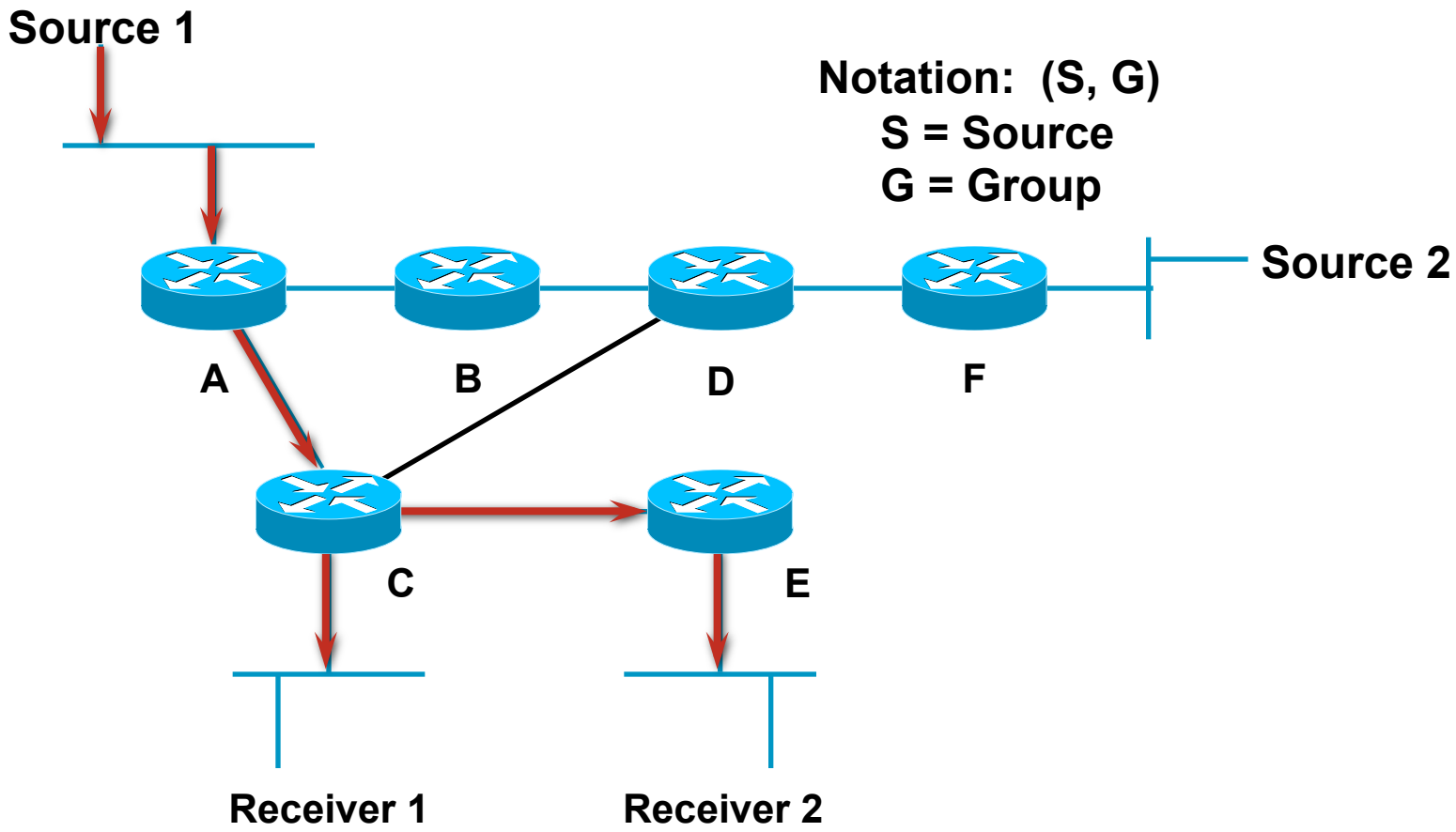
Reverse Path Forwarding (RPF)

- RPF Calculation
  - What if we have equal-cost paths?
    - We can't use both.
  - Tie-Breaker
    - Use highest Next-Hop IP address.



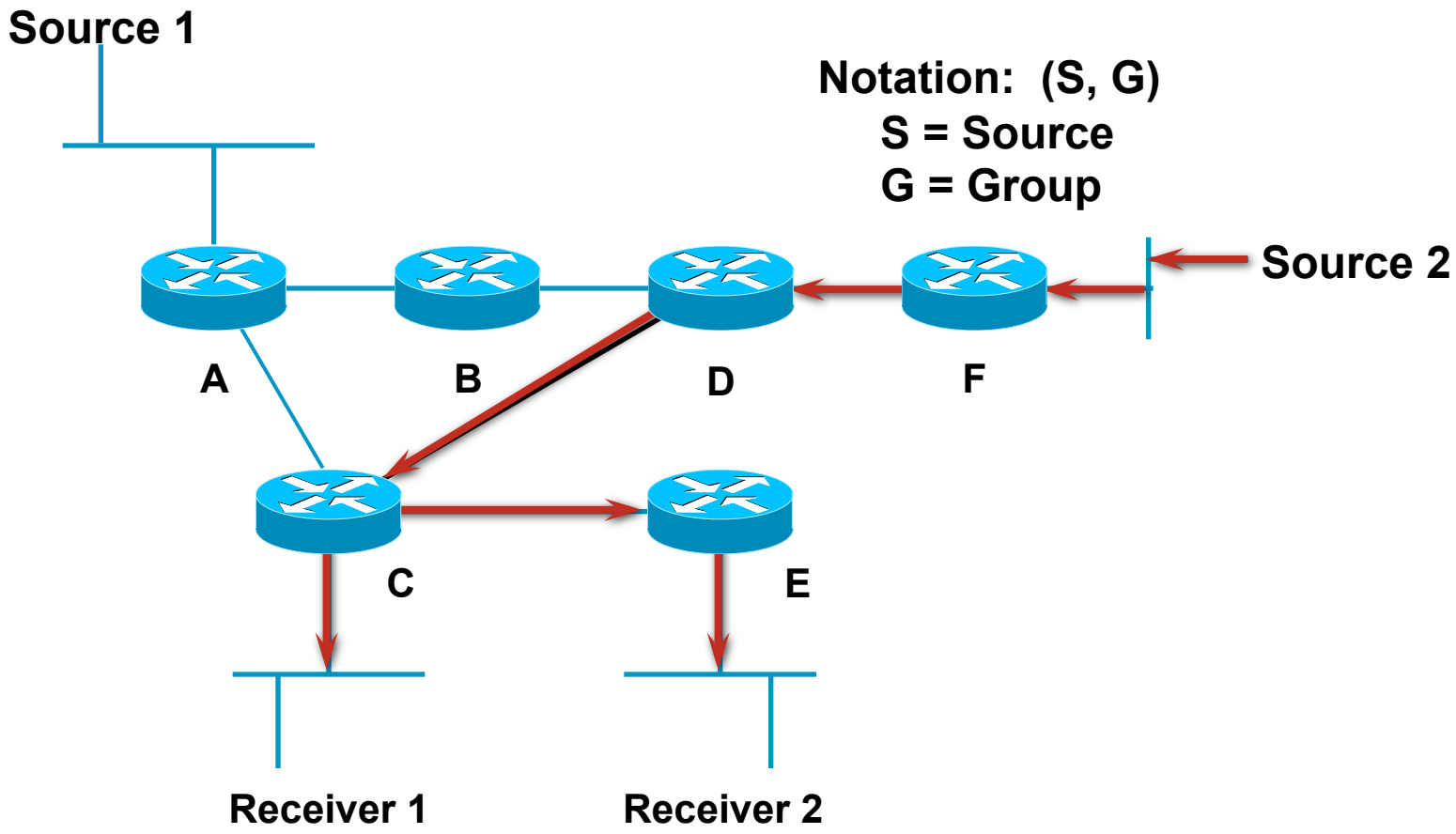
Multicast Distribution Trees

Shortest Path or Source Distribution Tree



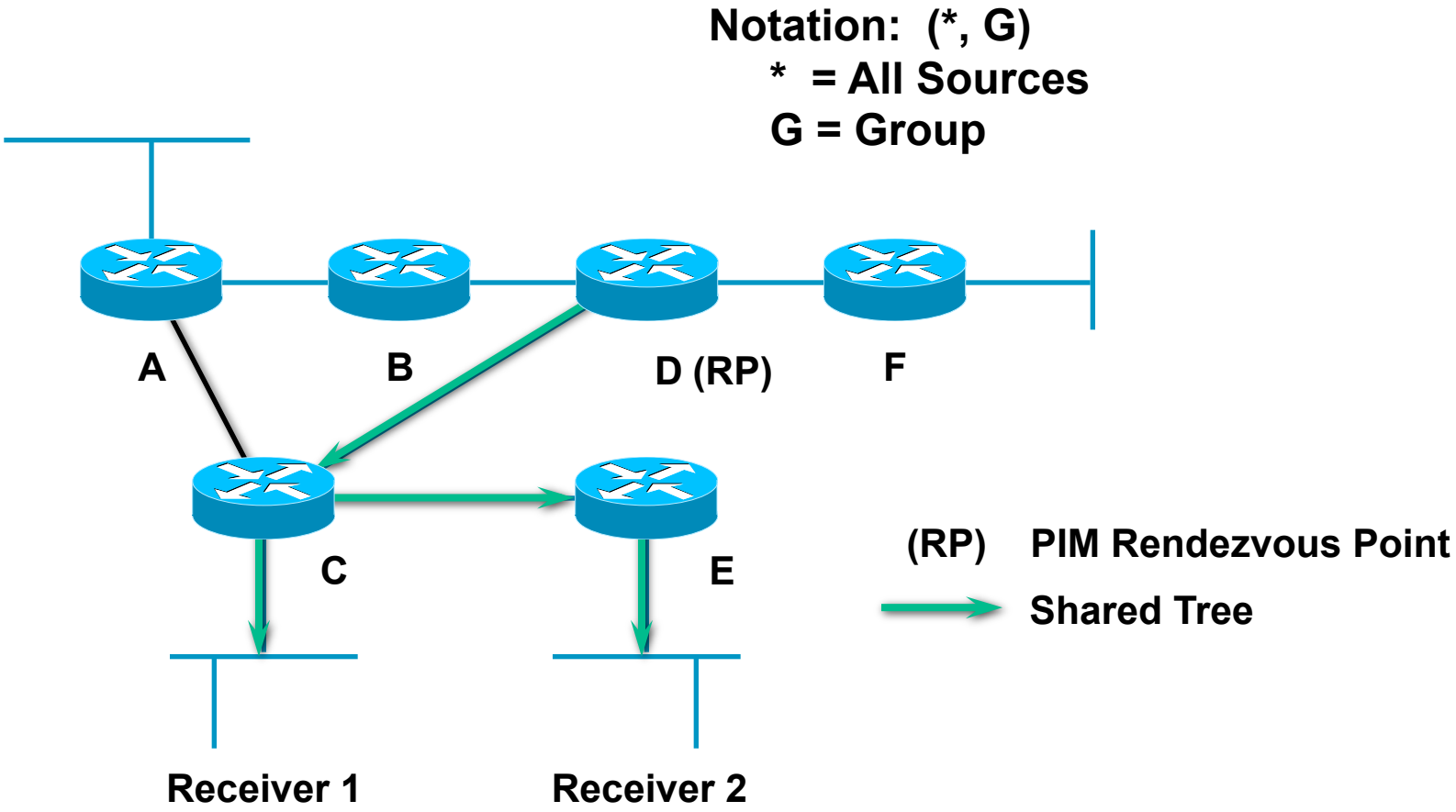
Multicast Distribution Trees

Shortest Path or Source Distribution Tree



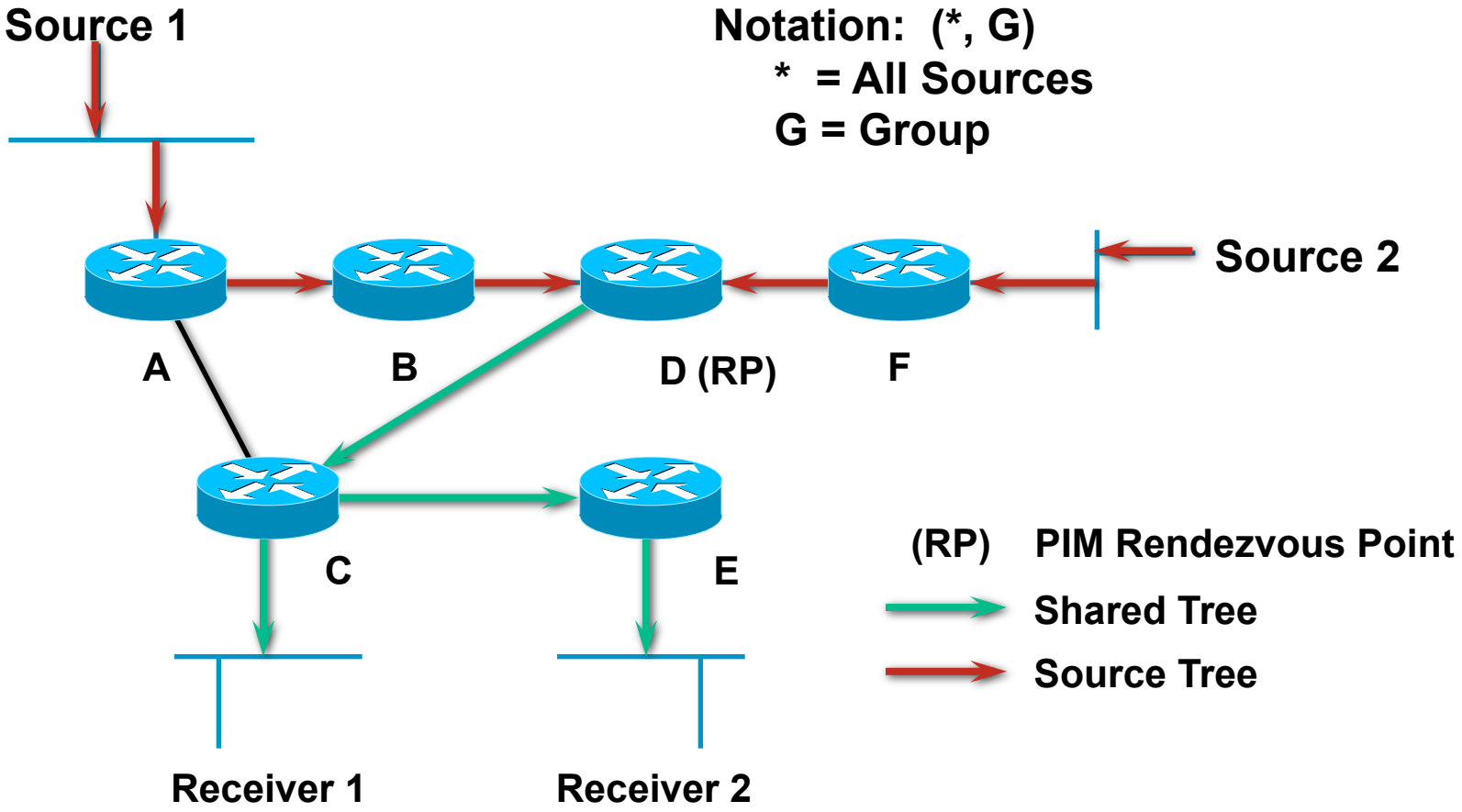
Multicast Distribution Trees

Shared Distribution Tree



Multicast Distribution Trees

Shared Distribution Tree



Multicast Distribution Trees

Characteristics of Distribution Trees

- Source or Shortest Path trees
  - Uses more memory O(S x G) but you get optimal paths from source to all receivers; minimizes delay
- Shared trees
  - Uses less memory O(G) but you may get sub-optimal paths from source to all receivers; may introduce extra delay

Multicast Tree creation

- PIM Join/Prune Control Messages
  - Used to create/remove Distribution Trees
- Shortest Path trees
  - PIM control messages are sent toward the Source
- Shared trees
  - PIM control messages are sent toward RP

PIM Protocol Variants

Major deployed PIM variants

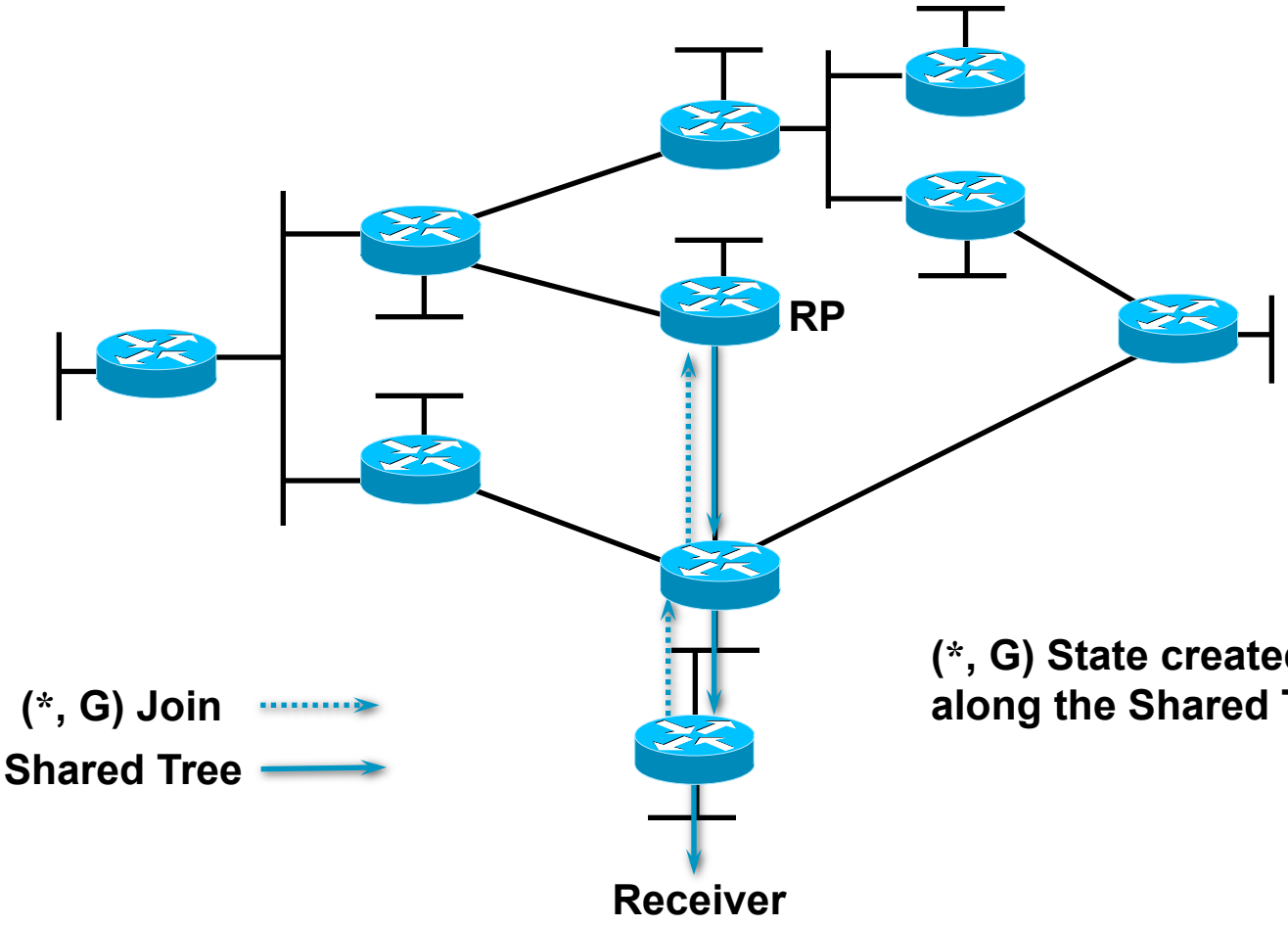
- PIM-SM
  - ASM

    Any Source Multicast / RP / SPT / Shared Tree
  - SSM

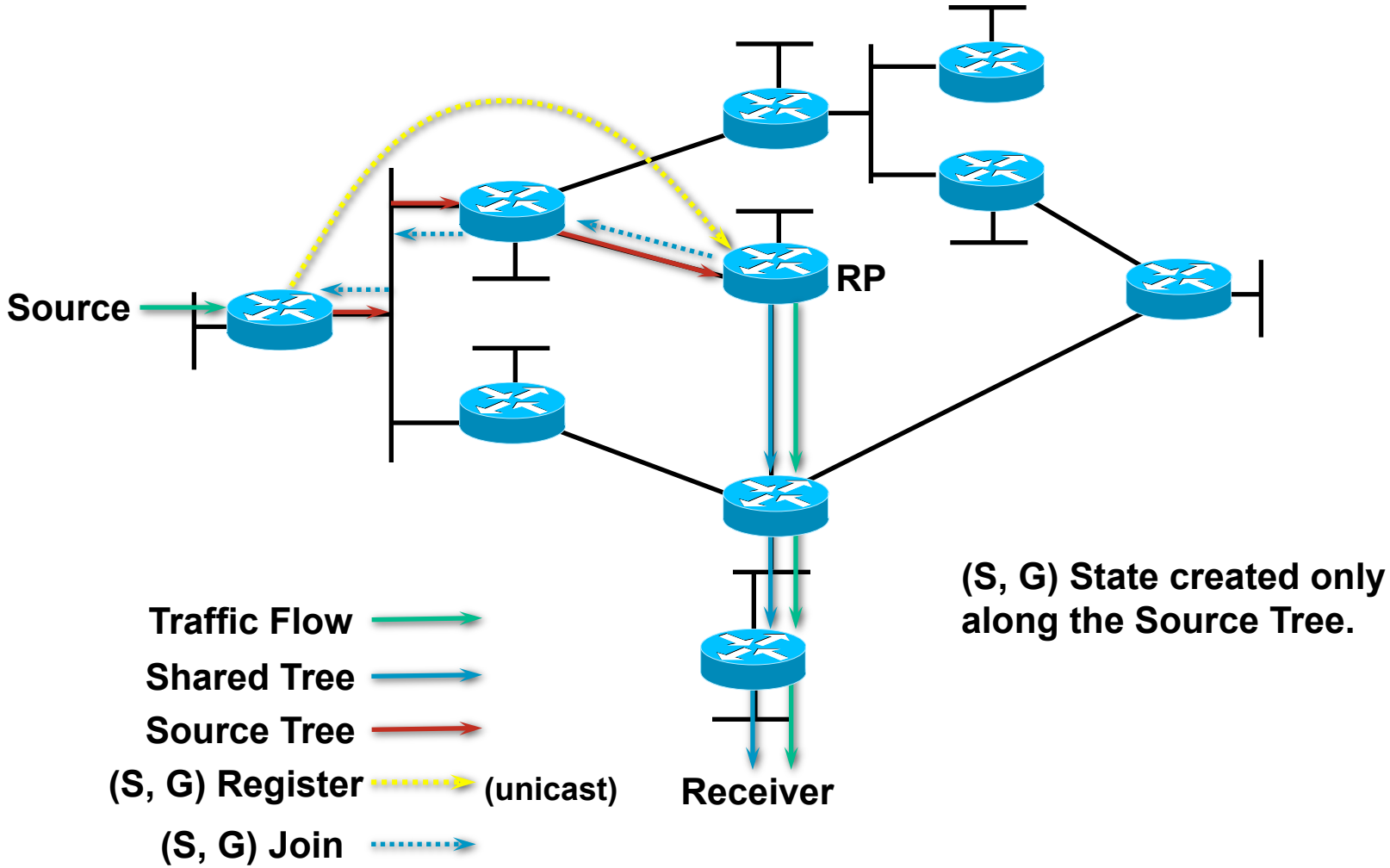
    Source Specific Multicast, no RP, SPT only
  - BiDir

    BiDirectional PIM, no SPT, Shared tree only

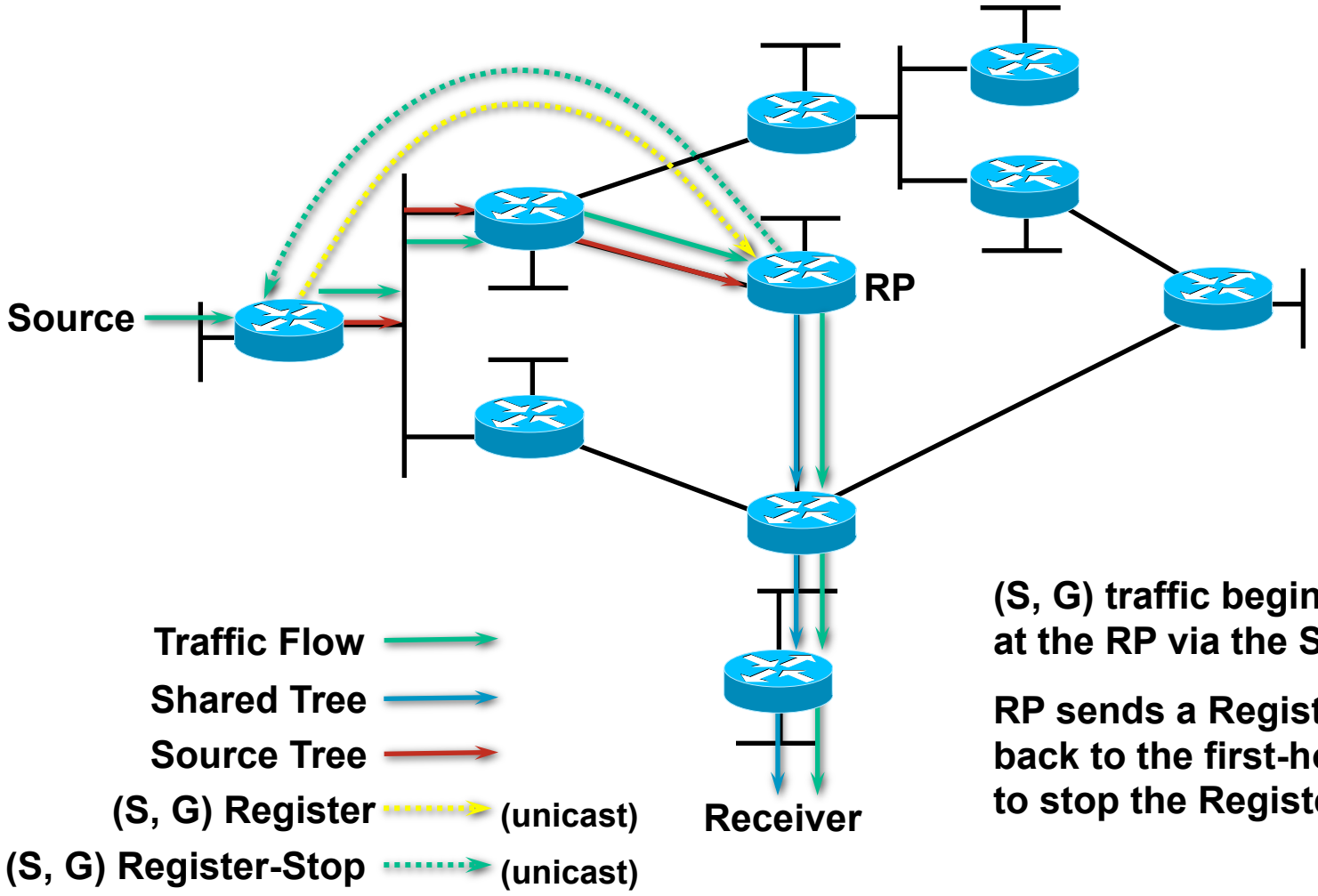
PIM-SM Shared Tree Join



PIM-SM Sender Registration



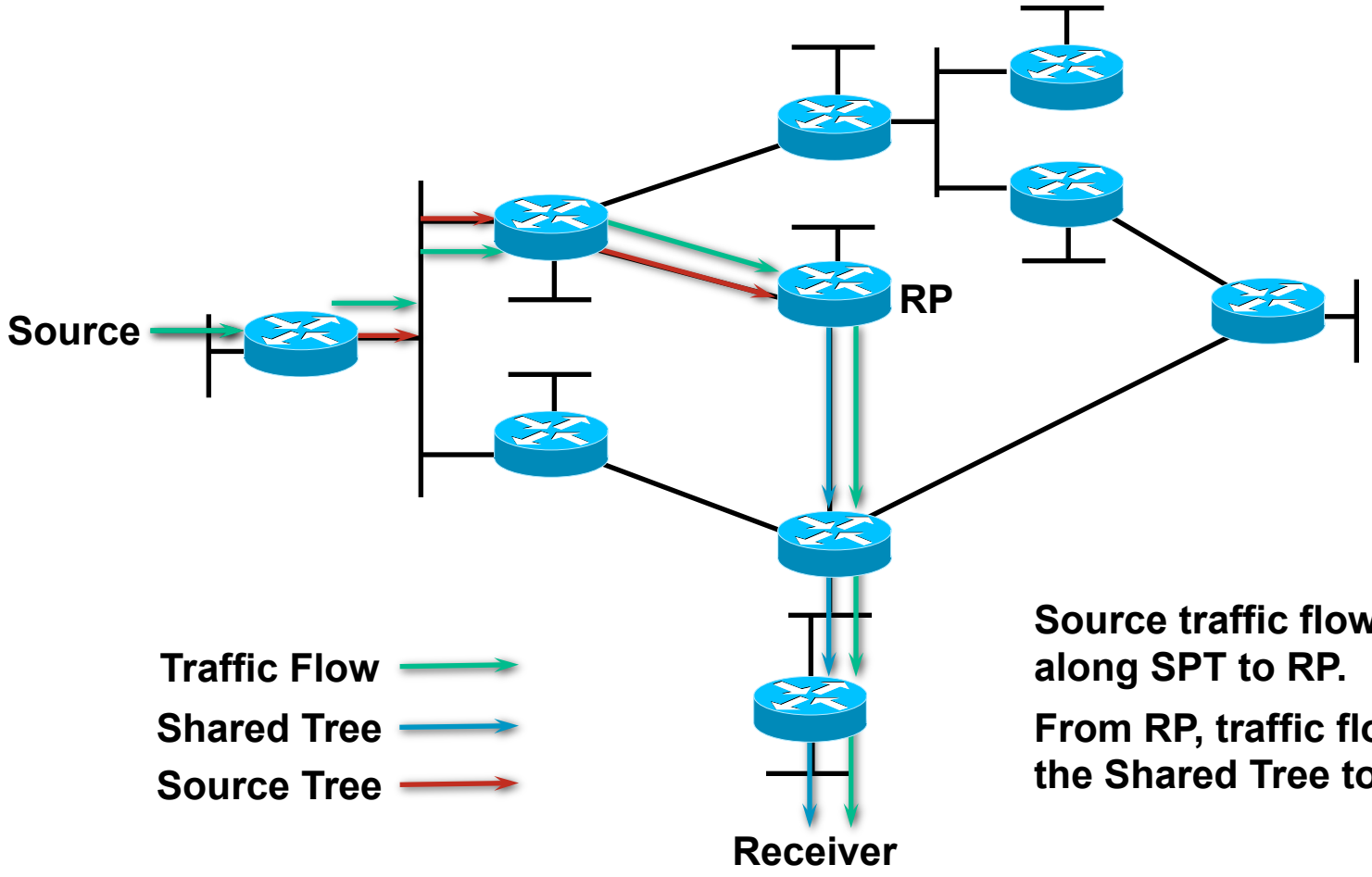
PIM-SM Sender Registration



(S, G) traffic begins arriving at the RP via the Source tree.

RP sends a Register-Stop back to the first-hop router to stop the Register process.

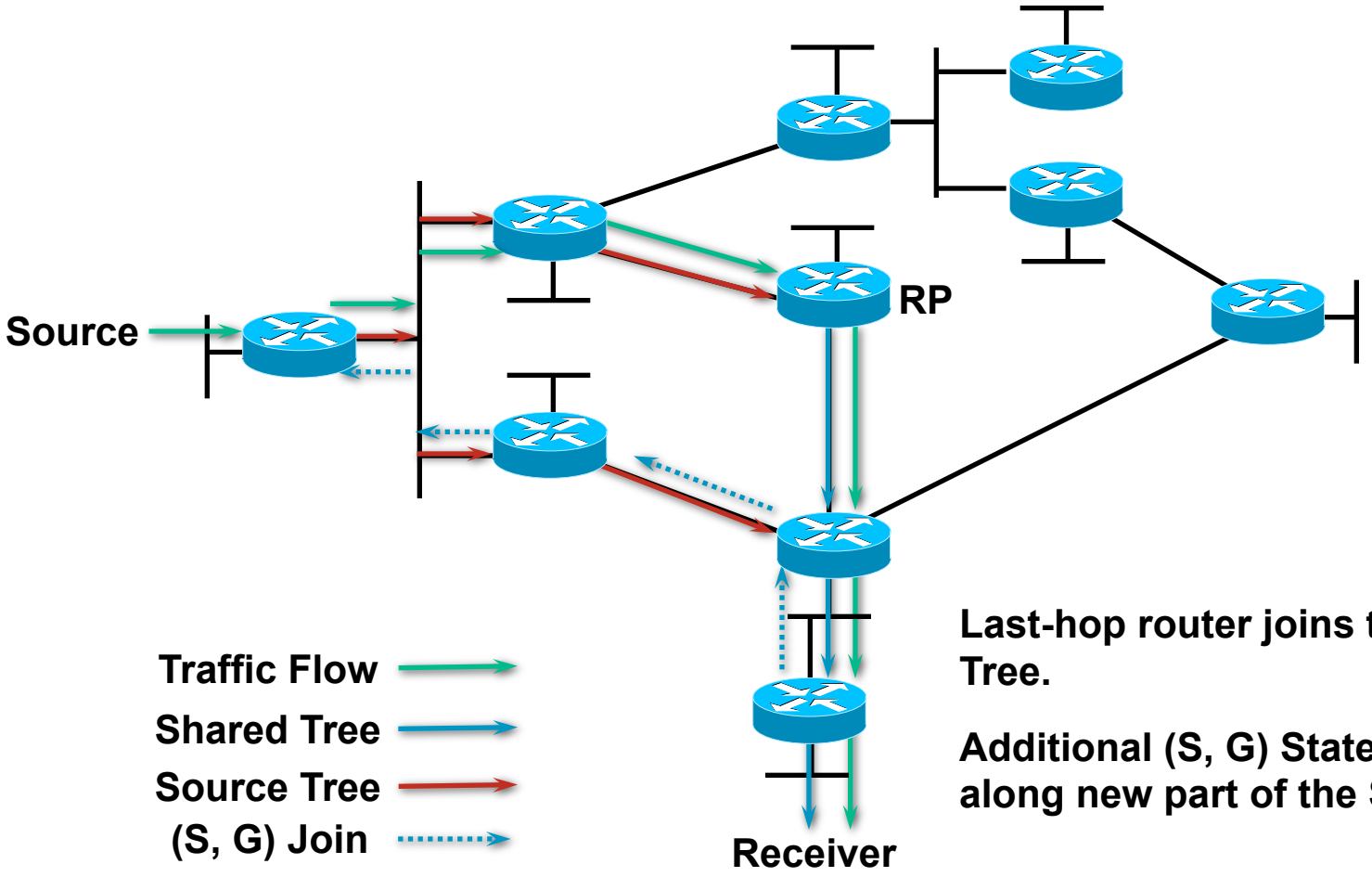
PIM-SM Sender Registration



Source traffic flows natively along SPT to RP.

From RP, traffic flows down the Shared Tree to Receivers.

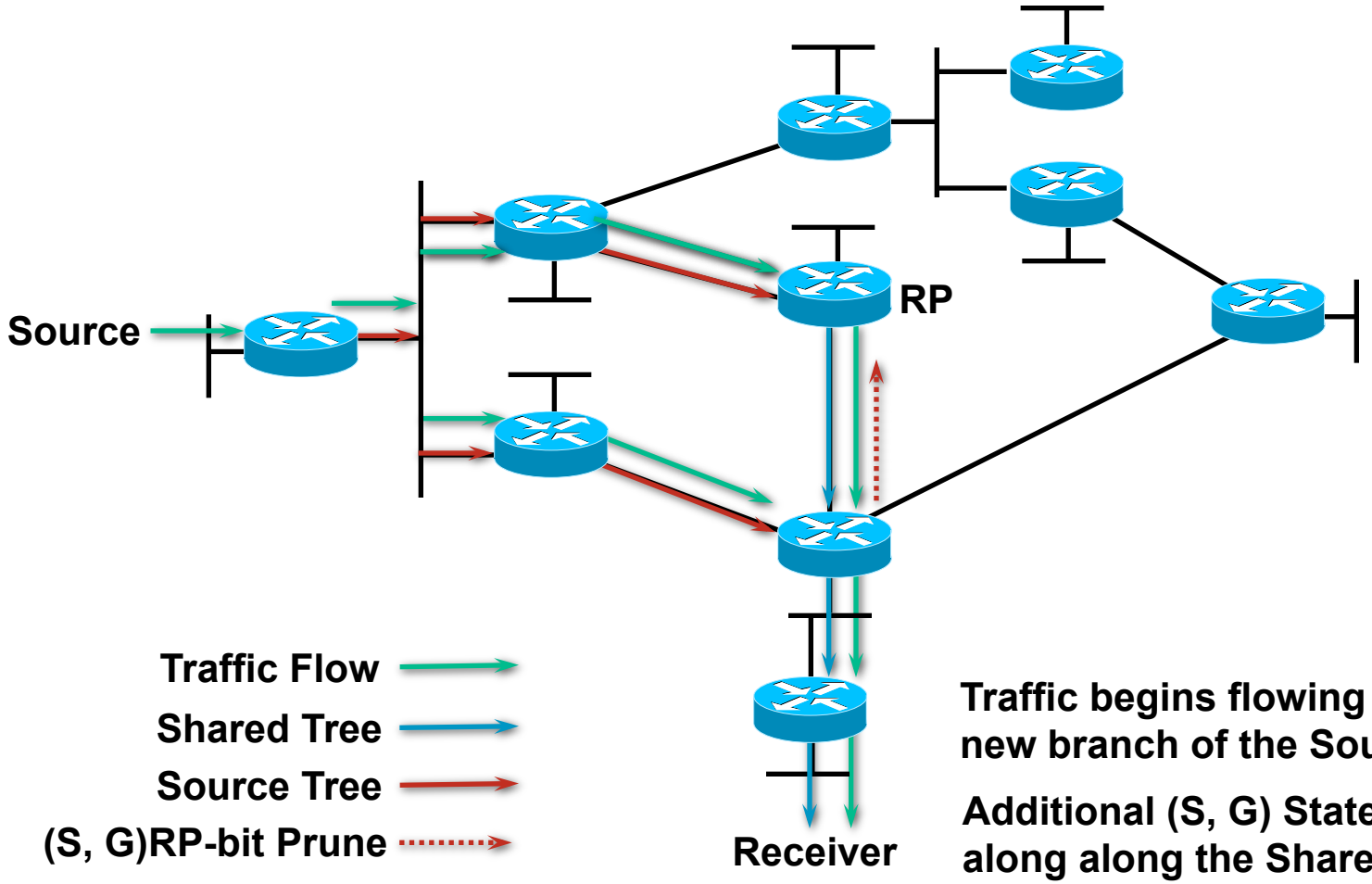
PIM-SM SPT Switchover



Last-hop router joins the Source Tree.

Additional (S, G) State is created along new part of the Source Tree.

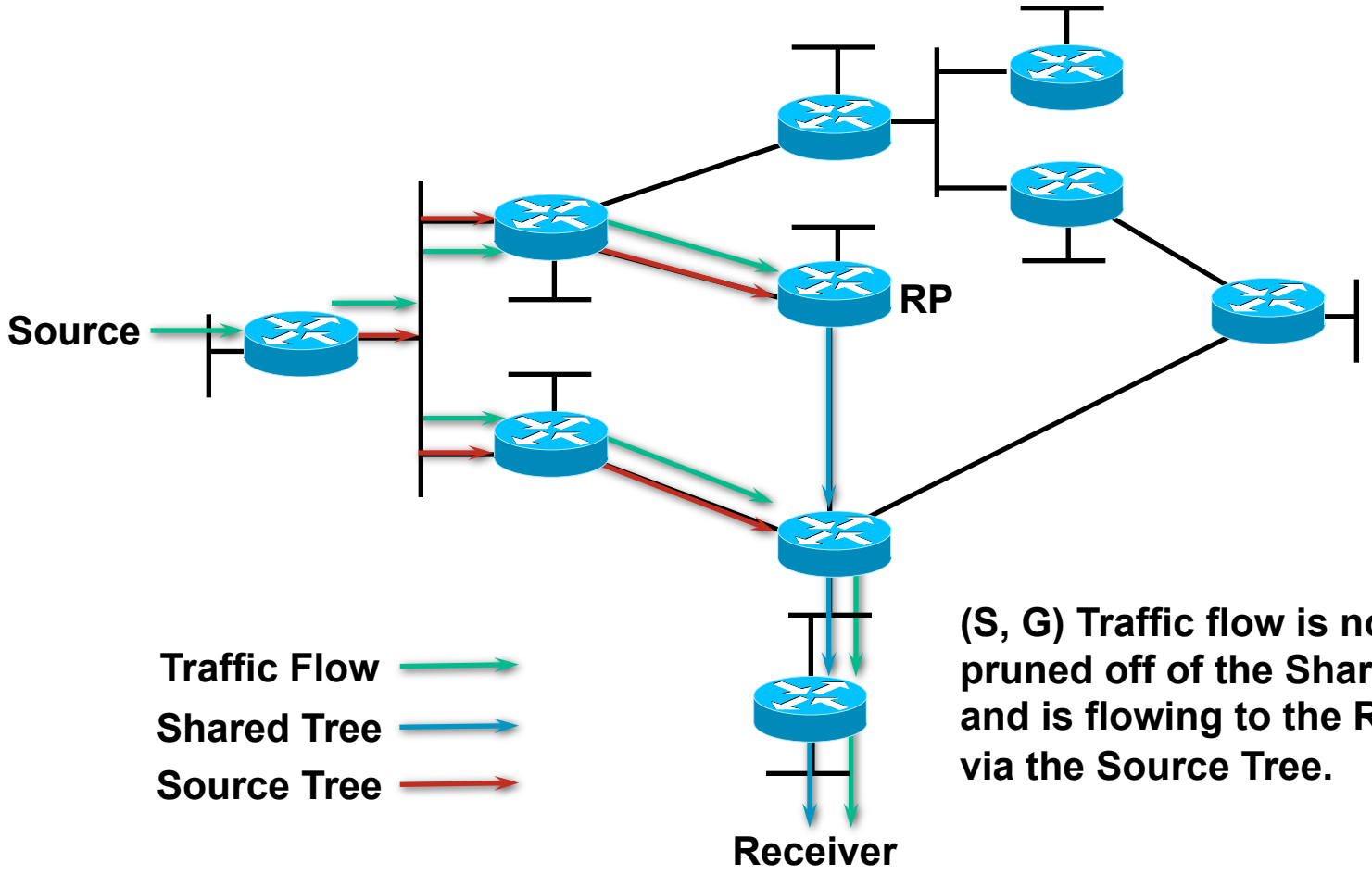
PIM-SM SPT Switchover



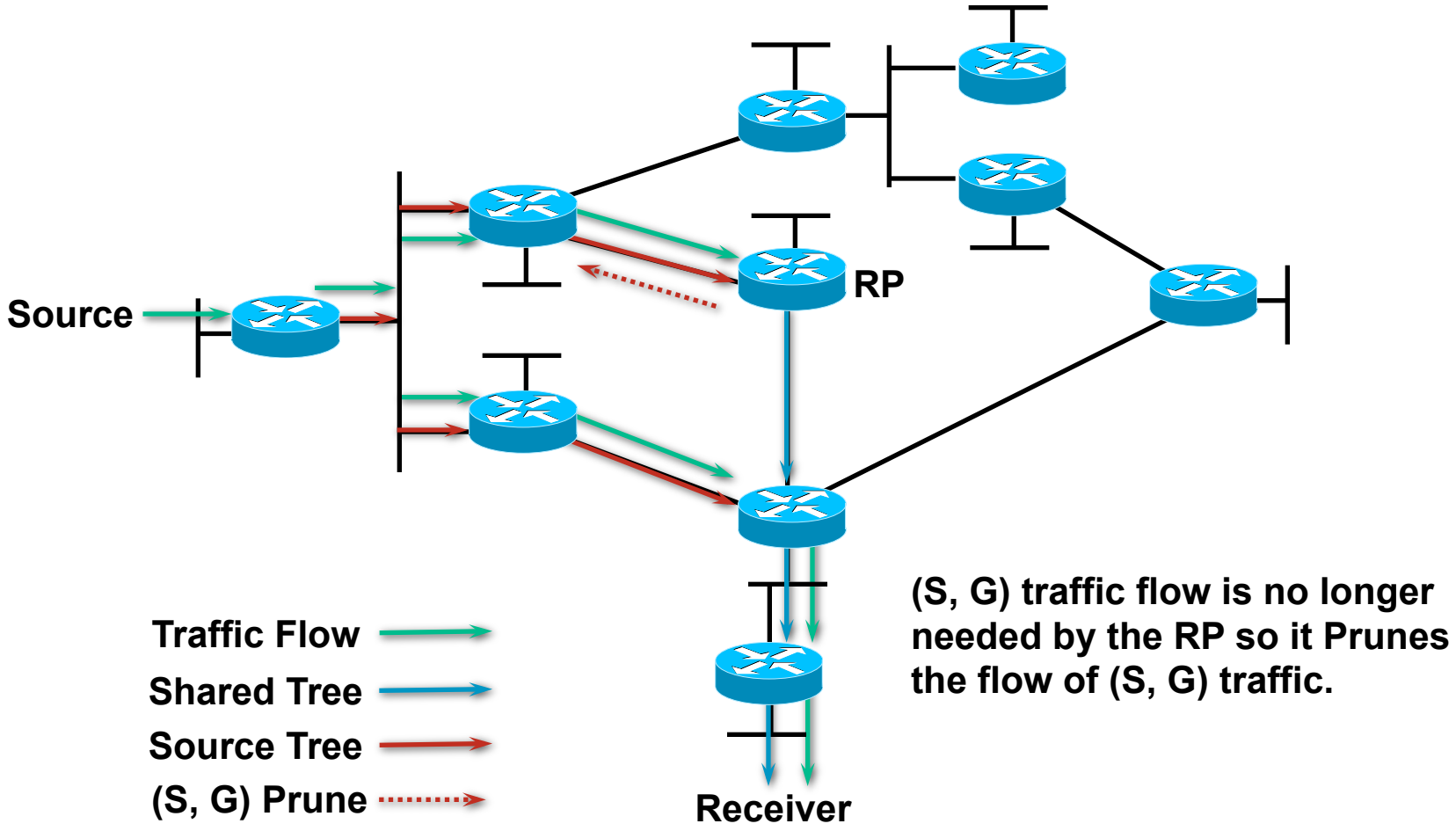
Traffic begins flowing down the new branch of the Source Tree.

Additional (S, G) State is created along along the Shared Tree to prune off (S, G) traffic.

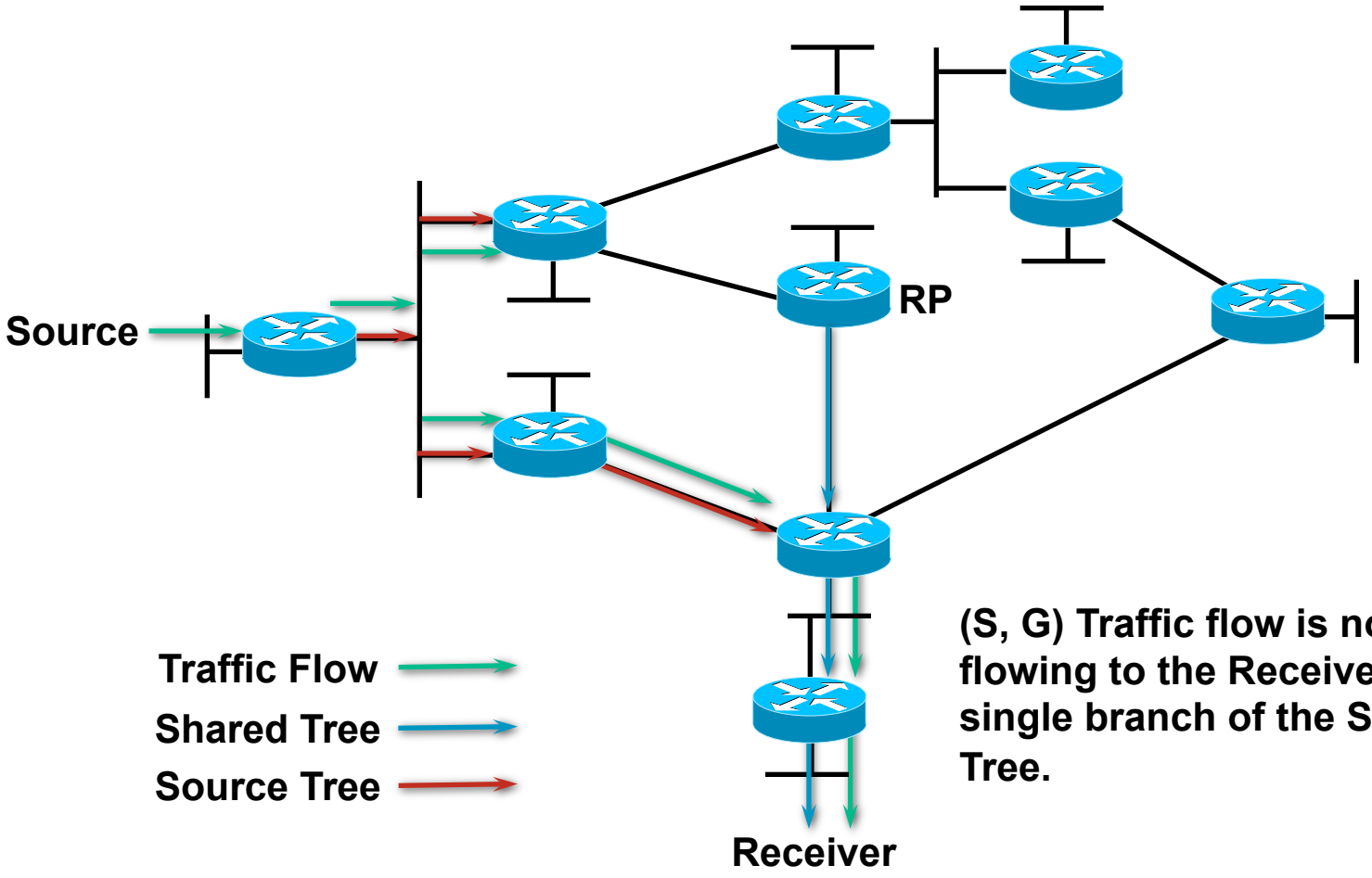
PIM-SM SPT Switchover



PIM-SM SPT Switchover



PIM-SM SPT Switchover



“ The default behavior of PIM-SM ASM is that routers with directly connected members will join the Shortest Path Tree as soon as they detect a new multicast source.”

PIM-SM Frequently Forgotten Fact

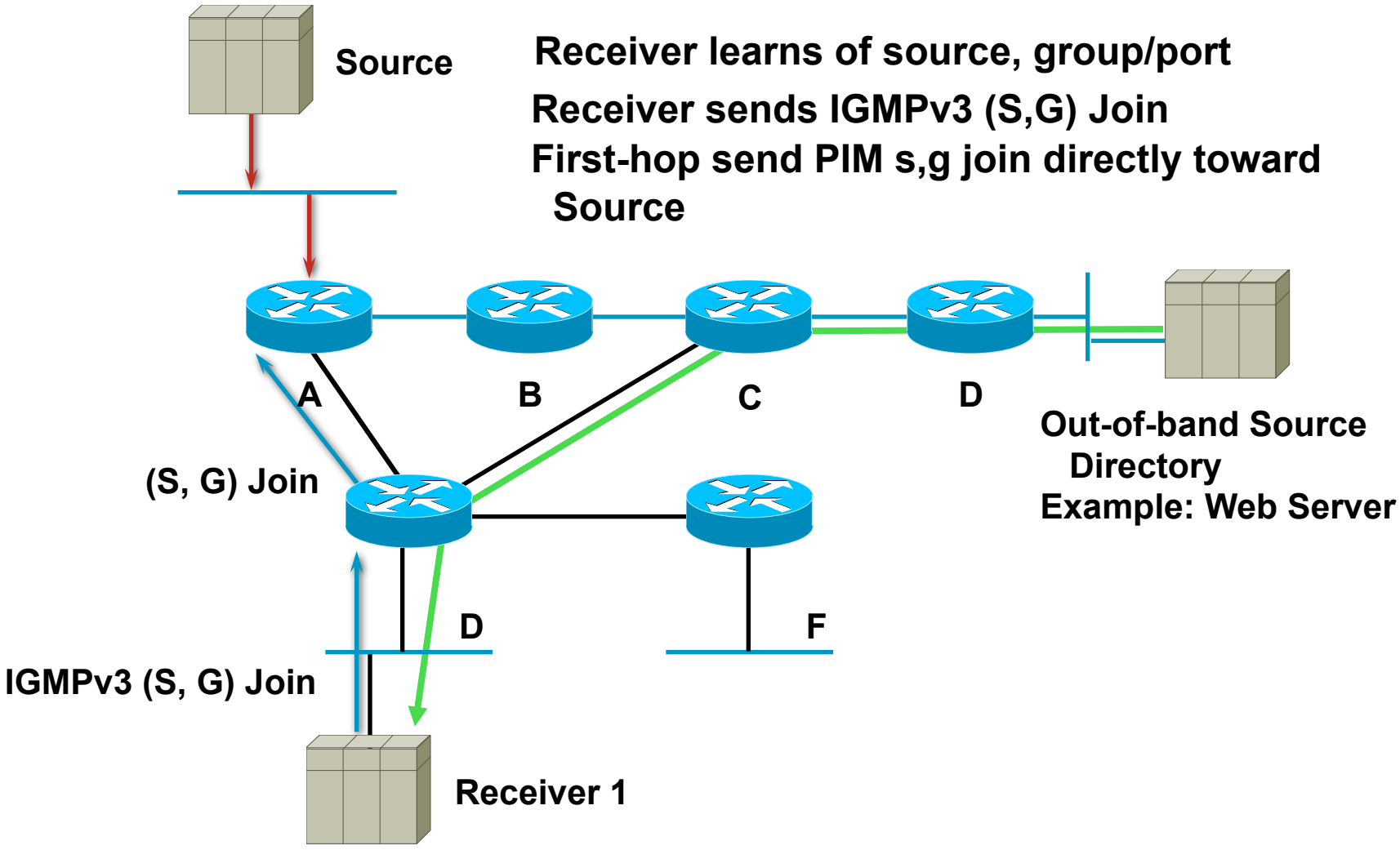
PIM-SM—Evaluation

- Effective for Sparse or “Dense” distribution of multicast receivers
- Advantages:
  - Traffic only sent down “joined” branches
  - Can switch to optimal source-trees for high traffic sources dynamically
  - Unicast routing protocol-independent
  - Basis for inter-domain multicast routing
    - When used with MBGP, MSDP and/or SSM

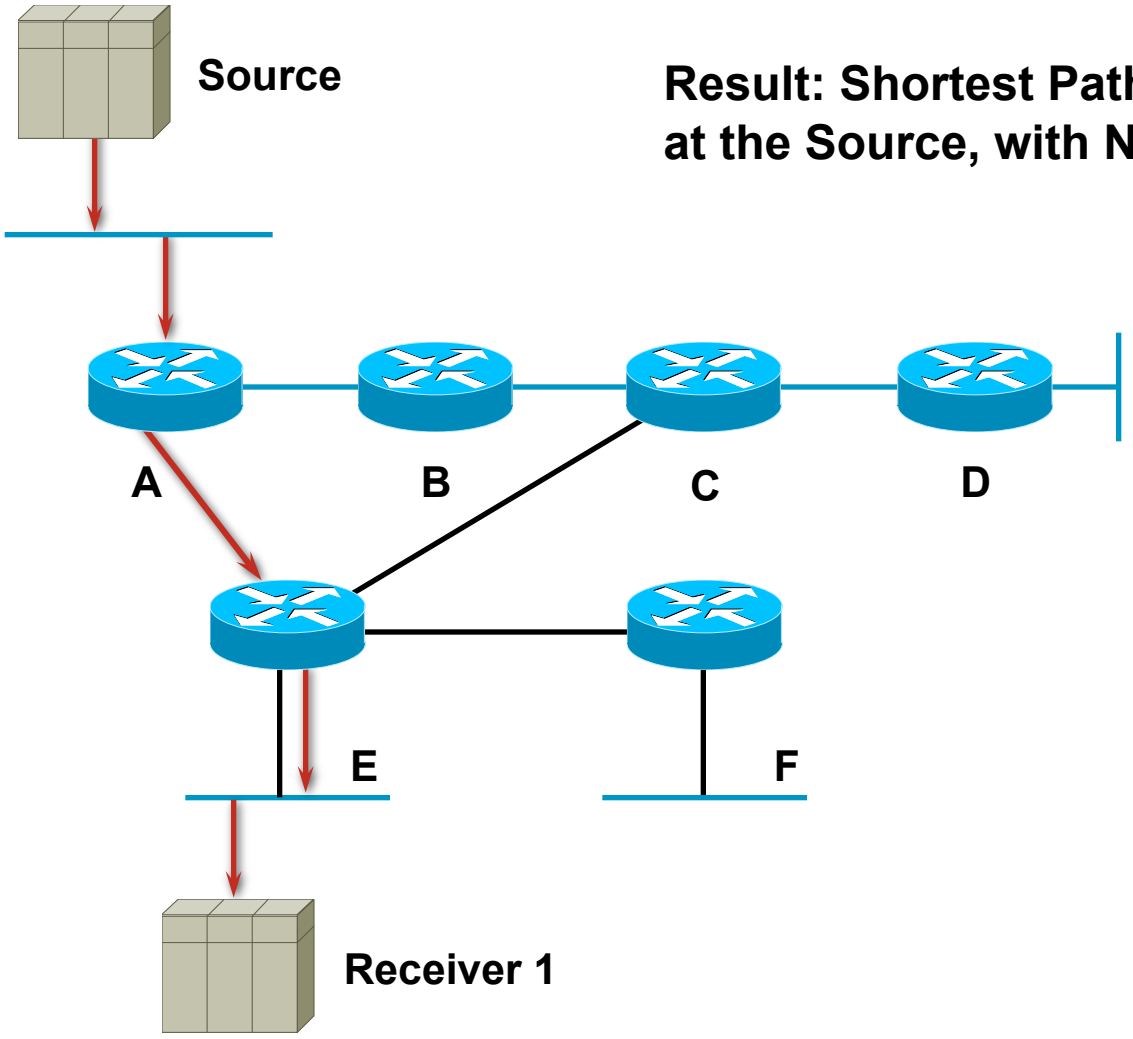
Source Specific Multicast

- Assume a One-to-Any Multicast Model.
  - Example: Video/Audio broadcasts, Stock Market data
- Why does ASM need a Shared Tree?
  - So that hosts and 1st hop routers can learn who the active source is for the group - Source Discovery
- What if this was already known?
  - Hosts could use IGMPv3 to signal exactly which (S,G) SPT to join.
  - The Shared Tree & RP wouldn't be necessary.
  - Different sources could share the same Group address and not interfere with each other.
- Result: Source Specific Multicast (SSM)
- RFC 3569 An Overview of Source-Specific Multicast (SSM)

PIM Source Specific Mode



PIM Source Specific Mode



Result: Shortest Path Tree rooted at the Source, with NO Shared Tree.

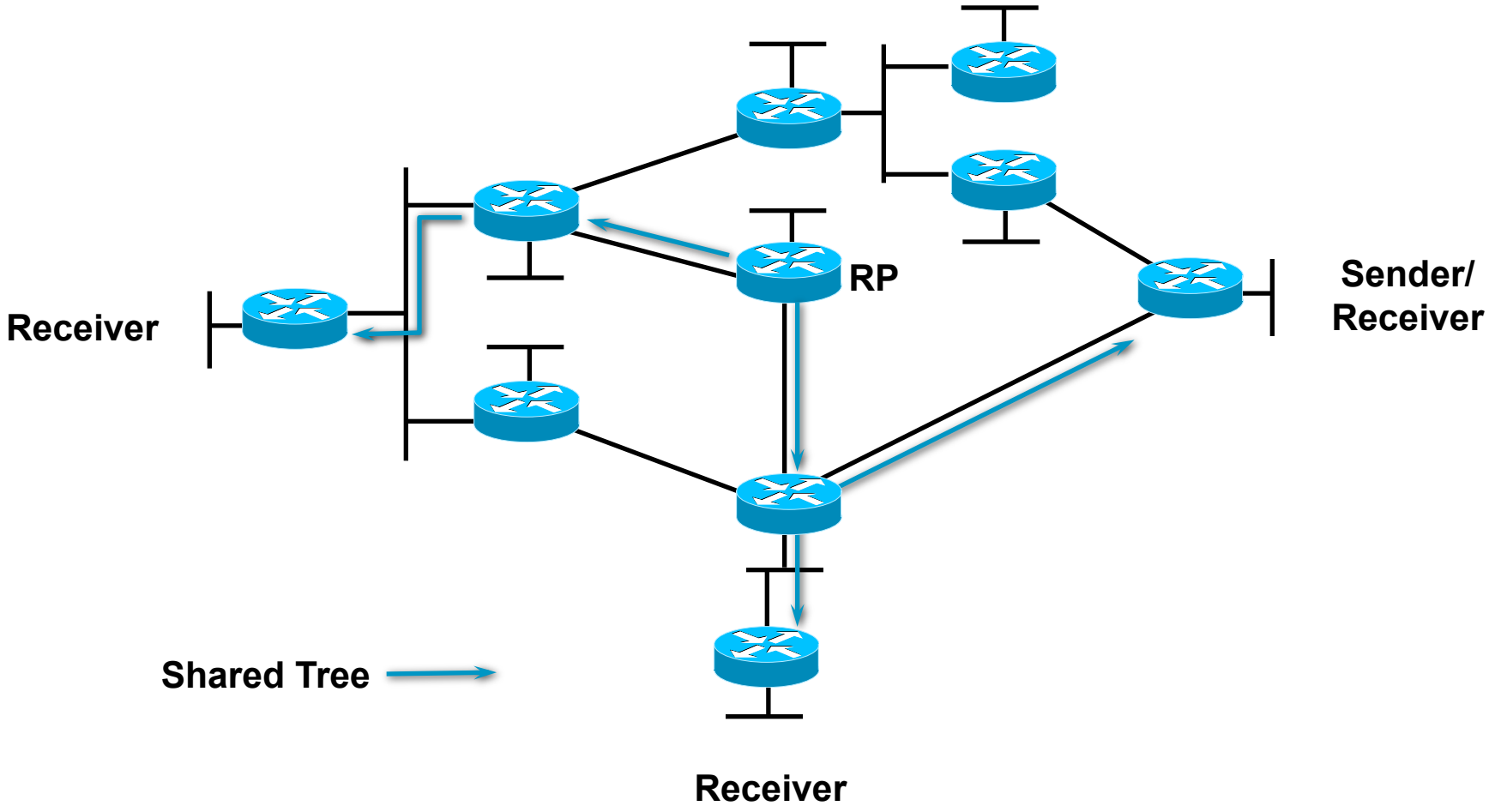
SSM - Evaluation

- Ideal for applications with one source sending to many receivers
- Uses a simplified subset of the PIM-SM protocol
  - Simpler network operation
- Solves multicast address allocation problems.
  - Flows differentiated by both source and group.
    - Not just by group.
  - Content providers can use same group ranges.
    - Since each (S,G) flow is unique.
- Helps prevent certain DoS attacks
  - “Bogus” source traffic:
    - Can’t consume network bandwidth.
    - Not received by host application.

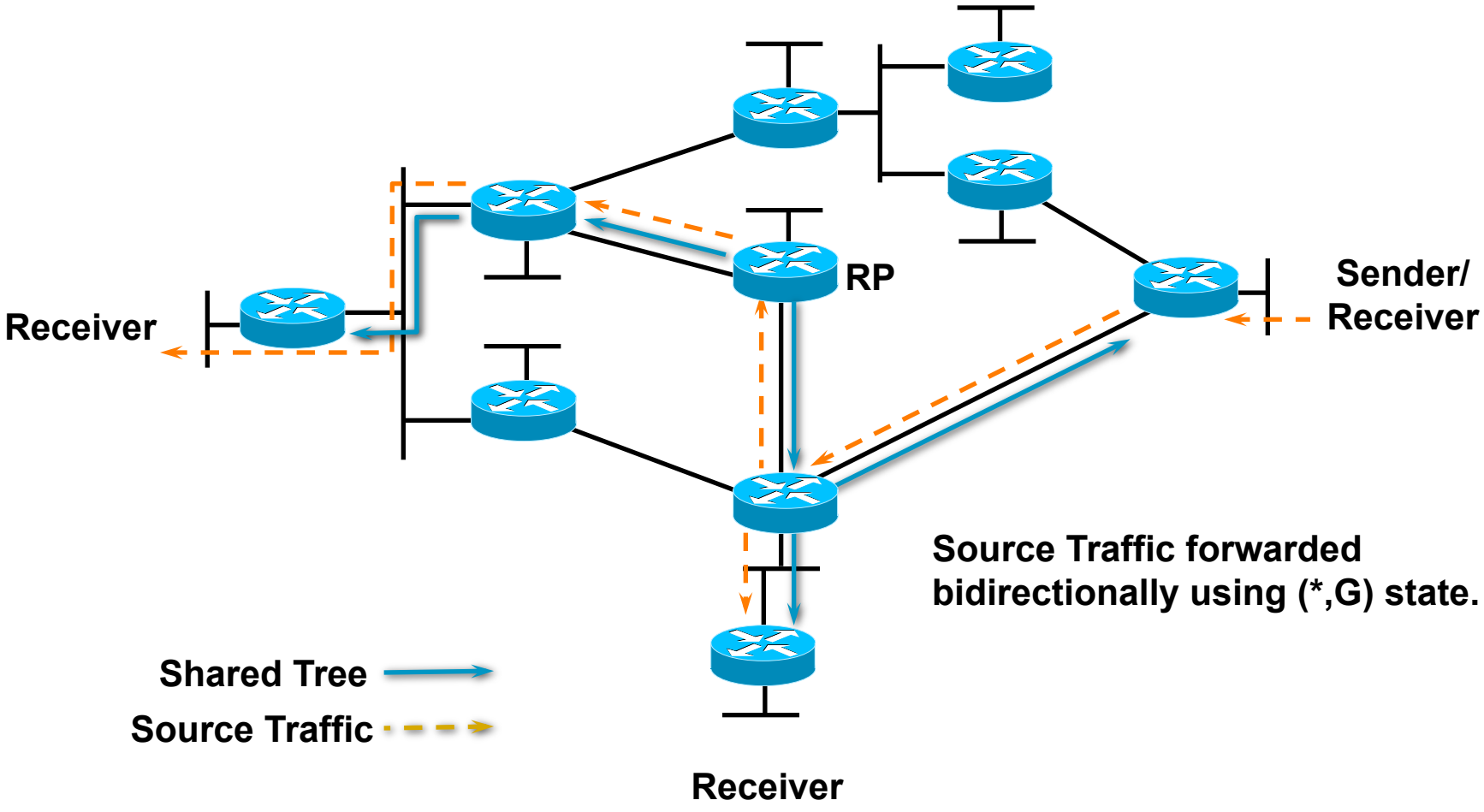
Many-to-Many State Problem

- Creates huge amounts of (S,G) state
  - State maintenance workloads skyrocket
    - High OIL fan-out makes the problem worse
  - Router performance begins to suffer
- Using Shared-Trees only.
  - Provides some (S,G) state reduction
    - Results in (S,G) state only along SPT to RP
    - Frequently still too much (S,G) state
    - Need a solution that only uses (*,G) state

Bidirectional PIM—Overview



Bidirectional PIM—Overview



Bidir PIM–Evaluation

- Ideal for Many to Many applications
- Drastically reduces network mroute state.
  - Eliminates ALL (S,G) state in the network.
    - SPT's between sources to RP eliminated.
    - Source traffic flows both up and down Shared Tree.
  - Allows Many-to-Any applications to scale.
    - Permits virtually an unlimited number of sources.

RP Choices

PIM-SM RP Requirements

- Group to RP mapping
  - Consistent in all routers within the PIM domain
- RP redundancy requirements
  - Elmiminate any single point of failure

How does the network know about the RP ?

- Static configuration
  - Manually on every router in the PIM domain
- AutoRP
  - Originally a cisco solution
  - Facilitated PIM-SM early transition
- BSR
  - draft-ietf-pim-sm-bsr

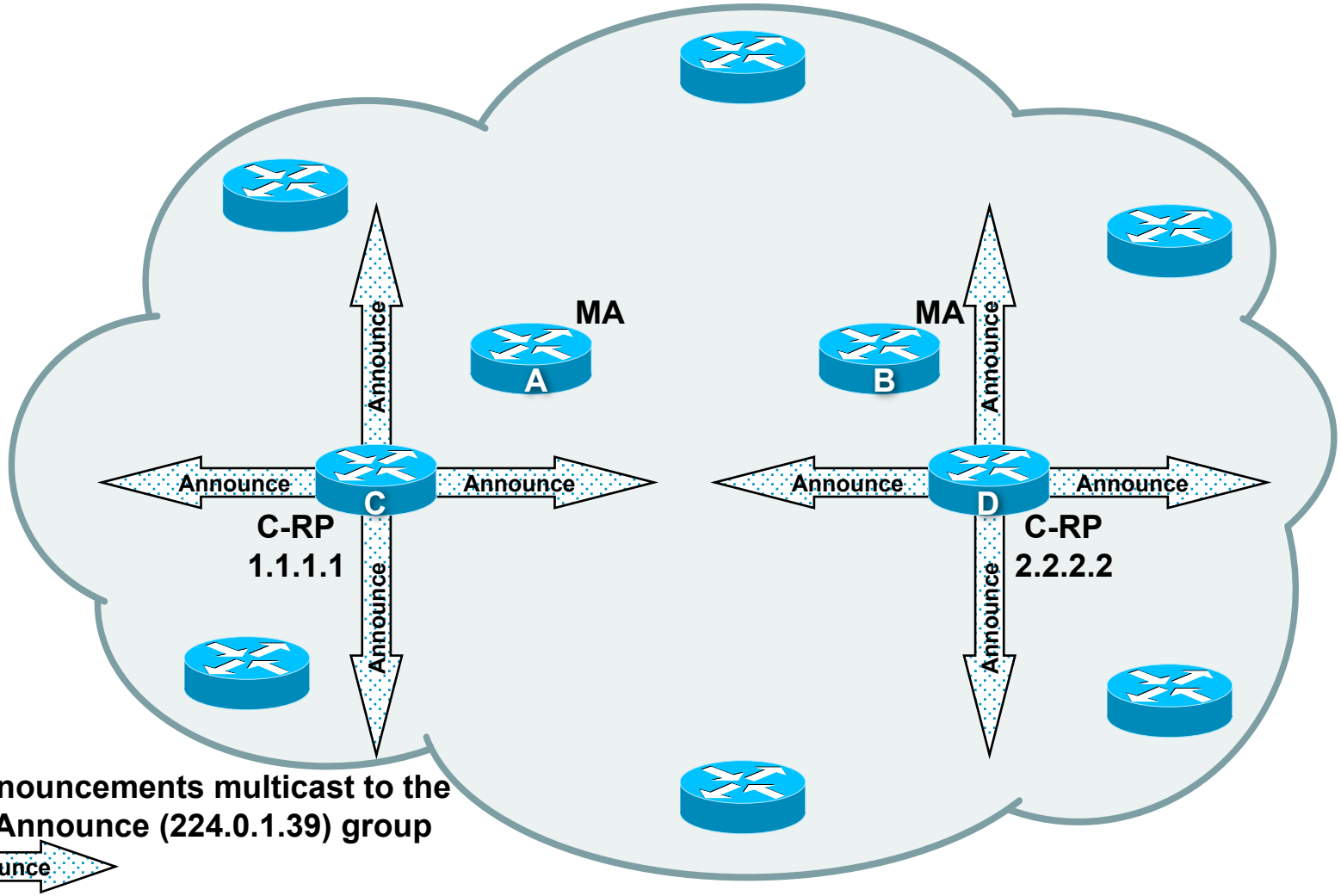
Static RP's

- Hard-configured RP address
  - When used, must be configured on every router
  - All routers must have the same RP address
  - RP fail-over not possible
    - Exception: If Anycast RPs are used.
- Command

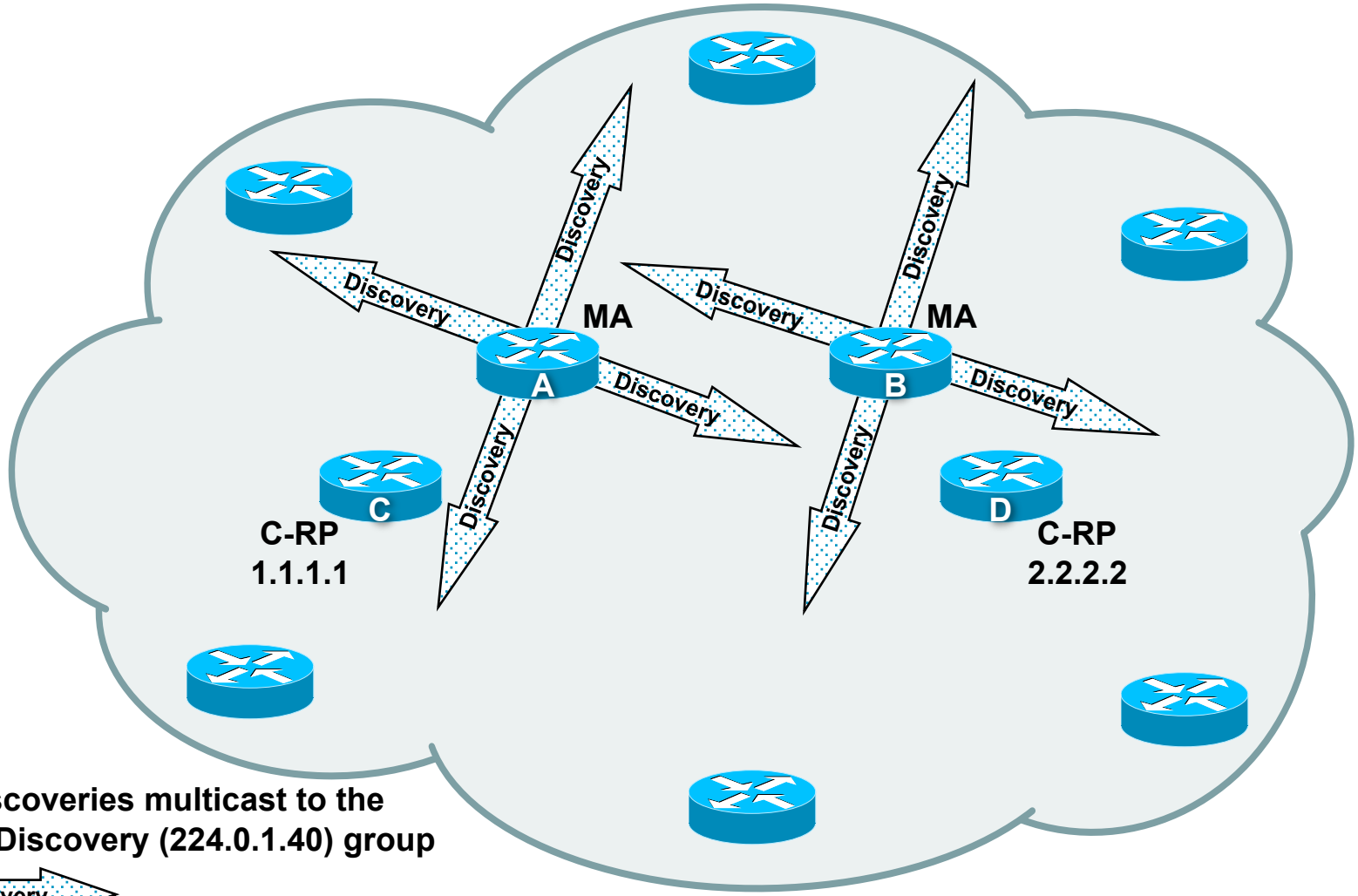
```
ip pim rp-address <address> [group-list <acl>] [override]
```

  - Optional group list specifies group range
    - Default: Range = 224.0.0.0/4 (Includes Auto-RP Groups!!!!)
  - Override keyword "overrides" Auto-RP information
    - Default: Auto-RP learned info takes precedence

Auto-RP—From 10,000 Feet



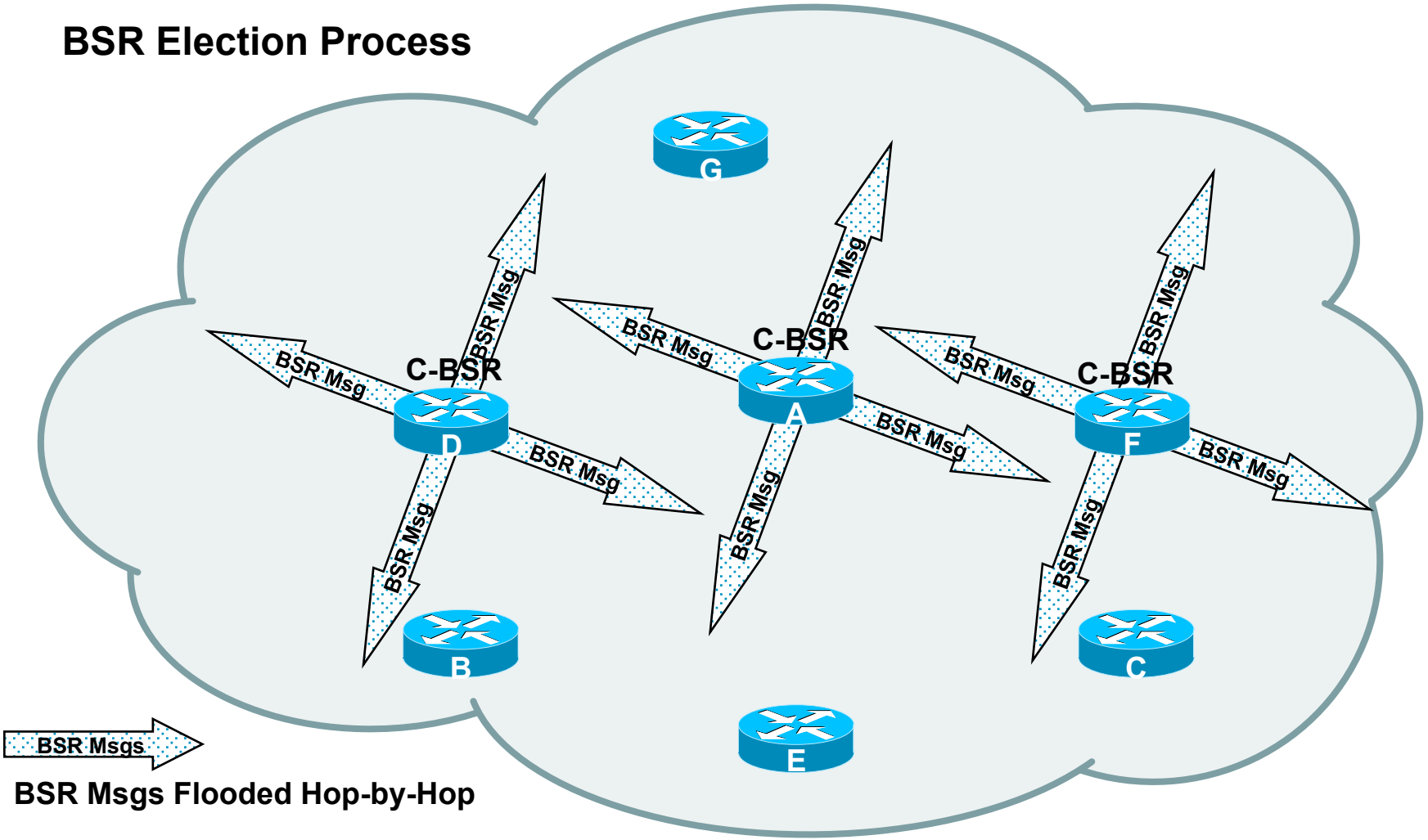
Auto-RP—From 10,000 Feet



RP-Discoveries multicast to the Cisco Discovery (224.0.1.40) group

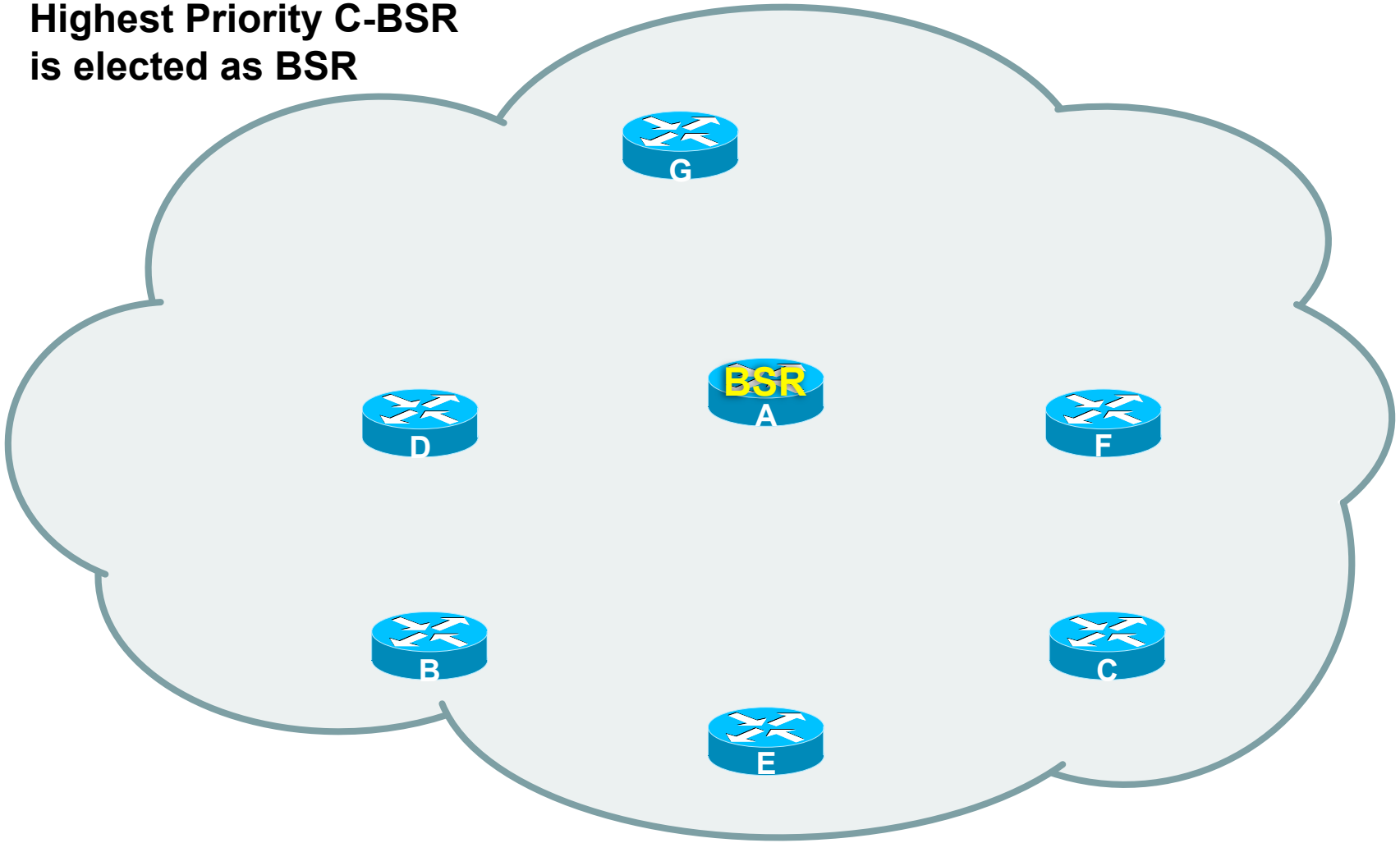
Discovery

BSR – From 10,000 Feet

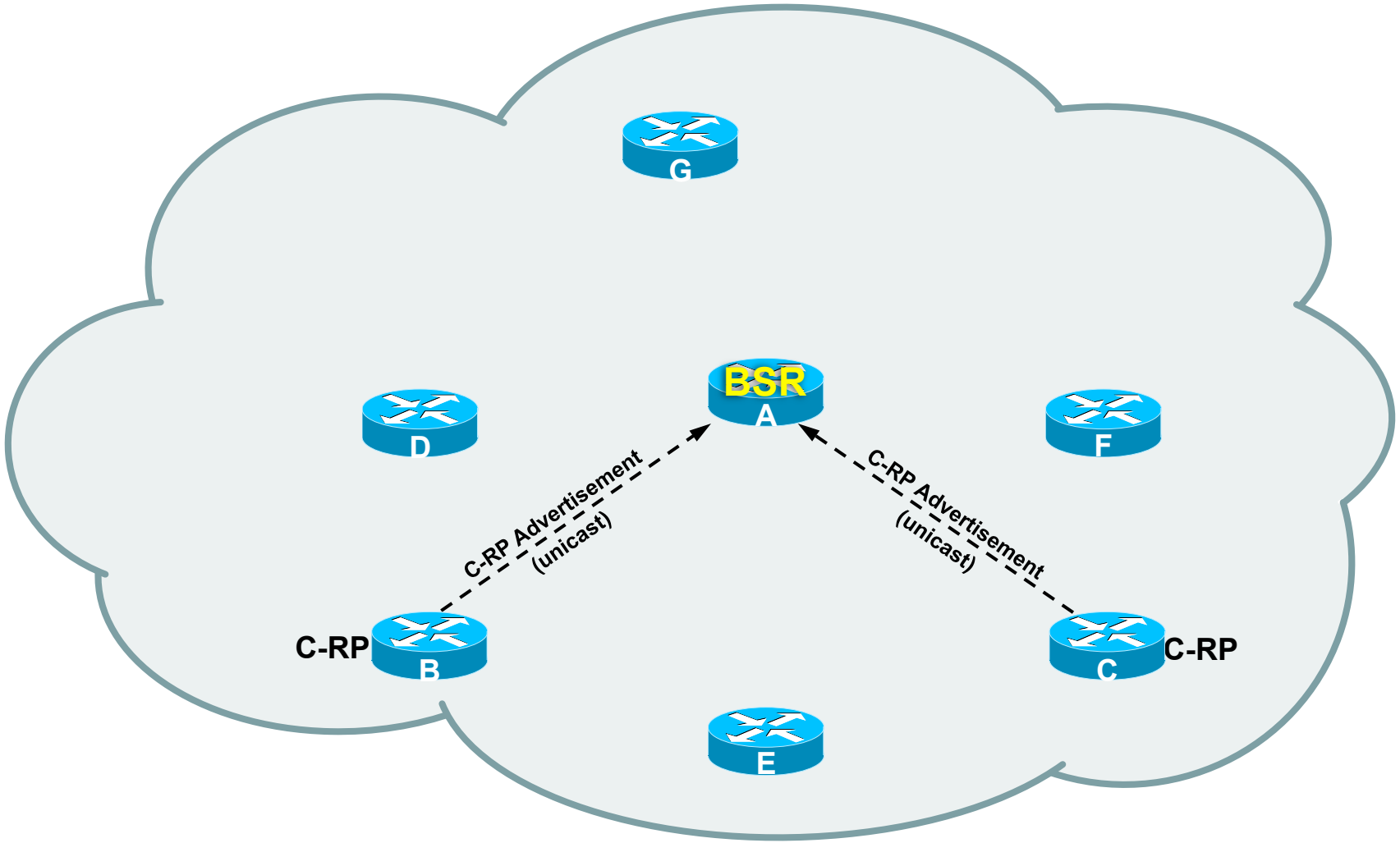


BSR – From 10,000 Feet

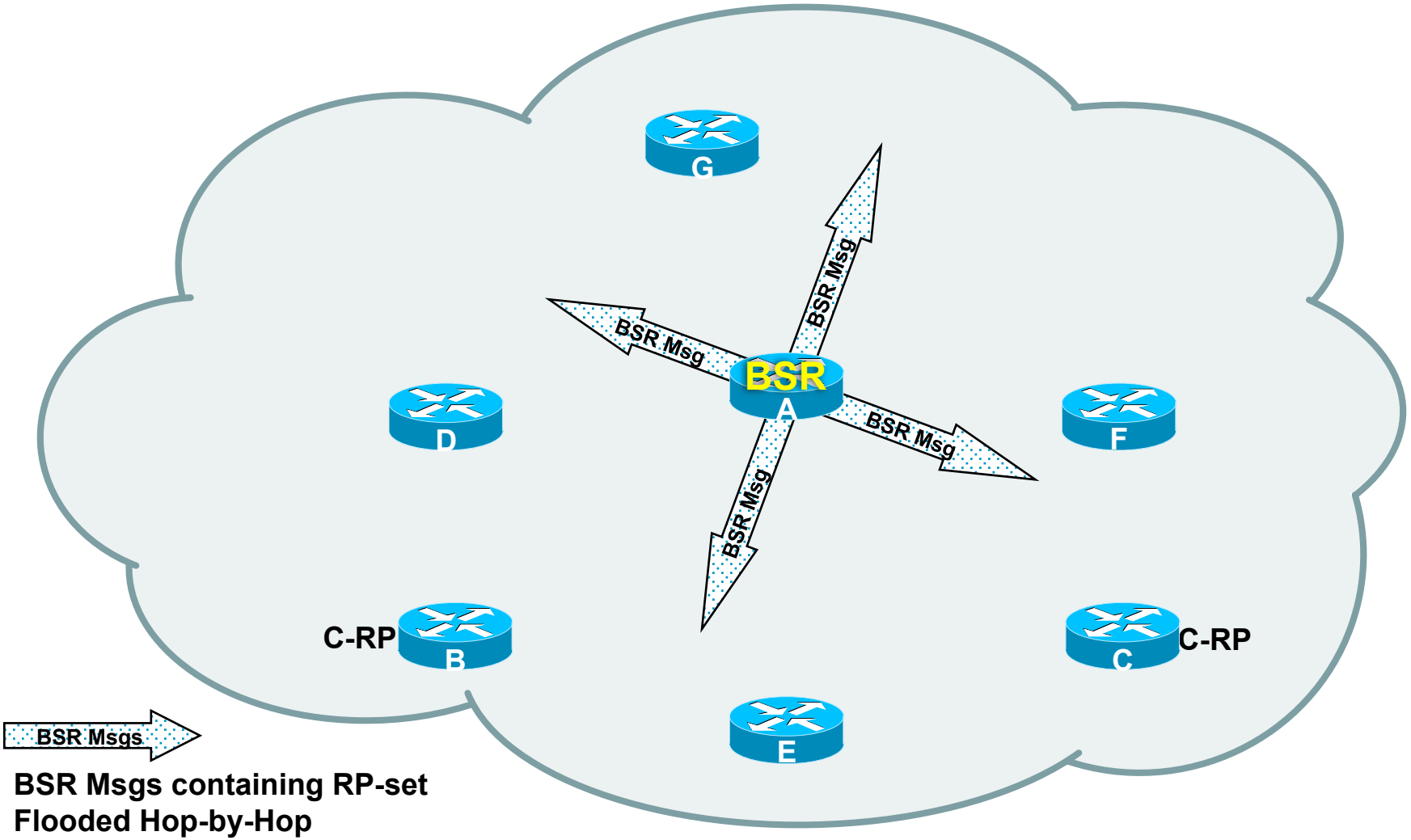
Highest Priority C-BSR is elected as BSR



BSR – From 10,000 Feet



BSR – From 10,000 Feet



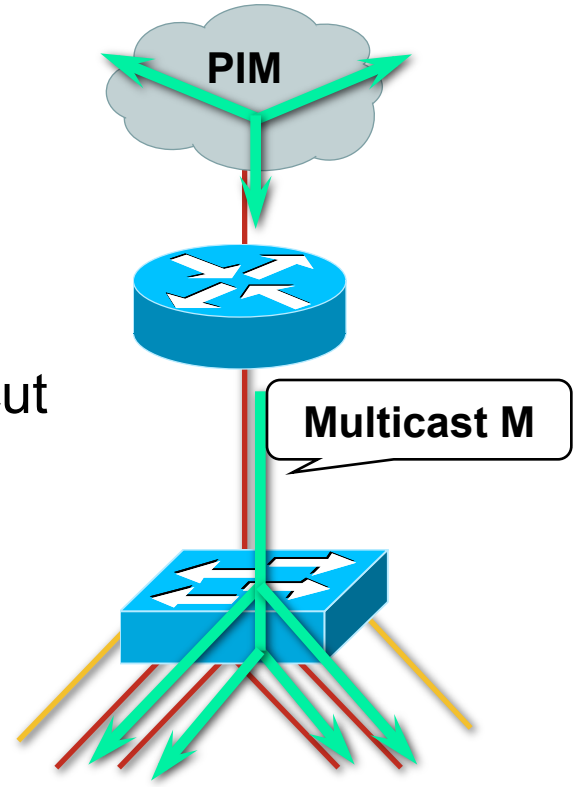
BSR Msgs containing RP-set
Flooded Hop-by-Hop

Multicast at Layer2

L2 Multicast Frame Switching

Problem: Layer 2 Flooding of Multicast Frames

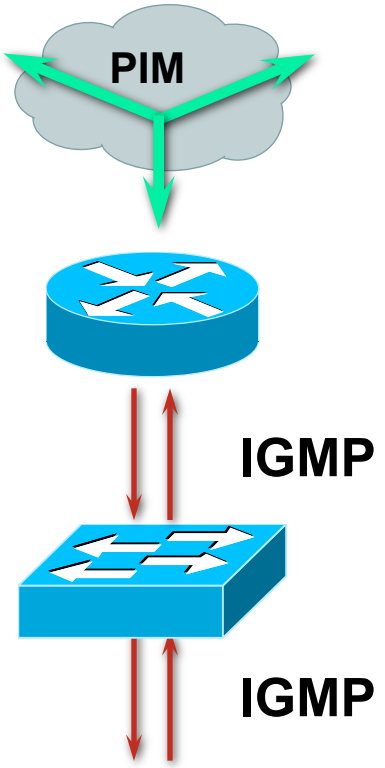
- Typical L2 switches treat multicast traffic as unknown or broadcast and must “flood” the frame to every port.
- Static entries can sometimes be set to specify which ports should receive which group(s) of multicast traffic.
- Dynamic configuration of these entries would cut down on user administration.



L2 Multicast Frame Switching

IGMPv1-v2 Snooping

- Switches become “IGMP” aware
- IGMP packets intercepted by the NMP or by special hardware ASICs
  - Requires special hardware to maintain throughput
- Switch must examine contents of IGMP messages to determine which ports want what traffic
  - IGMP membership reports
  - IGMP leave messages
- Impact on low-end Layer-2 switches:
  - Must process ALL Layer 2 multicast packets
  - Admin. load increases with multicast traffic load
  - Generally results in switch *Meltdown* !!!



L2 Multicast Frame Switching

- Impact of IGMPv3 on IGMP Snooping
  - IGMPv3 Reports sent to separate group (224.0.0.22)
    - Switches listen to just this group.
    - Only IGMP traffic – no data traffic.
    - Substantially reduces load on switch CPU.
    - Permits low-end switches to implement IGMPv3 Snooping
  - No Report Suppression in IGMPv3
    - Enables individual member tracking
  - IGMPv3 supports Source-specific Includes/ Excludes

Summary—Frame Switches

- IGMP snooping
  - Switches with Layer 3 aware Hardware/ASICs
    - High-throughput performance maintained
    - Increases cost of switches
  - Switches without Layer 3 aware Hardware/ASICs
    - Suffer serious performance degradation or even ***Meltdown!***
      - Shouldn't be a problem when IGMPv3 is implemented

Interdomain IP Multicast

MBGP Overview

- MBGP: Multiprotocol BGP
  - Defined in RFC 2858 (extensions to BGP)
  - Can carry different types of routes
    - Unicast
    - Multicast
  - Both routes carried in same BGP session
  - Does not propagate multicast state info
    - That's PIMs job
  - Same path selection and validation rules
    - AS-Path, LocalPref, MED, …

MBGP Overview

- Separate BGP tables maintained
  - Unicast prefixes for unicast forwarding
  - Unicast prefixes for multicast RPF checking
- AFI = 1, Sub-AFI = 1
  - Contains unicast prefixes for unicast forwarding
  - Populated with BGP unicast NLRI
- AFI = 1, Sub-AFI = 2
  - Contains unicast prefixes for RPF checking
  - Populated with BGP multicast NLRI

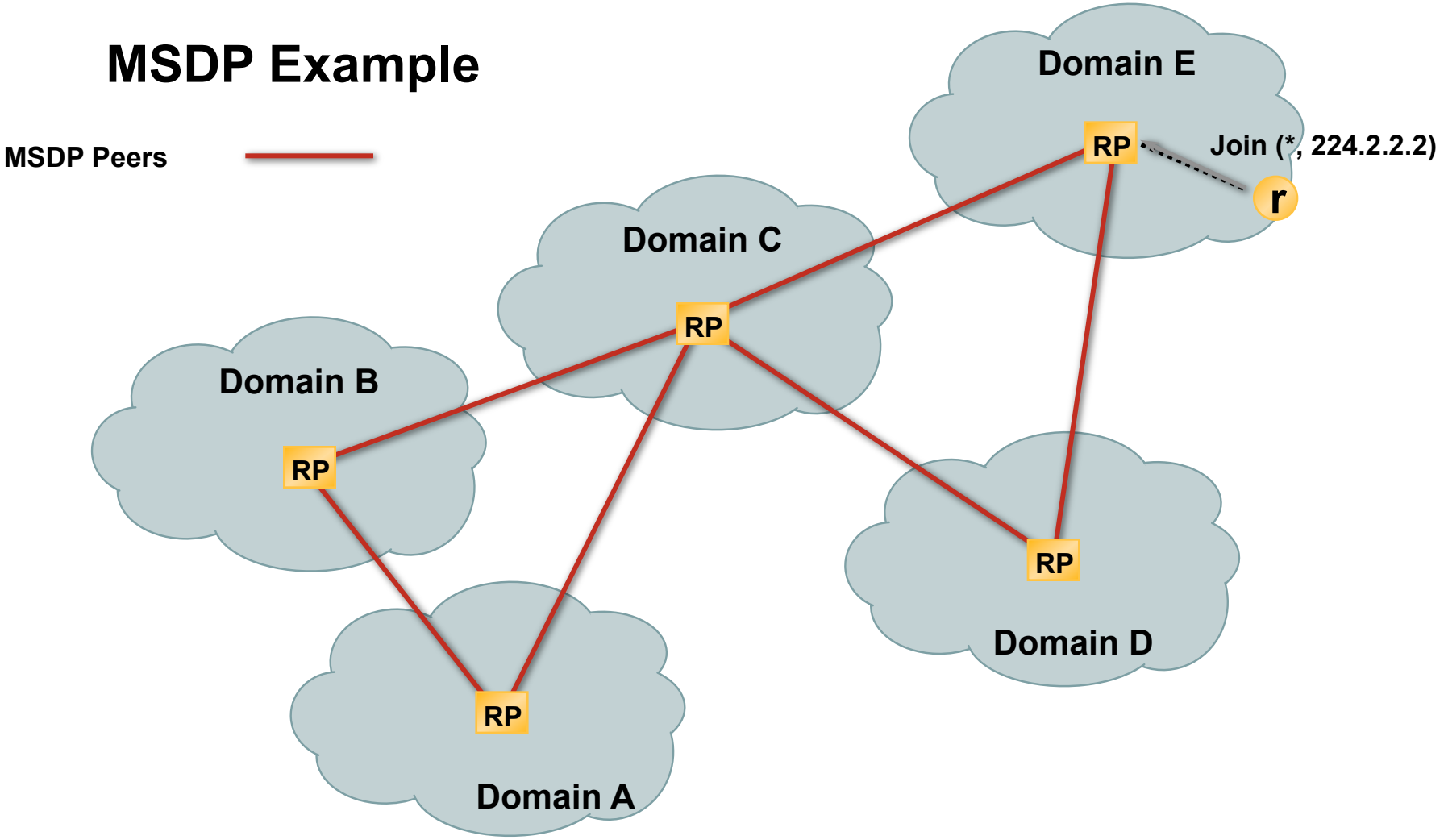
MBGP Overview

- MBGP allows divergent paths and policies
  - Same IP address holds dual significance
    - Unicast routing information
    - Multicast RPF information
  - For same IPv4 address two different NLRI with different next-hops
  - Can therefore support both congruent and incongruent topologies

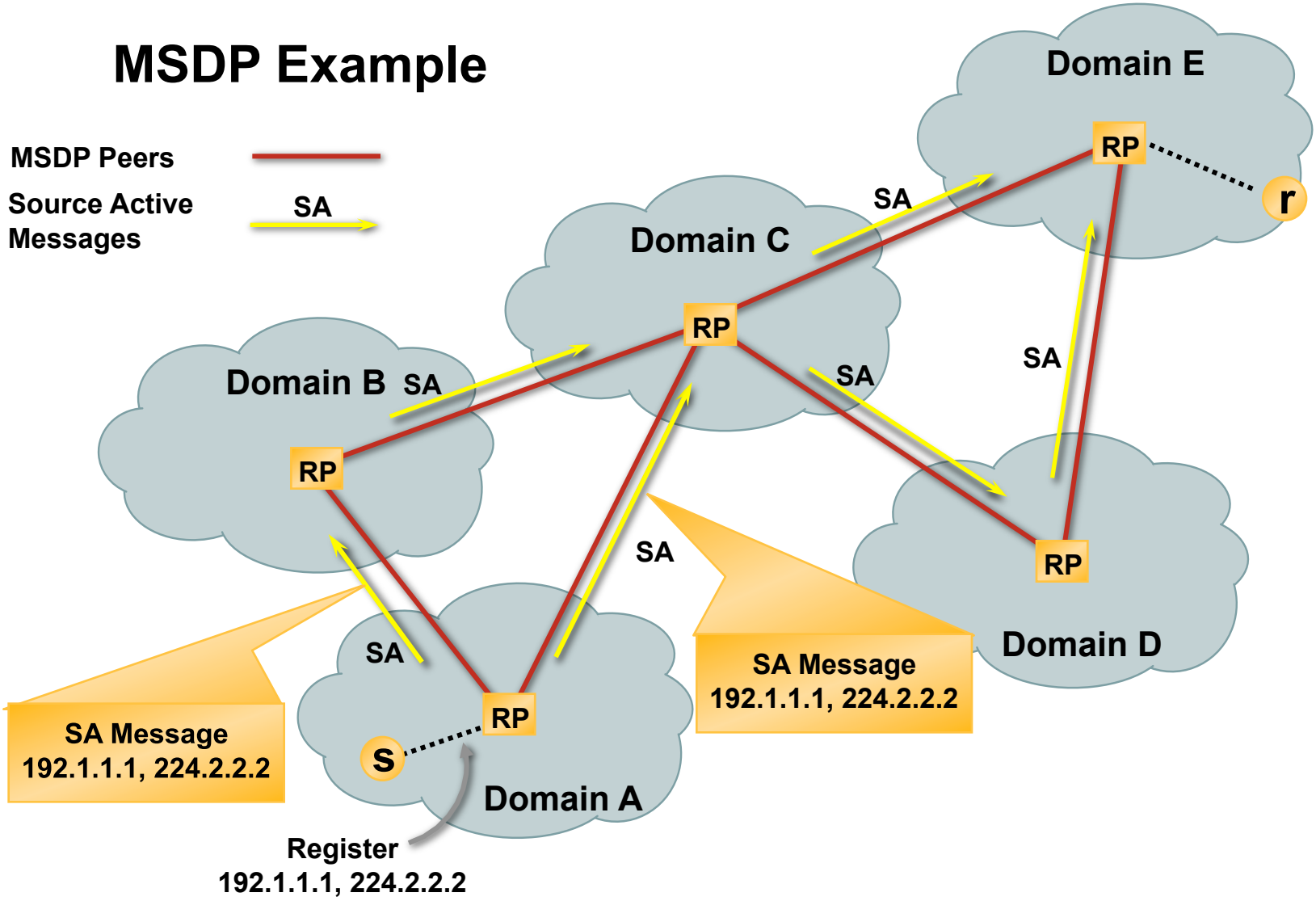
MSDP

- RFC 3618
- PIM-SM ASM only
  - RP's knows about all sources in their domain
    - Sources cause a "PIM Register" to the RP
    - Tell RP's in other domains of it's sources
      - Via MSDP SA (Source Active) messages
  - RP's know about receivers in a domain
    - Receivers cause a "(*, G) Join" to the RP
    - RP can join the source tree in the peer domain
      - Via normal PIM (S, G) joins
  - MSDP required for interdomain ASM source discovery

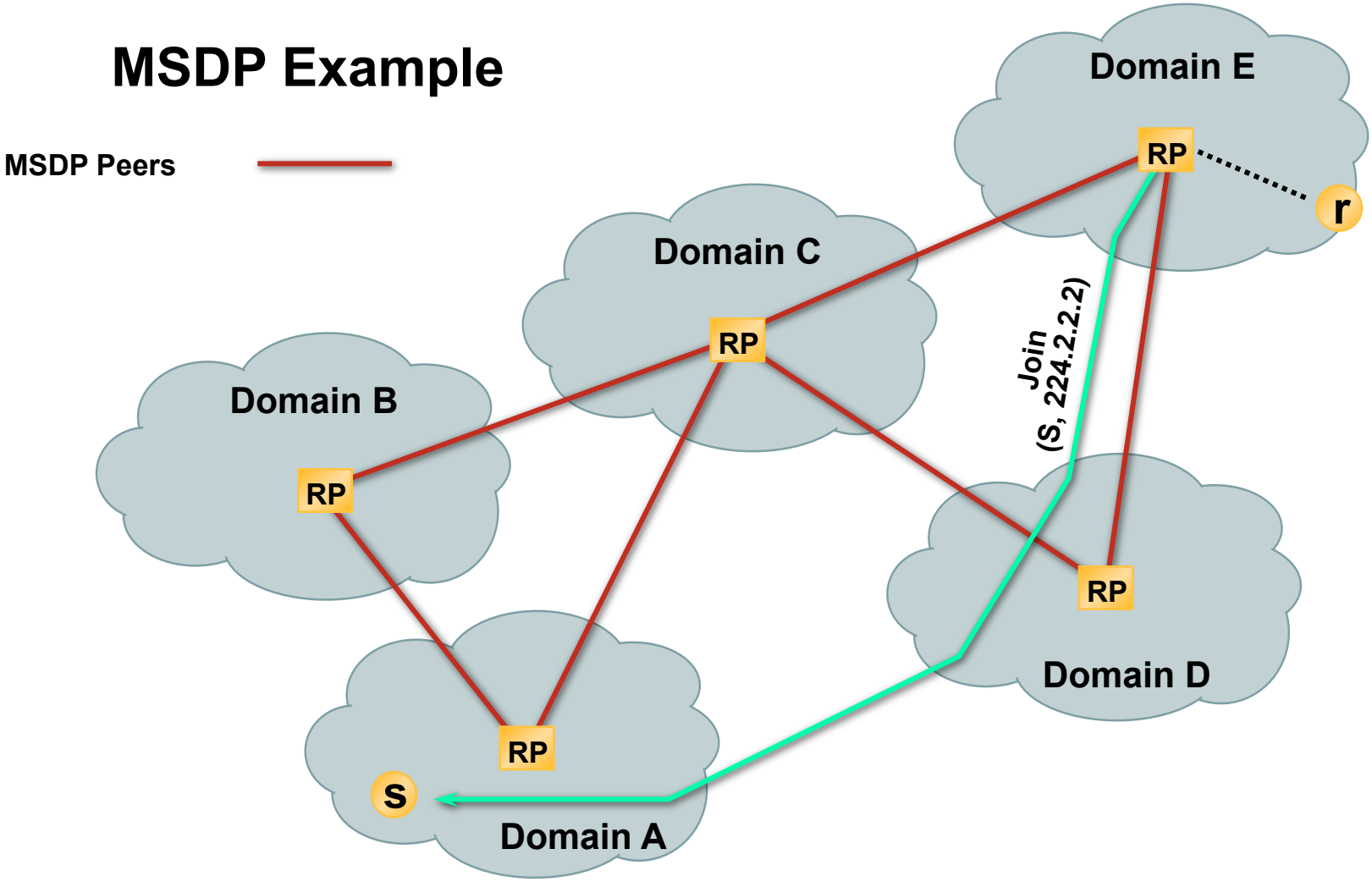
MSDP Overview



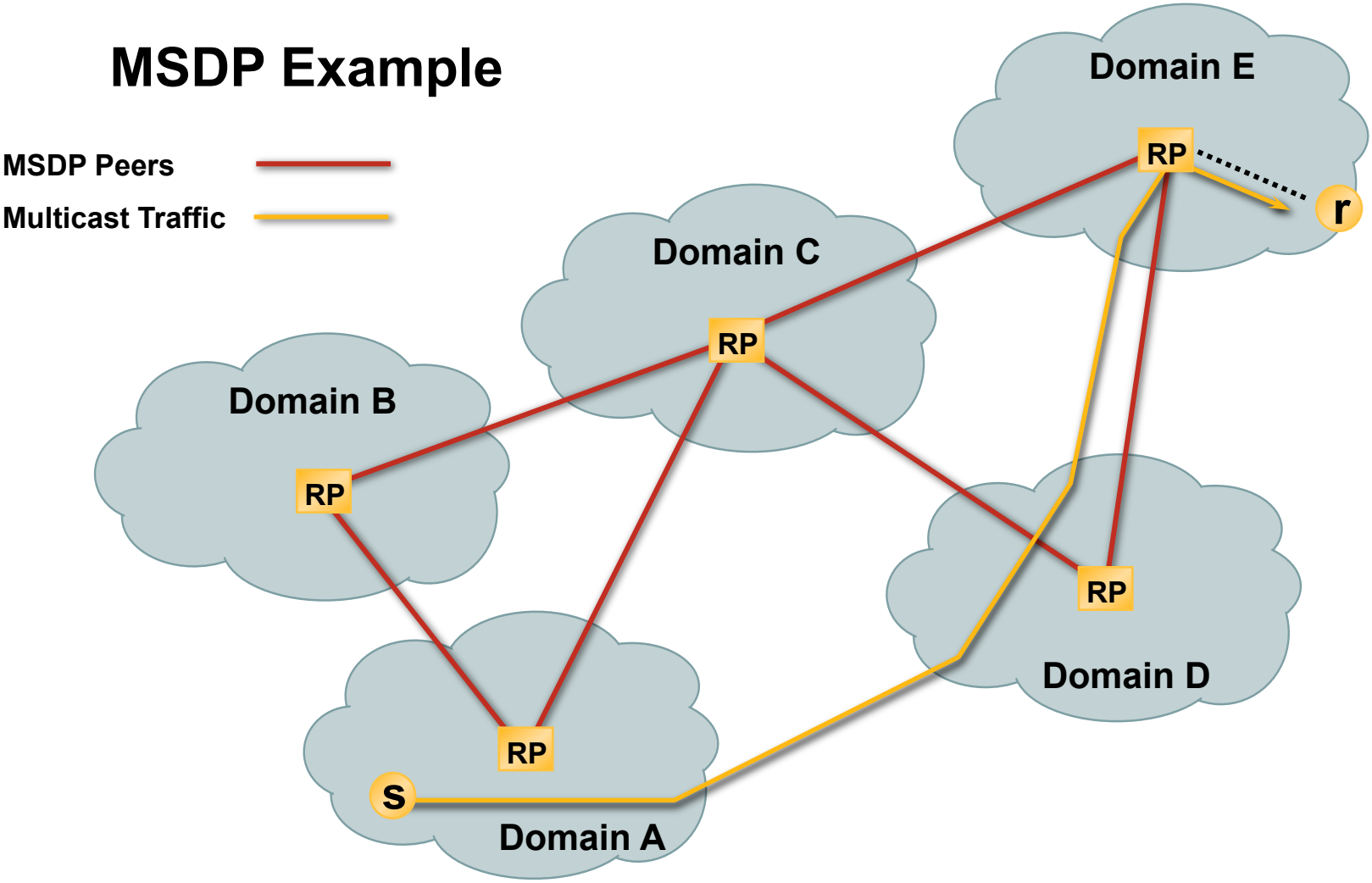
MSDP Overview



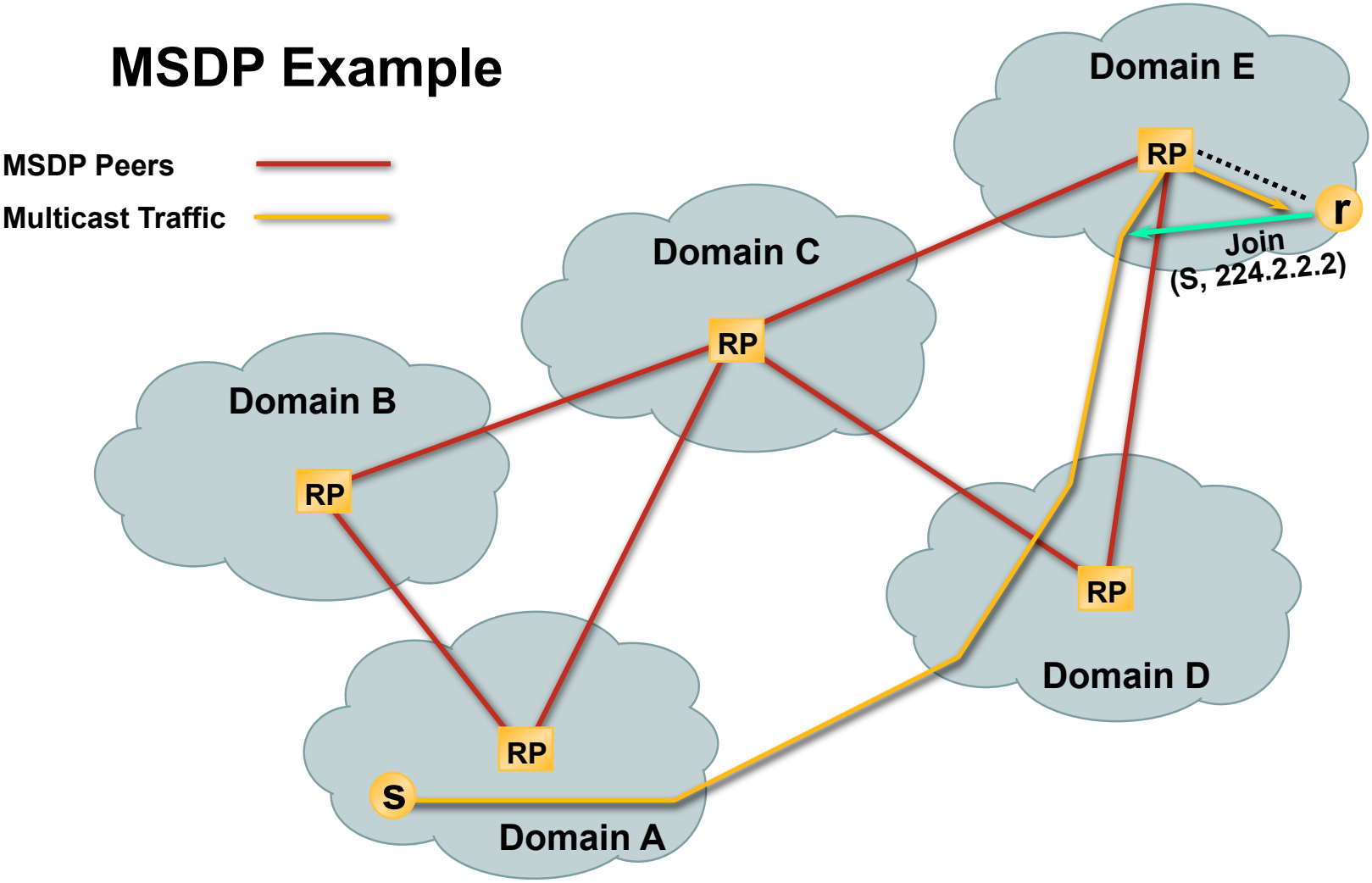
MSDP Overview



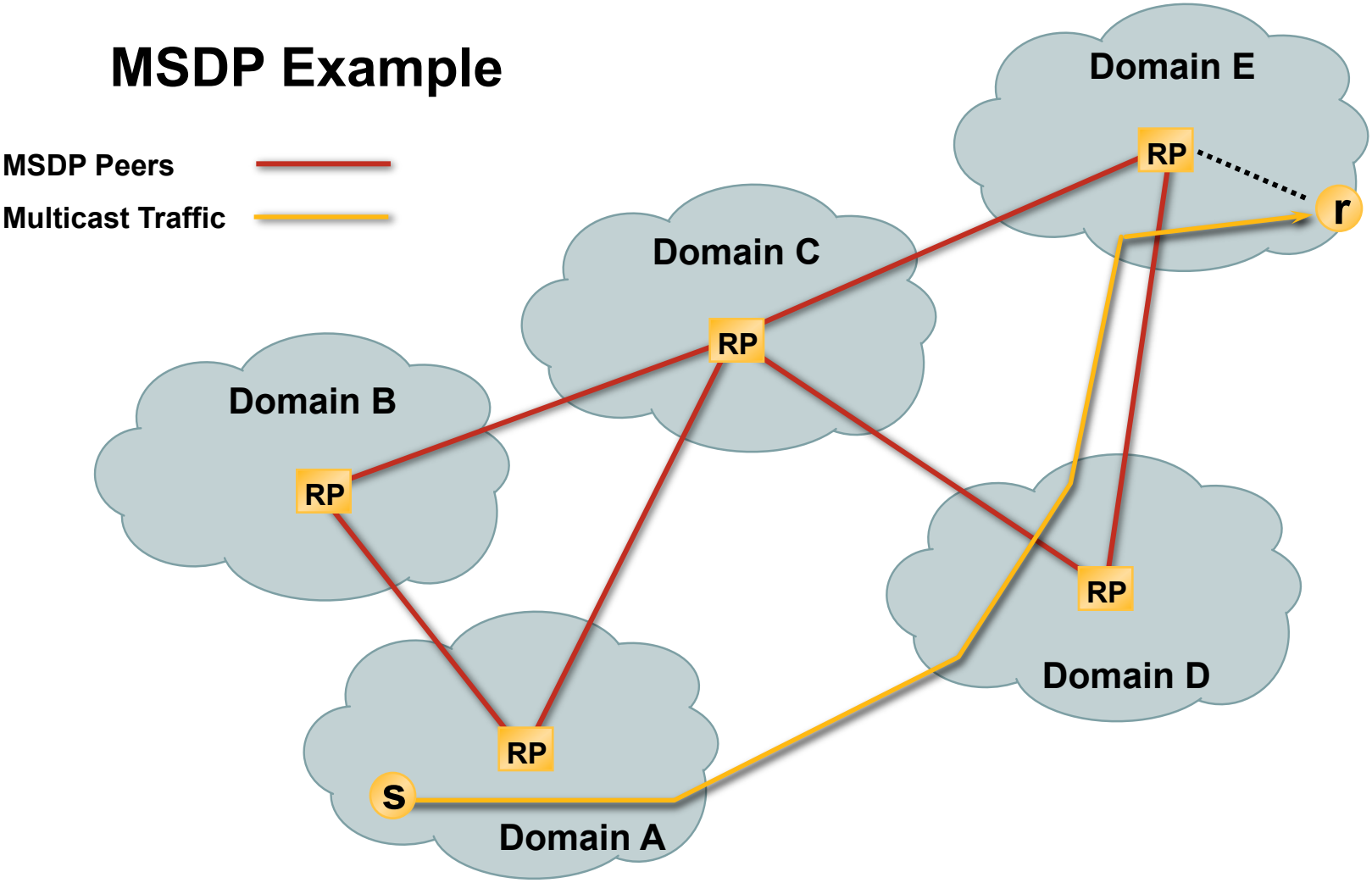
MSDP Overview



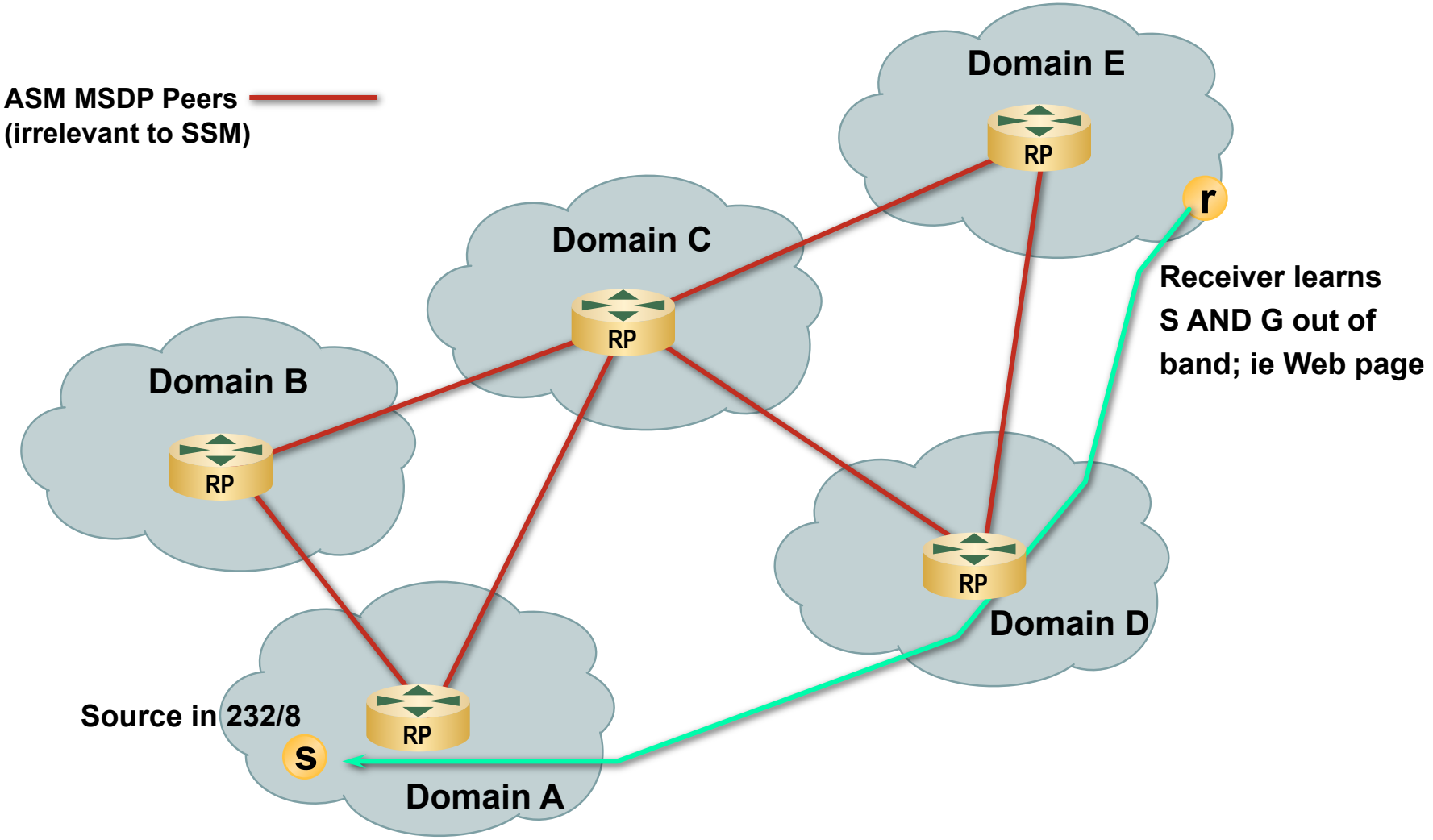
MSDP Overview



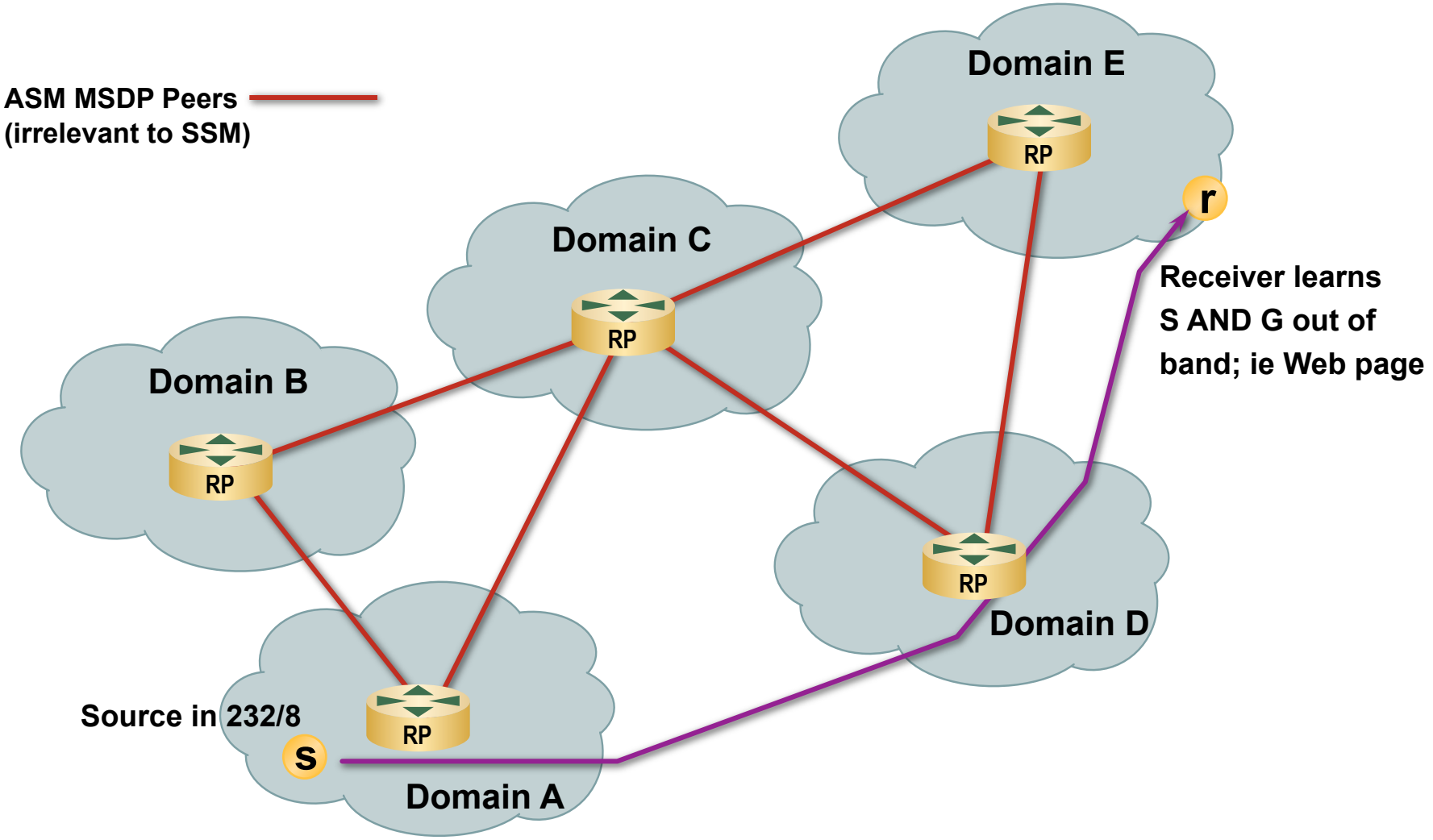
MSDP Overview



MSDP wrt SSM – Unnecessary!



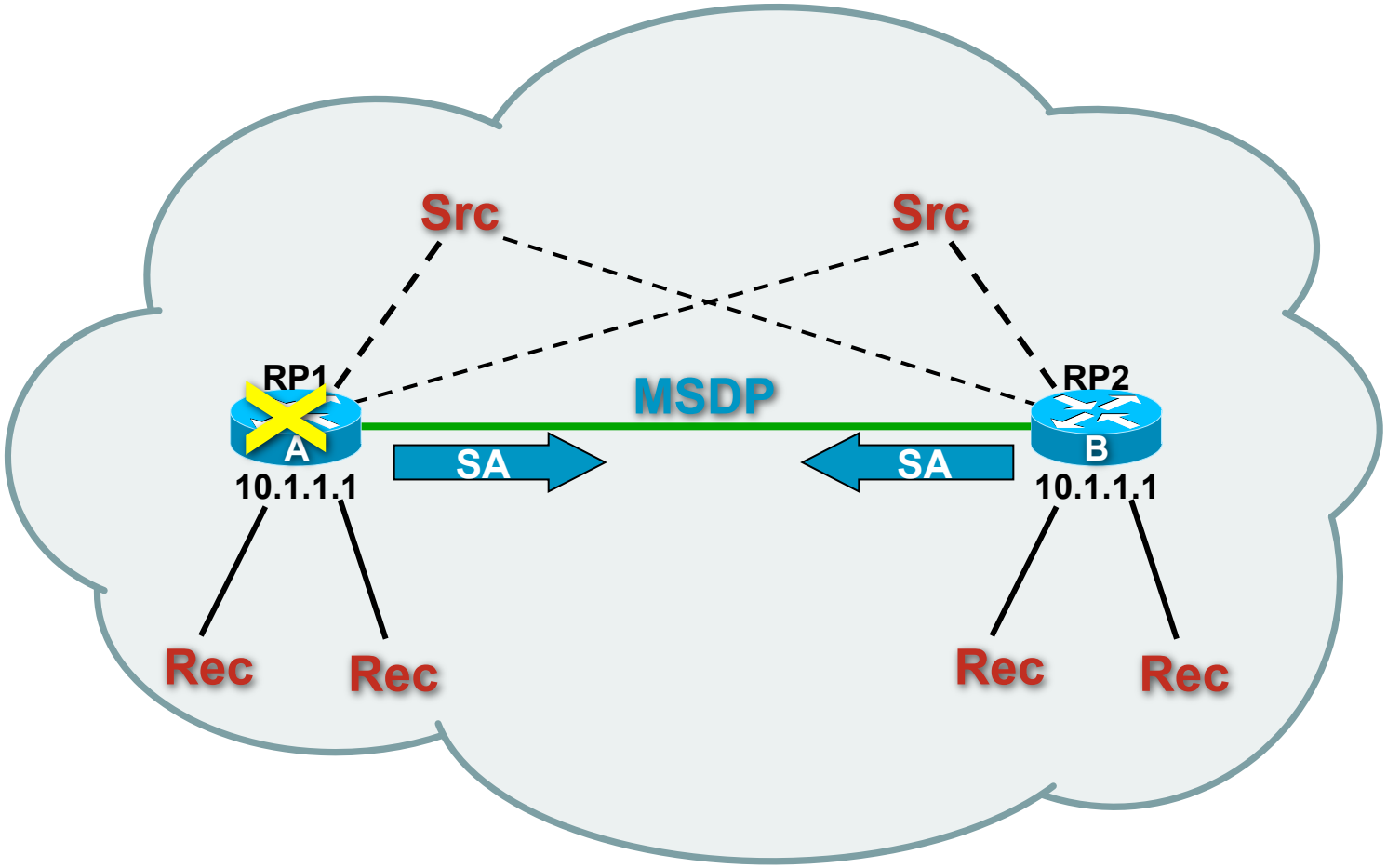
MSDP wrt SSM – Unnecessary!



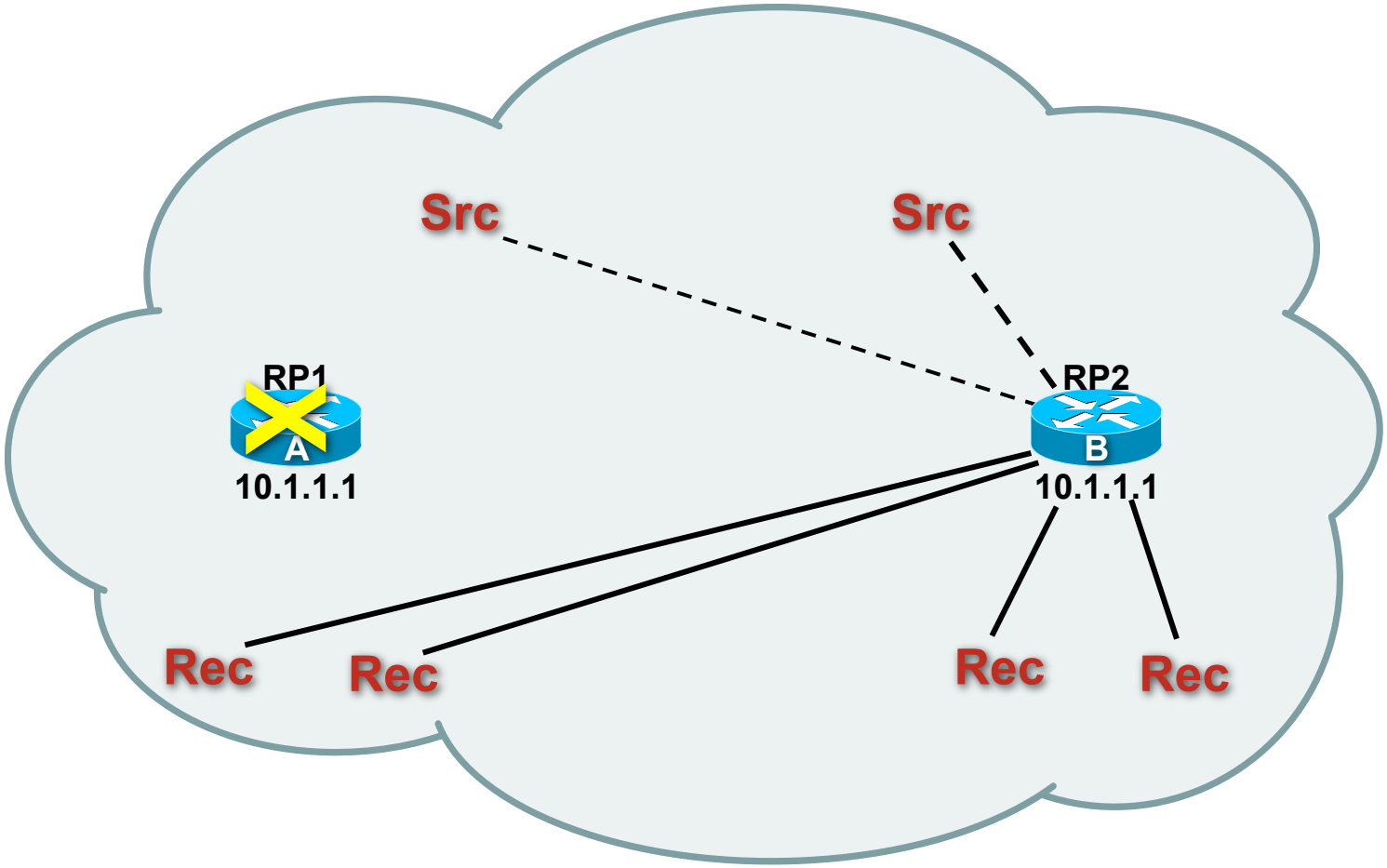
Anycast RP – Overview

- Redundant RP technique for PIM-SM ASM which uses MSDP for RP synchronization
- Uses single defined RP address
  - Two or more routers have same RP address
    - RP address defined as a Loopback Interface.
    - Loopback address advertised as a Host route.
  - Senders & Receivers Join/Register with closest RP
    - Closest RP determined from the unicast routing table.
    - Because RP is statically defined.
- MSDP session(s) run between all RPs
  - Informs RPs of sources in other parts of network
  - RPs join SPT to active sources as necessary

Anycast RP – Overview



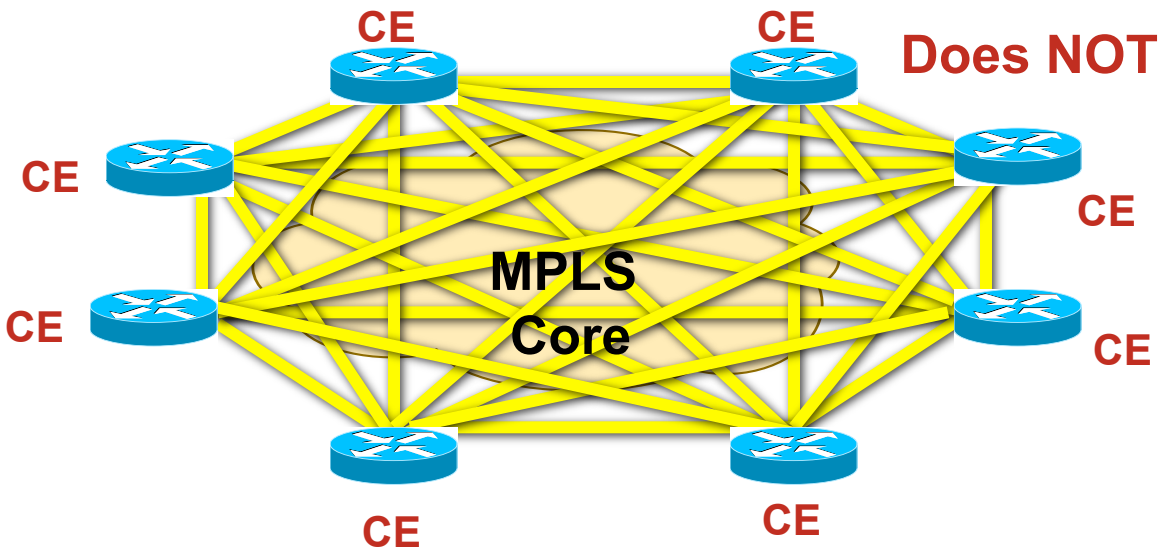
Anycast RP – Overview



Latest Additions

Multicast VPN – Customer Requirement

- MPLS VPN customers want to run multicast within their VPNs
- Multicast deployment is expanding
- MPLS VPNs do not support multicast today
- Multicast options in MPLS VPNs today
  - –GRE tunnels from CE to CE



Multicast VPN (MVPN)

- Allows an ISP to provide its MPLS VPN customers the ability to transport their Multicast traffic across MPLS packet-based core networks
- Requires IPmc enabled in the core
- MPLS may still be used to support unicast
- A scalable architecture solution for MPLS networks based on native multicast deployment in the core

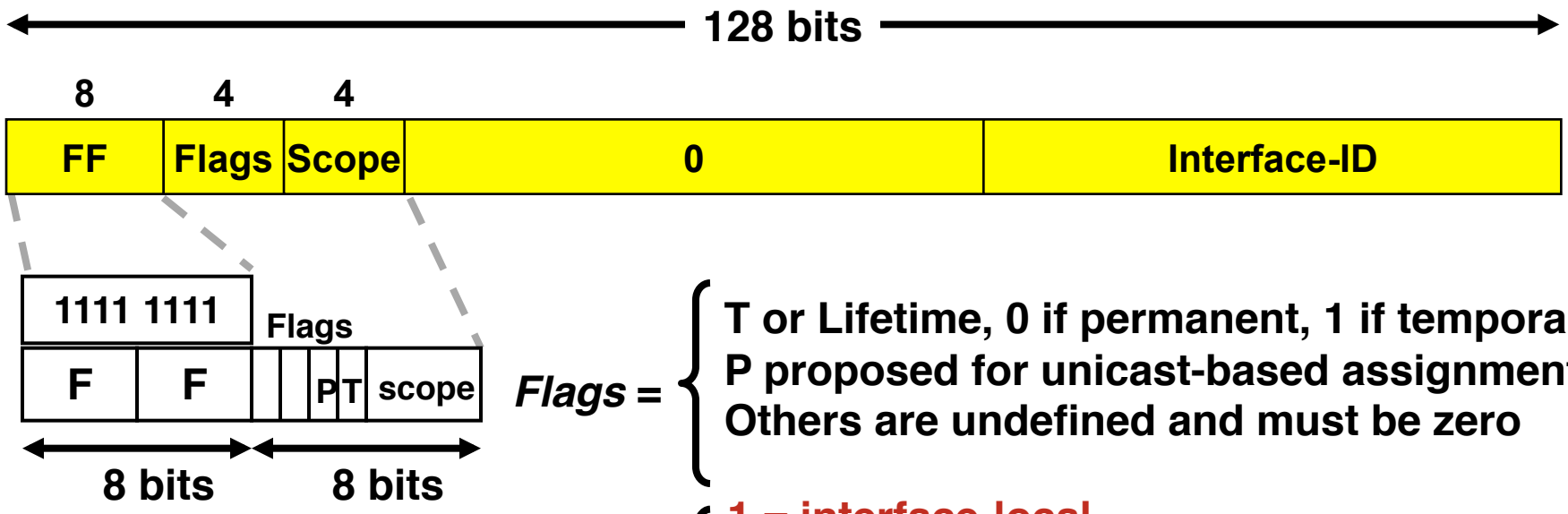
Multicast VPN (MVPN)

- Uses draft-rosen-vpn-mcast encapsulation and signaling to build MVPN Multicast VPN (MVPN)'
  - GRE encapsulation
  - PIM inside PIM
- Not universally deployed
  - Not all VPN providers offer MVPN services

IPv4 versus IPv6 Multicast

IP Service	IPv4 Solution	IPv6 Solution
Address Range	32-bit, class D	128-bit (112-bit Group)
Routing	Protocol Independent All IGPs,and BGP4+	Protocol Independent All IGPs,and BGP4+ with v6 mcast SAFI
Forwarding	PIM-DM, PIM-SM: ASM, SSM, BiDir	PIM-SM: ASM, SSM, BiDir
Group Management	IGMPv1, v2, v3	MLDv1, v2
Domain Control	Boundary/Border	Scope Identifier
Inter-domain Solutions	MSDP across Independent PIM Domains	Single RP within Globally Shared Domains

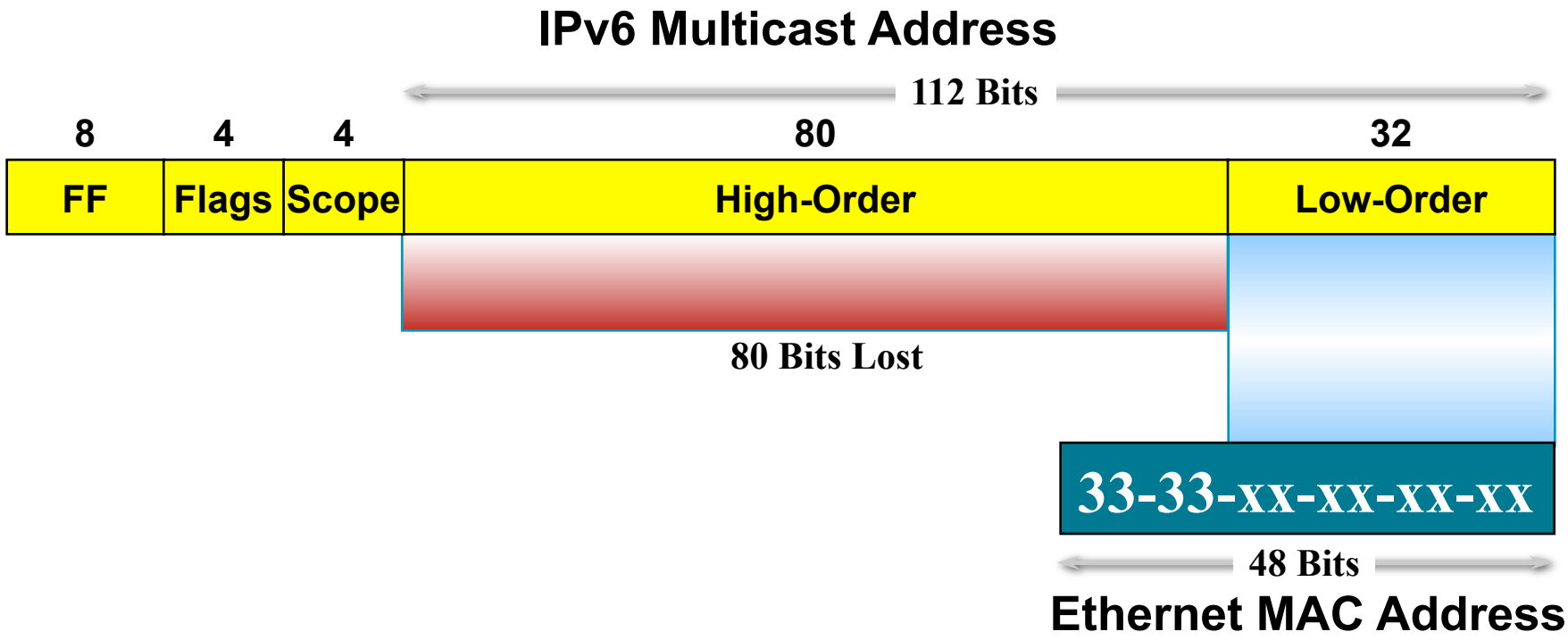
IPv6 Multicast Addresses (RFC 3513)



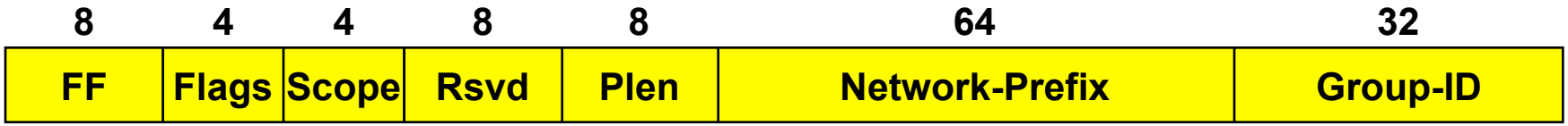
Flags =
- T or Lifetime, 0 if permanent, 1 if temporary
- P proposed for unicast-based assignments
- Others are undefined and must be zero

Scope =
- 1 = interface-local
- 2 = link
- 4 = admin-local
- 5 = site
- 8 = organization
- E = global

IPv6 Layer 2 Multicast Addressing Mapping



Unicast-based Multicast addresses



- RFC 3306 – Unicast-based Multicast Addresses
  - Similar to IPv4 GLOP Addressing
  - Solves IPv6 global address allocation problem.
  - Flags = 00PT

    P = 1, T = 1 => Unicast-based Multicast address
- Example:
  - Content Provider's Unicast Prefix

    1234:5678:9abc::/64
  - Multicast Address

    FF36:0030:1234:5678:9abc::0001

IP Routing for Multicast

- RPF based on reachability to v6 source same as with v4 multicast
- RPF still protocol independent:
  - Static routes, mroutes
  - Unicast RIB: BGP, ISIS, OSPF, EIGRP, RIP, etc
  - Multi-protocol BGP (mBGP)
    - Support for v6 mcast sub-address family
    - Provide translate function for non-supporting peers

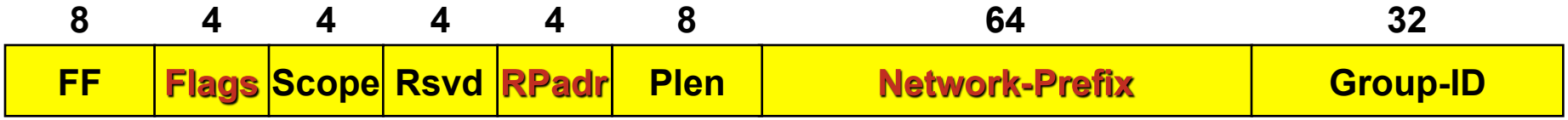
IPv6 Multicast Forwarding

- PIM-Sparse Mode (PIM-SM)
  - draft-ietf-pim-sm-v2-new-11.txt,
- PIM-Source Specific Mode (SSM)
  - RFC3569 SSM overview (v6 SSM needs MLDv2)
  - unicast prefix based multicast addresses ff30::/12
  - SSM range is ff3X::/32
    - Current allocation is from ff3X::/96
- PIM-bidirectional Mode (PIM-bidir)
  - draft-ietf-pim-bidir-07.txt

RP mapping mechanisms for IPv6 ASM

- Static RP assignment
- BSR
- Auto-RP – no current plans
- Embedded RP

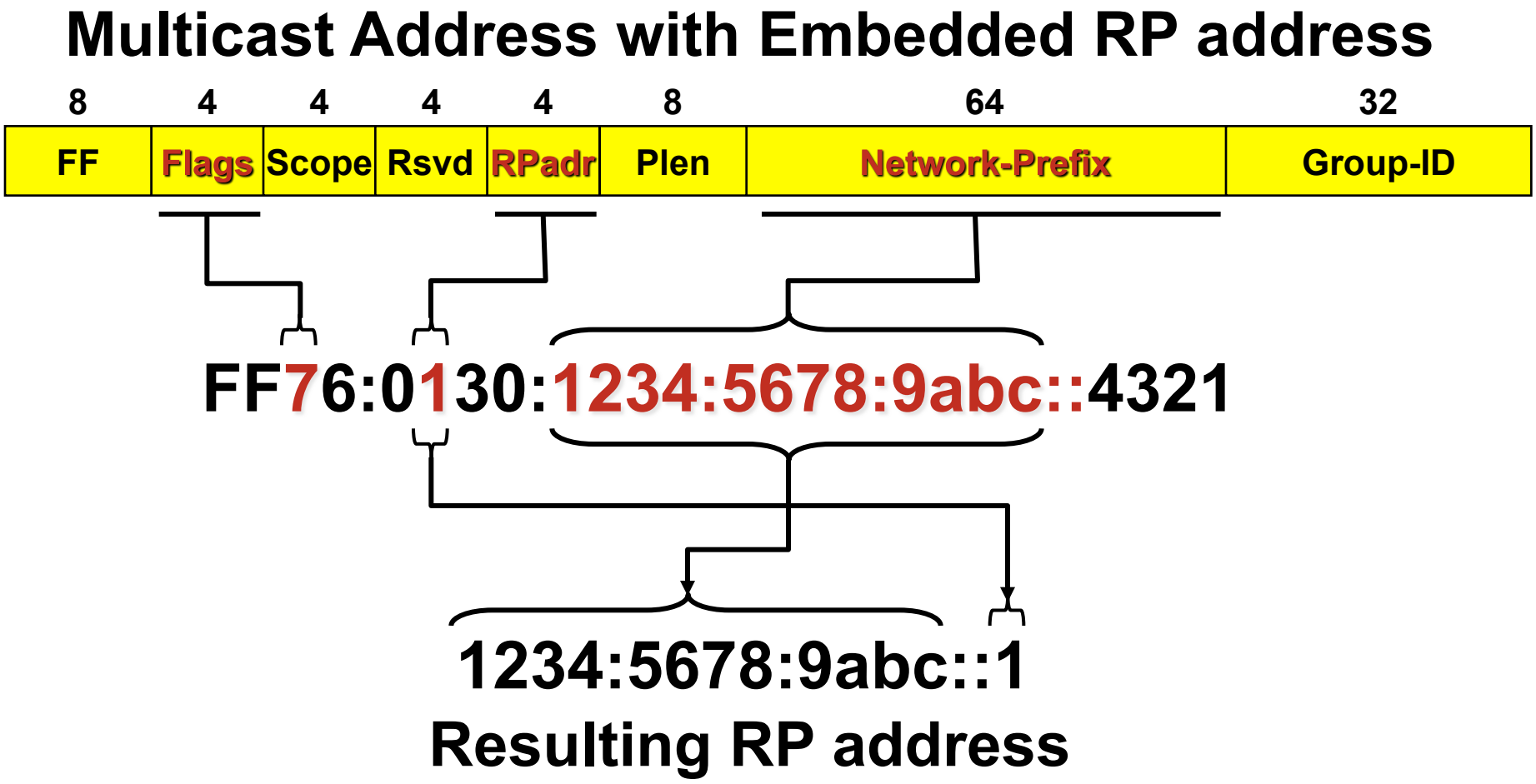
Embedded RP Addressing – RFC3956



- Proposed new multicast address type
  - Uses Unicast-Based Multicast addresses (RFC 3306)
- RP Address is embedded in multicast address.
- Flag bits = 0RPT

  R = 1, P = 1, T = 1 => Embedded RP Address
- Network-Prefix::RPadr = RP address
- For each Unicast prefix you own, you now also own:
  - 16 RPs for each of the 16 Multicast Scopes (256 total) with 2^32 multicast groups assigned to each RP (2^40 total)

Embedded RP Addressing – Example



Multicast Listener Discover – MLD

- MLD is equivalent to IGMP in IPv4
- MLD messages are transported over ICMPv6
- Version number confusion:
  - MLDv1 corresponds to IGMPv2
    - RFC 2710
  - MLDv2 corresponds to IGMPv3, needed for SSM
    - RFC 3810
- MLD snooping
  - draft-ietf-magma-snoop-12.txt

Now you know…

- Why Multicast?
- Multicast Fundamentals
- PIM Protocols
- RP choices
- Multicast at Layer 2
- Interdomain IP Multicast
- Some Latest Additions