MULTICAST Tutorial

RedIRIS/Red.es

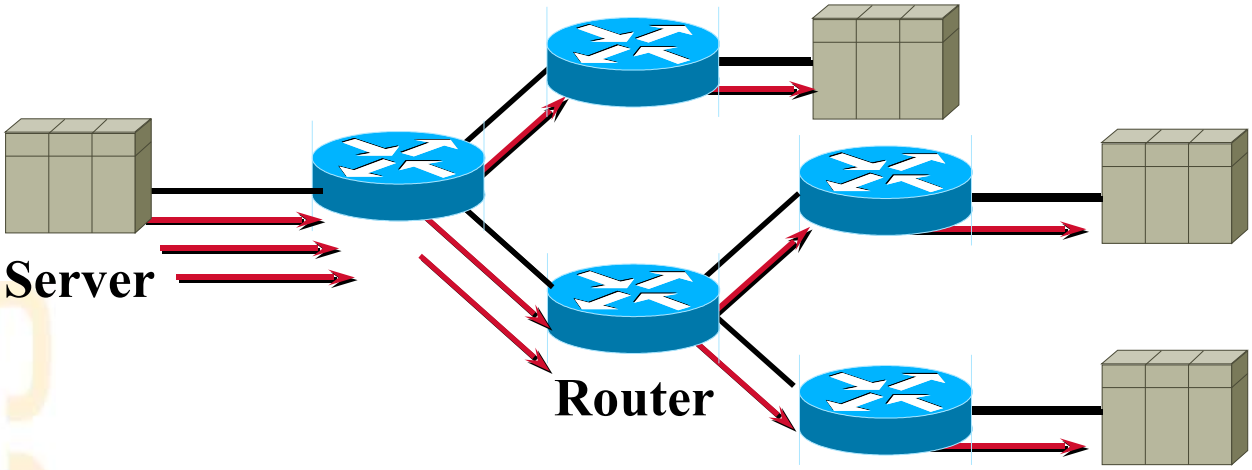
October 7th, 2004

Agenda

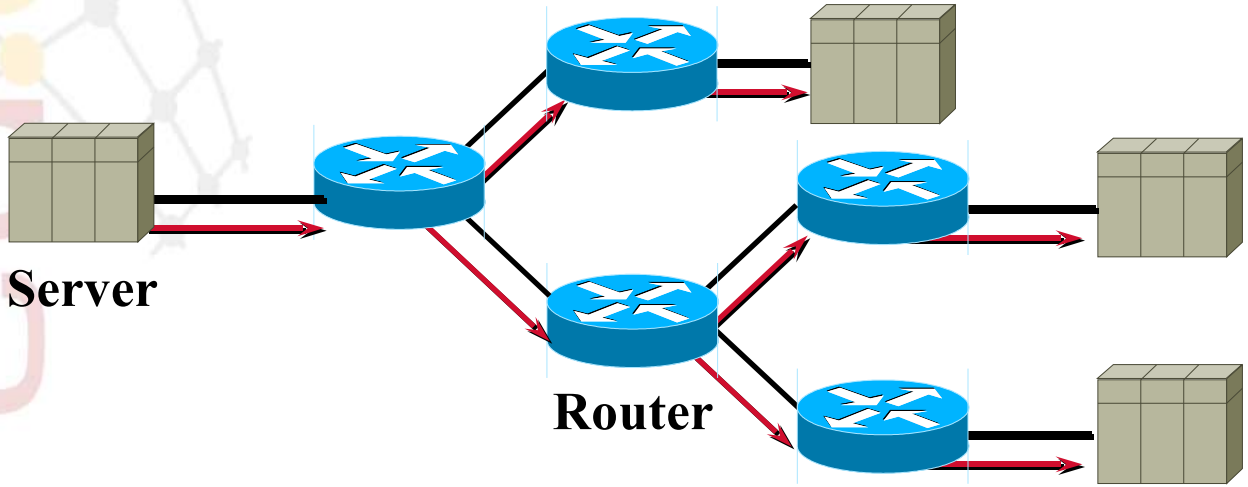
- Introduction
- Multicast addressing
- Group Membership Protocol
- PIM-SM / SSM
- MSDP
- MBGP

- Introduction
- Multicast addressing
- Group Membership Protocol
- PIM-SM / SSM
- MSDP
- MBGP

What is Multicasting?



Unicast



Multicast

Multicast Uses

Any Applications with multiple receivers

- 1-to-many or many-to-many

Live Video distribution

- Seminars, conferences, workshops ....

Collaborative groupware

e-Learning

Periodic Data Delivery - "Push" technology

- stock quotes, sports scores, magazines, newspapers
- advertisements

Server/Web-site replication

Reducing Network/Resource Overhead

- more efficient to establish multicast tree rather then multiple point-to-point links

Resource Discovery

A bit of history

In 1995 the first mcast network was born: MBone

DVMRP (Distance Vector Multicast Routing Protocol) was the protocol used

- DVMRP subnetworks was interconnected through the unicast Internet infrastructure with tunnels
- Flood and Prune technology
- Very successful in academic circles

Problem

- **DVMRP can't scale to Internet sizes**
  - ❑ Distance vector-based routing protocol
  - ❑ Periodic updates
    - Full table refresh every 60 seconds
  - ❑ Table sizes
    - Internet > 40,000 prefixes at that moment
  - ❑ Scalability
    - Too many tunnels, hop-count till 32 hops, etc

=> In 1997, a native protocol is developed, Protocol Independent Multicast

The evolution

PIM Dense mode

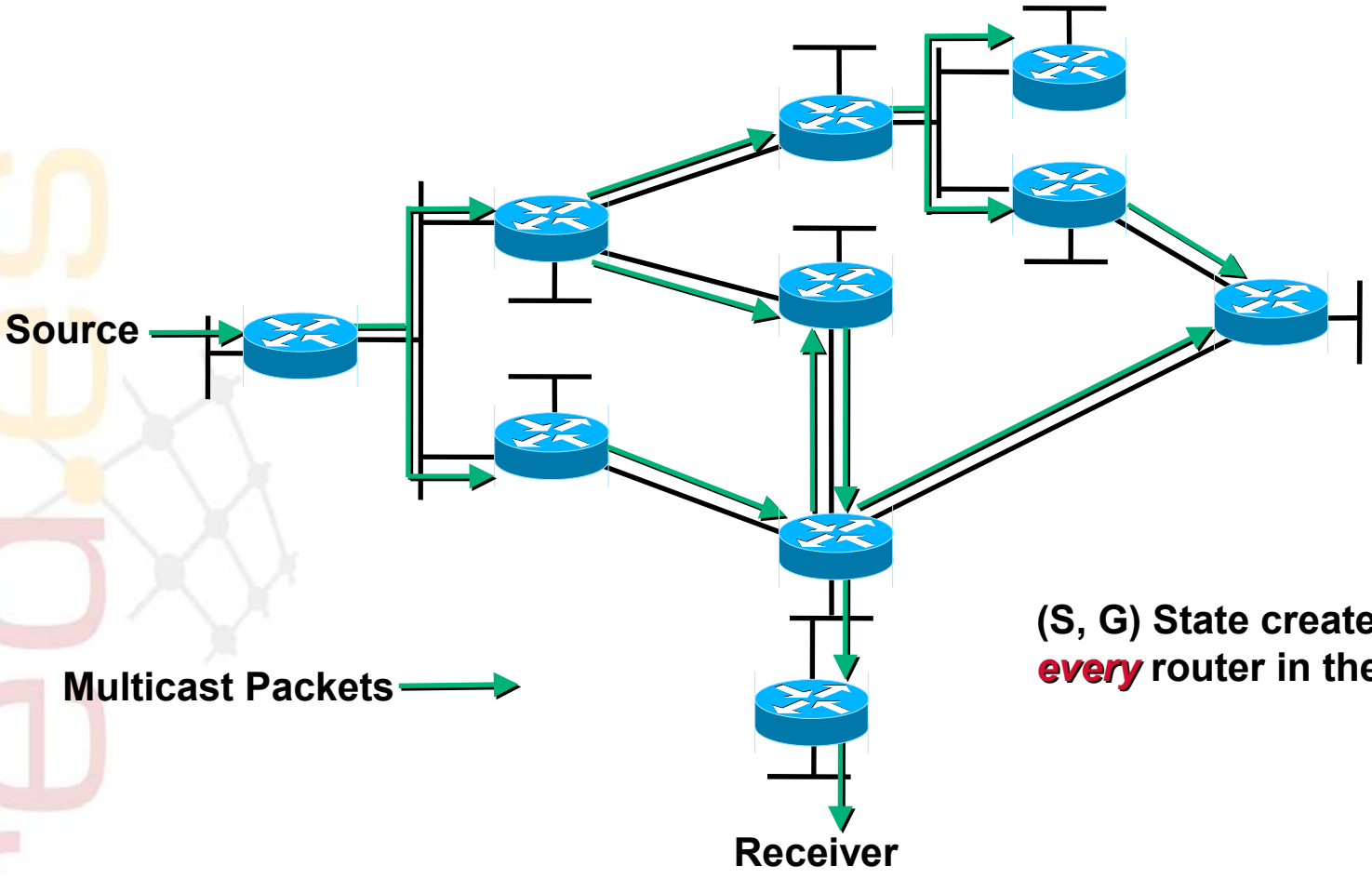
- Flood and Prune behavior very inefficient
  - Can cause problems in certain network topologies
- Creates (S, G) state in EVERY router
  - Even when there are no receivers for the traffic
- Complex Assert mechanism
  - To determine which router in a LAN will forward the traffic
- No support for shared trees

The evolution

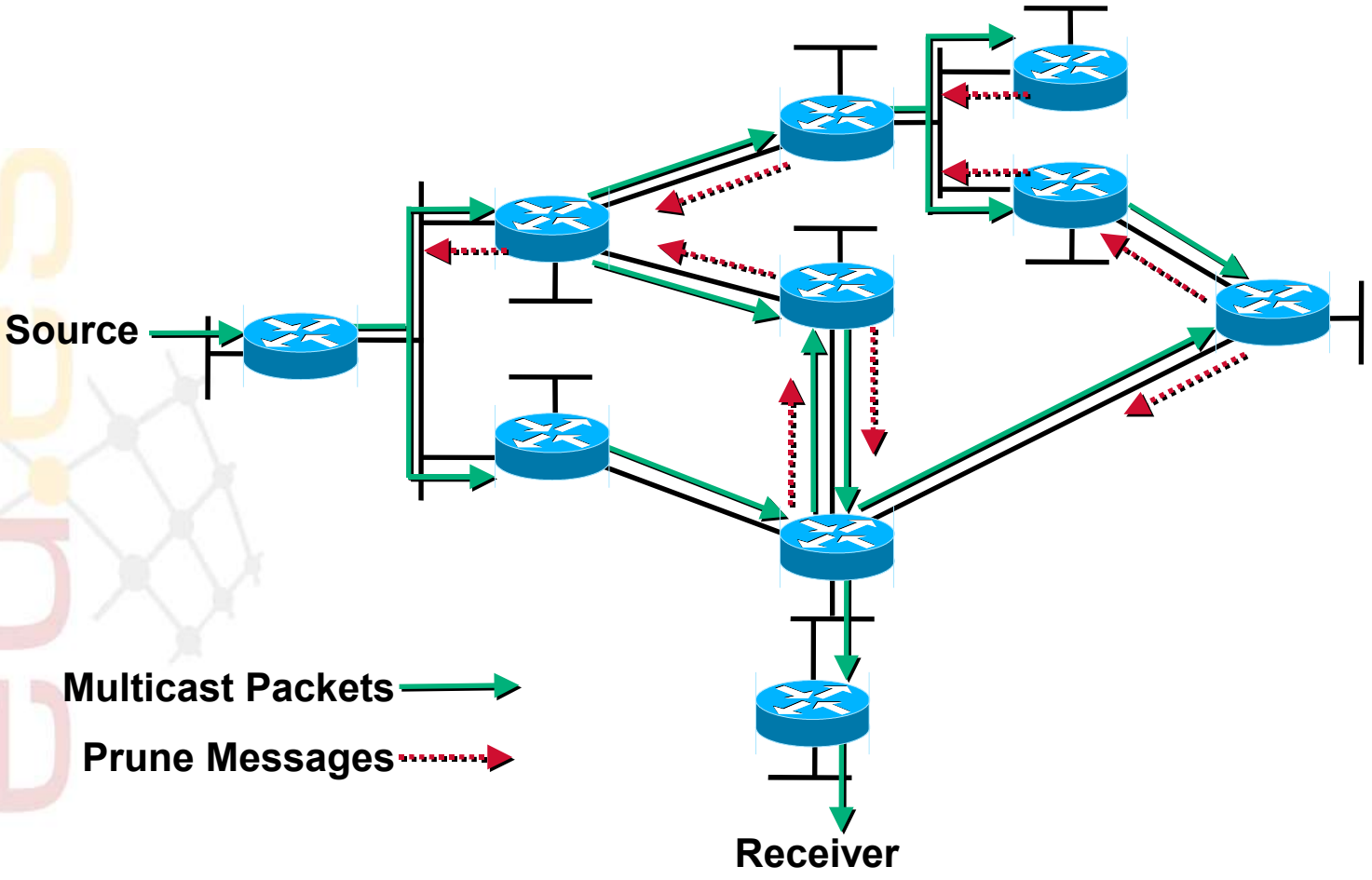
PIM Sparse mode

- Must configure a Rendezvous Point (RP)
  - Statically (on every Router)
  - Using Auto-RP or BSR (Routers learn RP automatically)
- Very efficient
  - Uses Explicit Join model
  - Traffic only flows to where it's needed
  - Router state only created along flow paths
- Scales better than dense mode
  - Works for *both* sparsely or densely populated networks

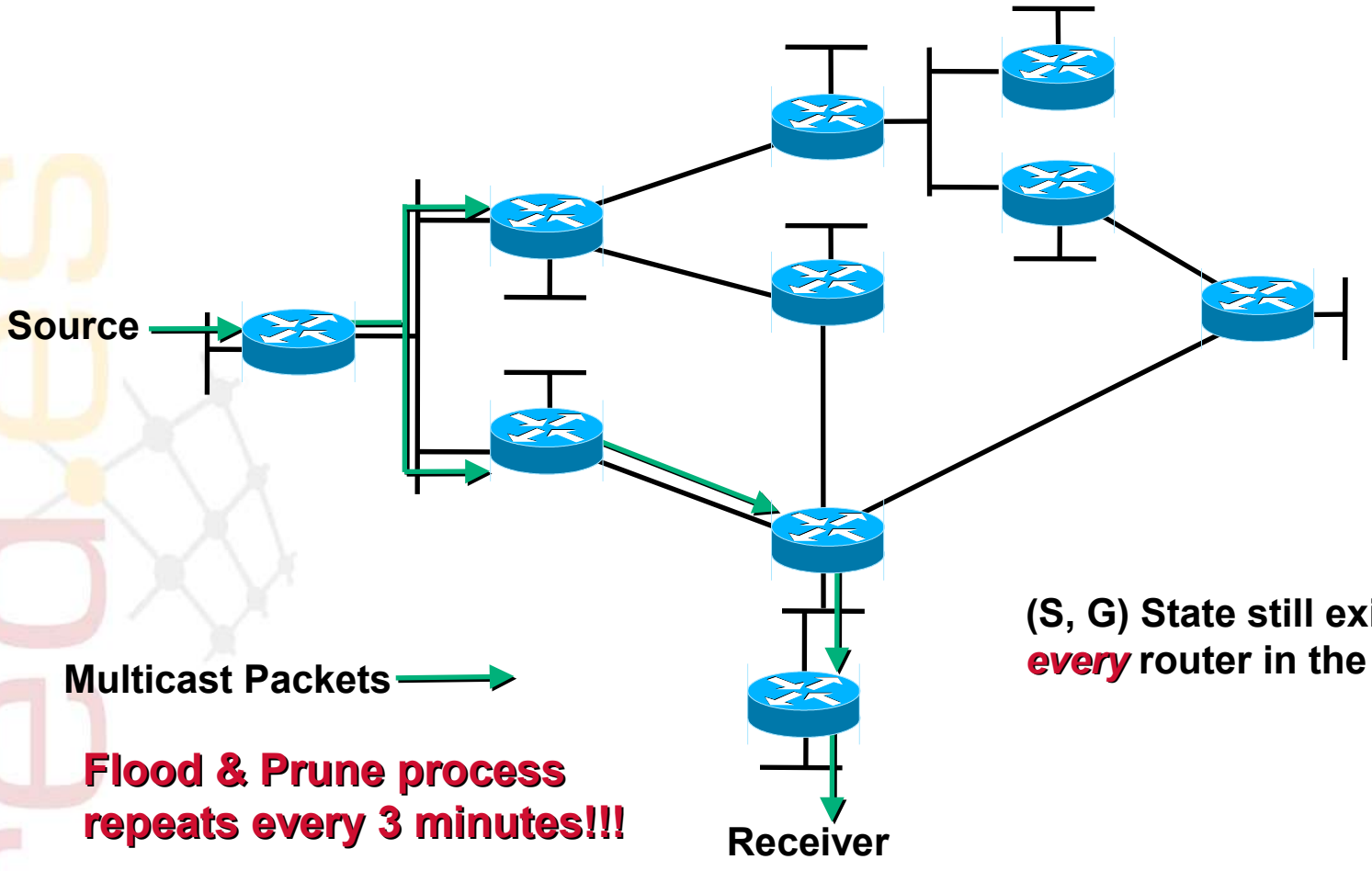
Initial Flooding



Pruning Unwanted Traffic



Results After Pruning



(S, G) State still exists in *every* router in the network!

Flood & Prune process repeats every 3 minutes!!!

(S,G) notation

- **For every multicast source there must be two pieces of information: the source IP address, S, and the group address, G.**

  - This is generally expressed as (S,G).
  - Also commonly used is (*,G) - every source for a particular group.
  - The router creates a table with the entries (*,G), (S,G).

IP Multicast building blocks

The SENDERS send

- Multicast Addressing - rfc1700
- class D (224.0.0.0 - 239.255.255.255)

The RECEIVERS inform the routers what they want to receive

- Internet Group Management Protocol (IGMP) - rfc2236 -> version 2

The routers make sure the STREAMS make it to the correct receiving nets.

- Multicast Routing Protocols (PIM-SM/SSM)
- RPF (reverse path forwarding) – against source address

Multicast 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.

Multicast Routing uses "Reverse Path Forwarding"

Reverse Path Forwarding (RPF)

What is RPF?

- A router forwards a multicast datagram only if received on the up stream interface to the source (i.e. it follows the distribution tree).

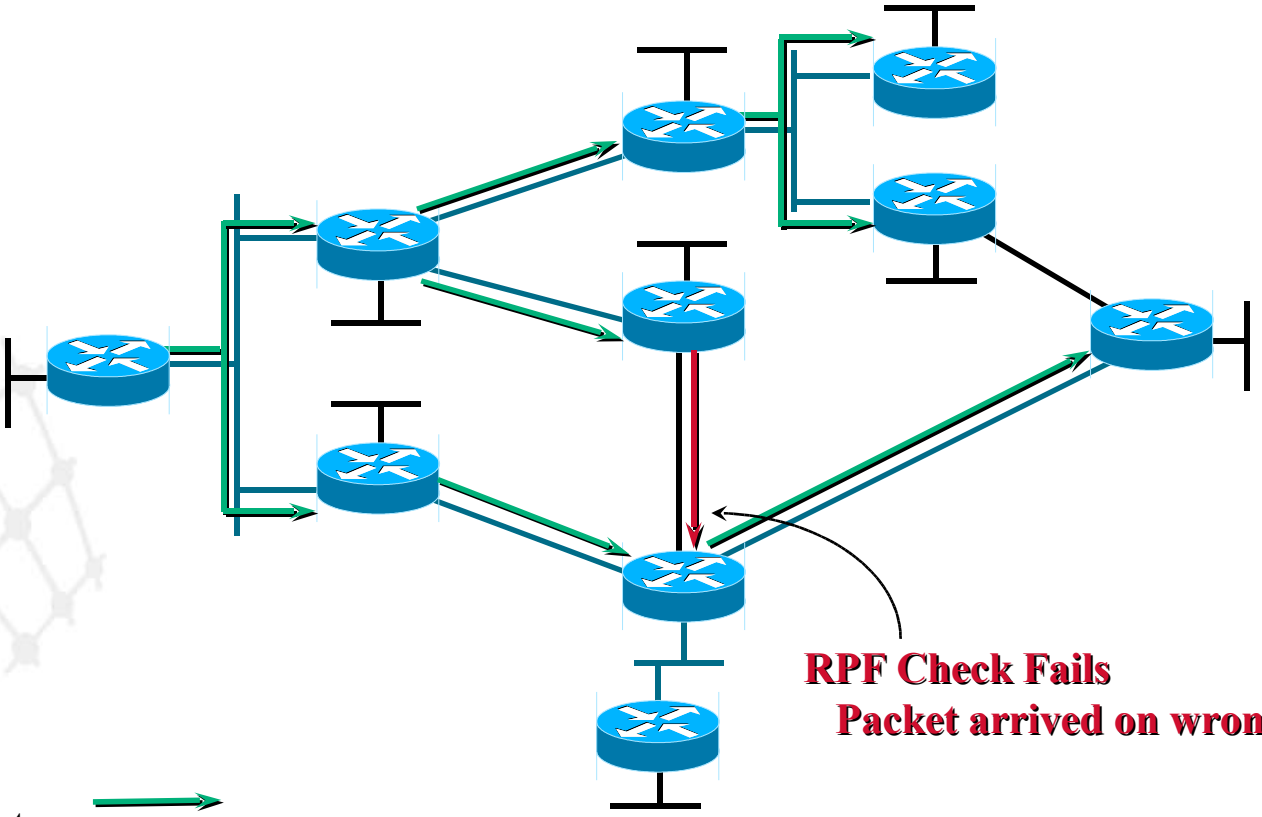
The RPF Check

- The source IP address of incoming multicast packets are checked against a unicast routing table.
- If the datagram arrived on the interface specified in the routing table for the source address; then the RPF check succeeds.
- Otherwise, the RPF Check fails.

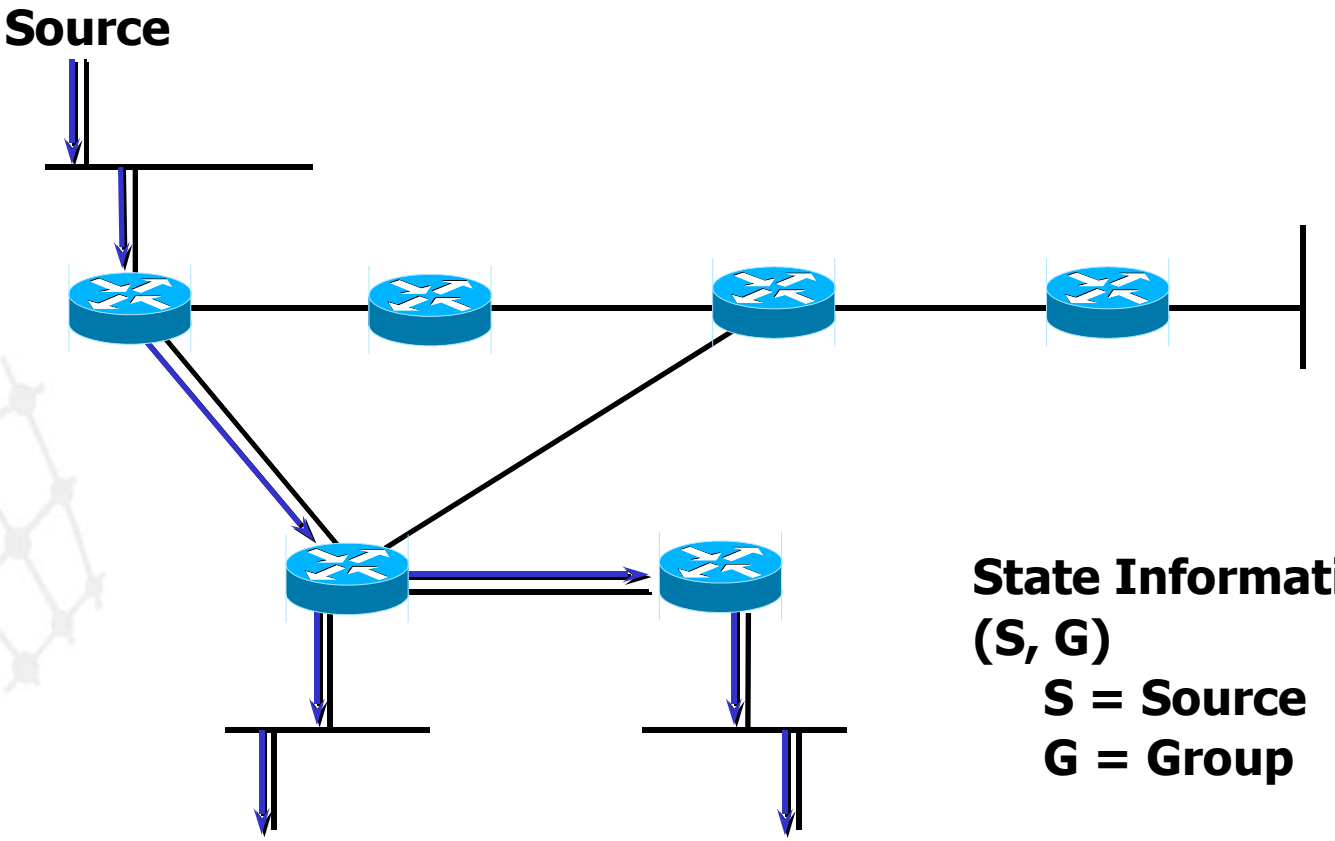
Multicast Forwarding

- **Multicast uses unicast routes to determine path back to source**
- **RPF checks ensures packets won't loop**
- **RPF checks are performed against routing table by default**
- **If multicast path is different from unicast path, then a multicast table will exist. It will be use for RPF check.**
- **Routes contain incoming interface**
  - Packets matching are forwarded
  - Packets mis-matching are dropped

Example: RPF Checking

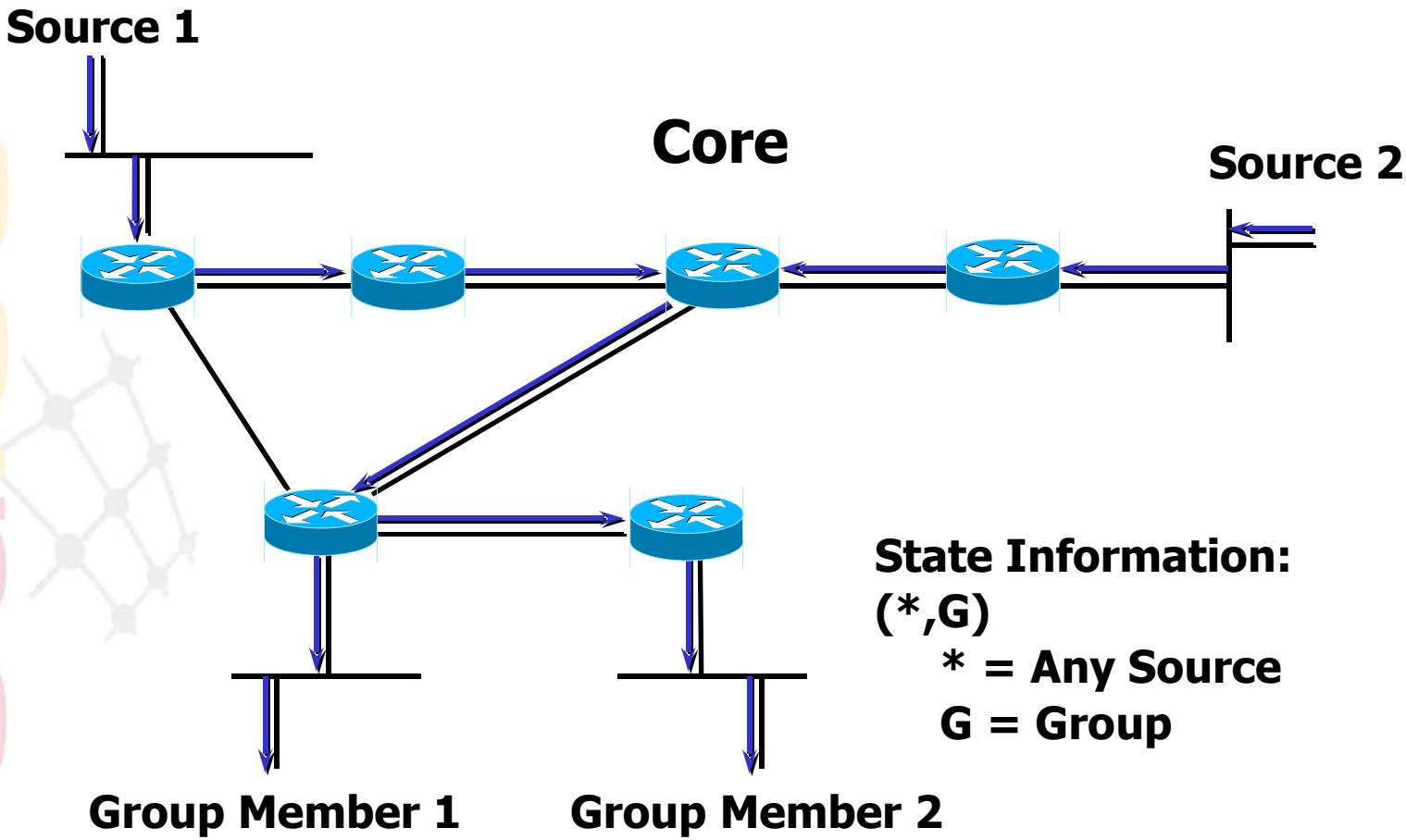


Shortest Path or Source Based Distribution Tree



State Information:
(S, G)
S = Source
G = Group

Shared or Core Based Distribution Tree



Multicast Distribution Trees

Source or Shortest Path trees

- More resource intensive; requires more states (S,G)
- You get optimal paths from source to all receivers, minimizes delay
- Best for one-to-many distribution

Shared or Core Based trees

- Uses less resources; less memory (*,G)
- You may get sub optimal paths from source to all receivers, depending on topology
- The RP (core) itself and its location may affect performance
- Best for many-to-many distribution
- May be necessary for source discovery (PIM-SM)

- Introduction
- Multicast addressing
- Group Membership Protocol
- PIM-SM / SSM
- MSDP
- MBGP

Multicast Addressing

IP Multicast Group Addresses

- 224.0.0.0–239.255.255.255
- Class “D” Address Space
  - High order bits of 1st Octet = “1110”
- TTL value defines scope and limits distribution
  - IP multicast packet must have TTL > interface TTL or it is discarded
  - values are: 0=host, 1=network, 32=same site, 64=same region, 128=same continent, 255=unrestricted
  - No longer recommended as a reliable scoping mechanism

Multicast Addressing

Administratively Scoped Addresses – RFC 2365

- 239.0.0.0–239.255.255.255
- Private address space
  - Similar to RFC 1918 unicast addresses
  - Not used for global Internet traffic
  - Used to limit “scope” of multicast traffic
  - Same addresses may be in use at different locations for different multicast sessions
- Examples
  - Site-local scope: 239.253.0.0/16
  - Organization-local scope: 239.192.0.0/14

Multicast Addressing

GLOP addresses

- Provides globally available private Class D space
- 233.x.x/24 per AS number
- RFC2770

How?

- AS number = 16 bits
  - Insert the 16 ASN into the middle two octets of 233/8

Online Glop Calculator:

www.shepfarm.com/multicast/glop.html

Multicast Addressing

http://www.iana.org/assignments/multicast-addresses

Examples of Reserved & Link-local Addresses

- 224.0.0.0 - 224.0.0.255 reserved & not forwarded
- 239.0.0.0 - 239.255.255.255 Administrative Scoping
- 232.0.0.0 - 232.255.255.255 Source-Specific Multicast
- 224.0.0.1 - All local hosts
- 224.0.0.2 - All local routers
- 224.0.0.4 - DVMRP
- 224.0.0.5 - OSPF
- 224.0.0.6 - Designated Router OSPF
- 224.0.0.9 - RIP2
- 224.0.0.13 - PIM
- 224.0.0.15 - CBT
- 224.0.0.18 - VRRP

- Introduction
- Multicast addressing
- Group Membership Protocol
- PIM-SM / SSM
- M-BGP
- MSDP

Internet Group Membership Protocol (IGMP)

How hosts tell routers about group membership

Routers solicit group membership from directly connected hosts

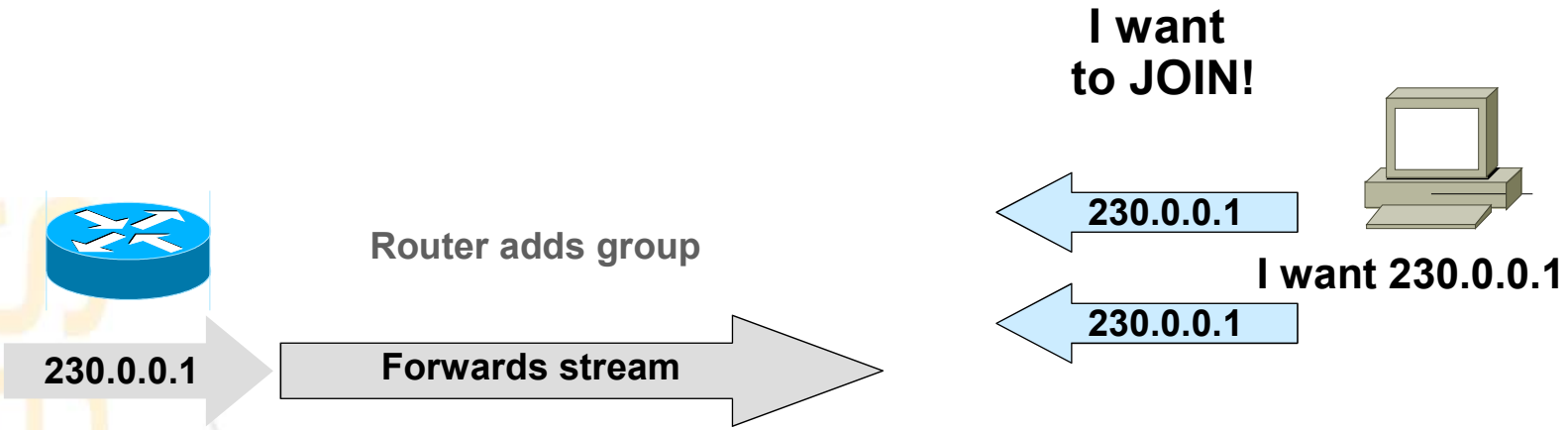
RFC 2236 specifies version 2 of IGMP

- Supported on every OS

IGMP version 3 is the latest version

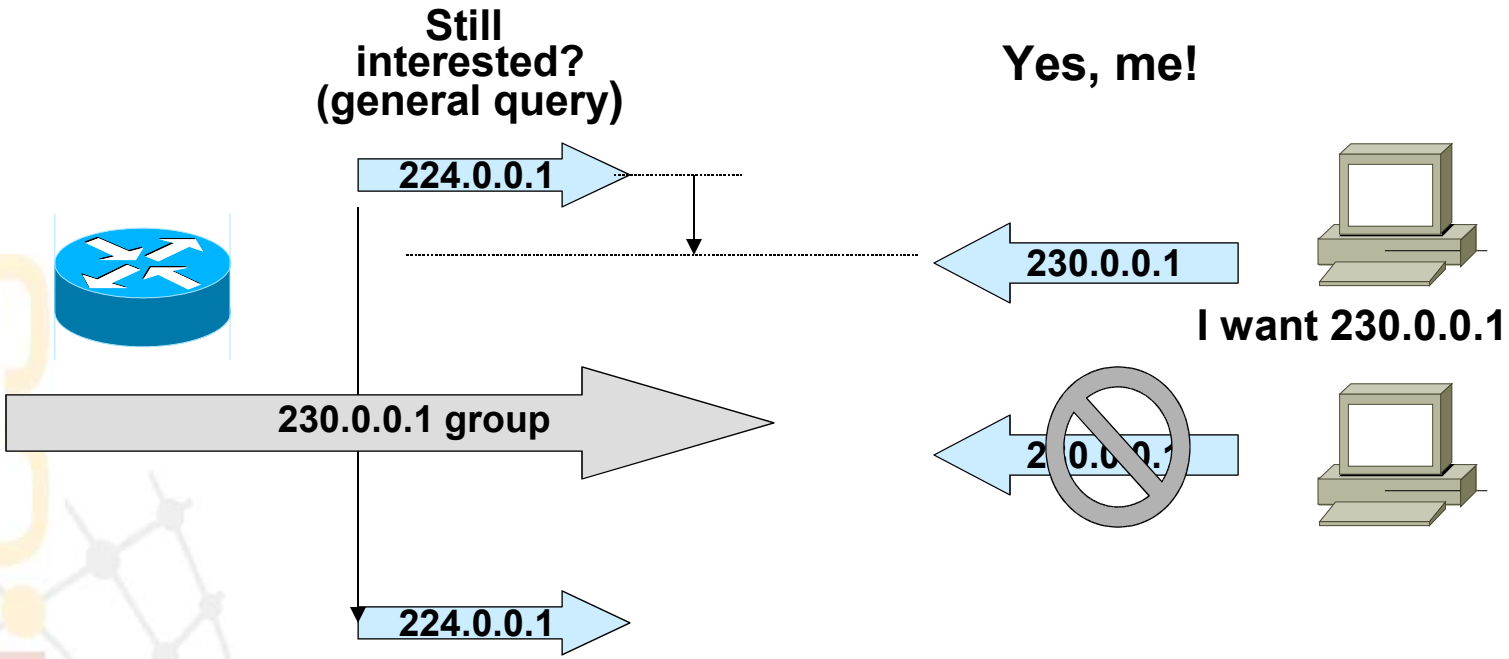
- RFC 3376
- provides source include-list capabilities (SSM!)
- Support?
  - Unix latest versions, Window XP

IGMPv2 Protocol Flow - Join a Group



- Router triggers group membership request to PIM.
- Hosts can send unsolicited *join* membership messages – called reports in the RFC (usually more than 1)
- Or hosts can join by responding to periodic query from router

IGMPv2 Protocol Flow - Querier

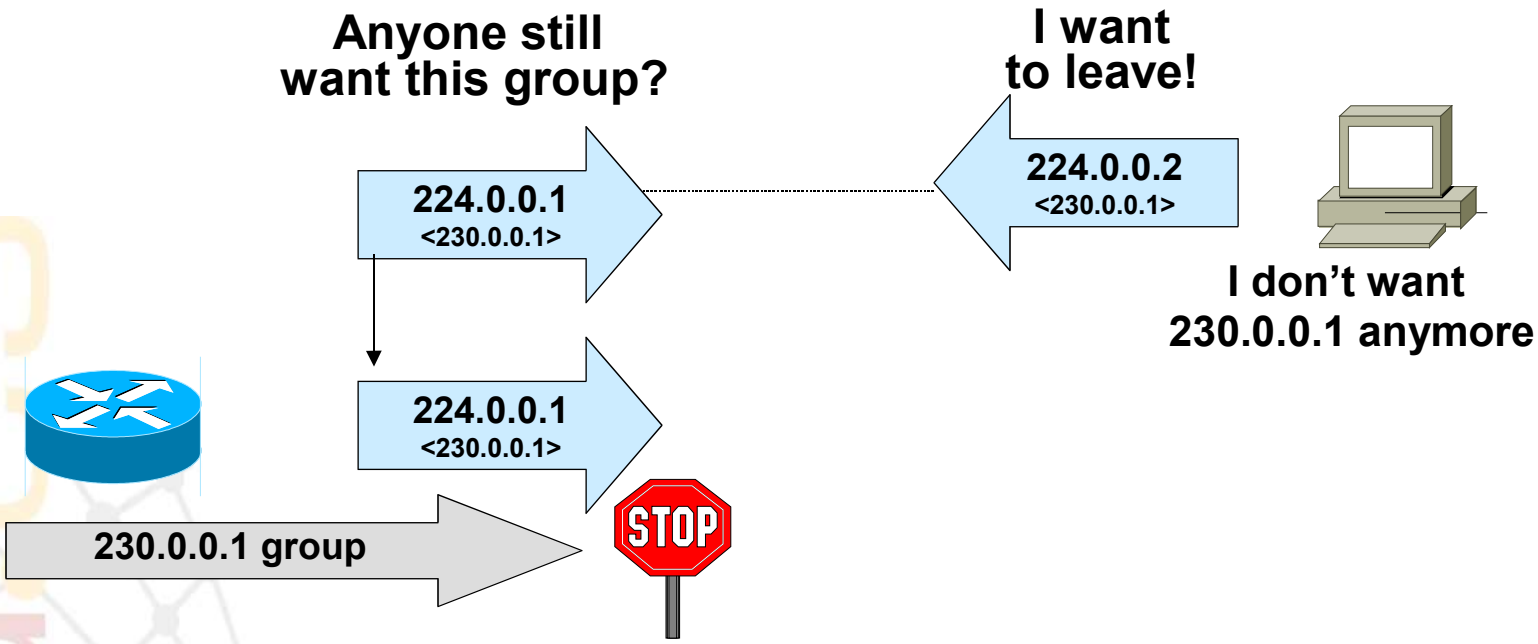


Hosts respond to *query* to indicate (new or continued) interest in group(s)

- only 1 host should respond per group
  - Hosts fall into idle-member state when same-group report heard.

After 260 sec with no response, router times out group

IGMPv2 Protocol Flow - Leave a Group

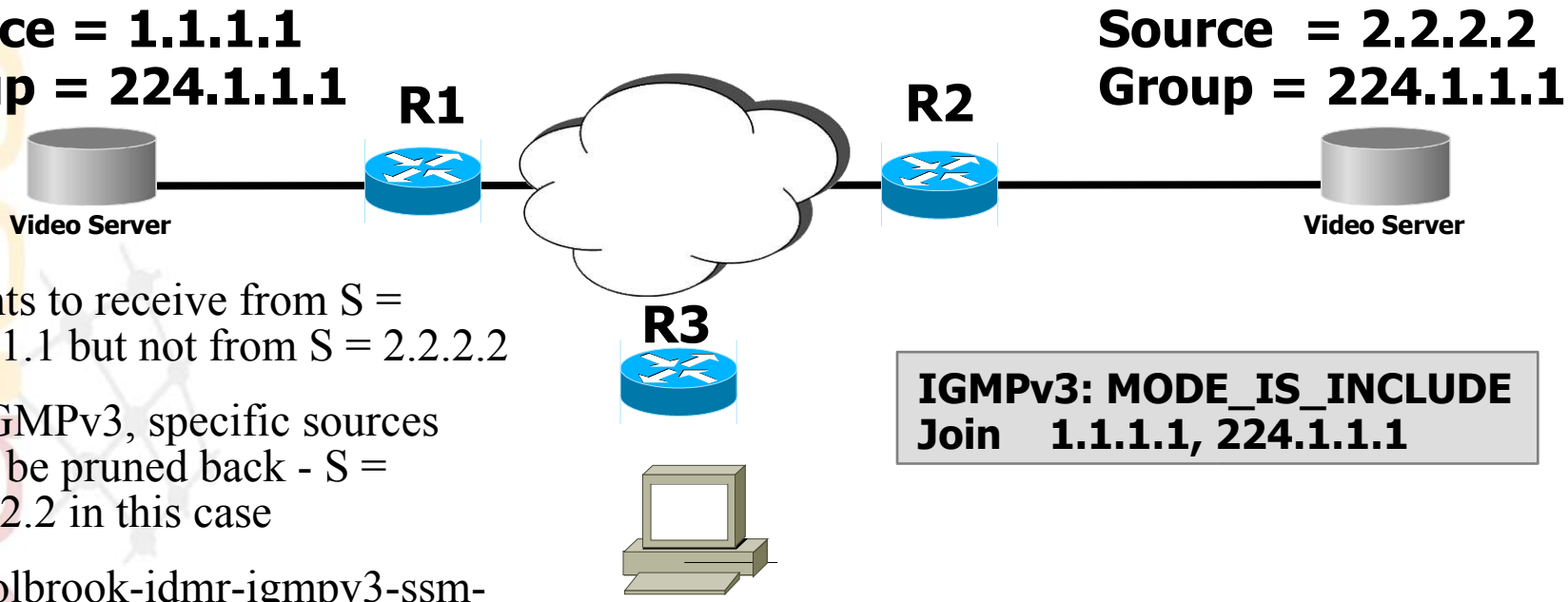


Hosts send *leave* messages to all routers group indicating group they're leaving.

- Router follows up with 2 *group-specific queries* messages

RFC 3376

Enables hosts to listen only to a specified subset of the hosts sending to the group



H1 wants to receive from S = 1.1.1.1 but not from S = 2.2.2.2

With IGMPv3, specific sources can be pruned back - S = 2.2.2.2 in this case

draft-holbrook-idmr-igmpv3-ssm-01.txt

IGMP Enhancements

IGMP Version 2

- multicast router with lowest IP address is elected querier
- Group-Specific Query message is defined. Enables router to transmit query to specific multicast address rather than to the "all-hosts" address of 224.0.0.1
- Leave Group message is defined. Last host in group wishes to leave, it sends Leave Group message to the "all-routers" address of 224.0.0.2. Router then transmits Group-Specific query and if no reports come in, then the router removes that group from the list of group memberships for that interface

IGMP Version 3

- Group-Source Report message is defined. Enables hosts to specify which senders it can receive or not receive data from.
- Group-Source Leave message is defined. Enables host to specify the specific IP addresses of a (source,group) that it wishes to leave.

- Introduction
- Multicast addressing
- Group Membership Protocol
- PIM-SM / SSM
- MSDP
- MBGP

Protocol Independent Multicast - sparse mode

- draft-ietf-pim-sm-v2-new-10.txt
  - Obsoletes RFC 2362
  - BSR removed from PIM spec.
- explicit join: assumes everyone does not want the data
- uses unicast routing table for RPF checking
- data and joins are forwarded to RP for initial rendezvous
- all routers in a PIM domain must have RP mapping
- when load exceeds threshold forwarding swaps to shortest path tree (default is first packet)
- state increases (not everywhere) as number of sources and number of groups increase
- source-tree state is refreshed when data is forwarded and with Join/Prune control messages

PIM Sparse-Mode :RP

Allows Source Trees or Shared Trees

Rendezvous Point (RP)

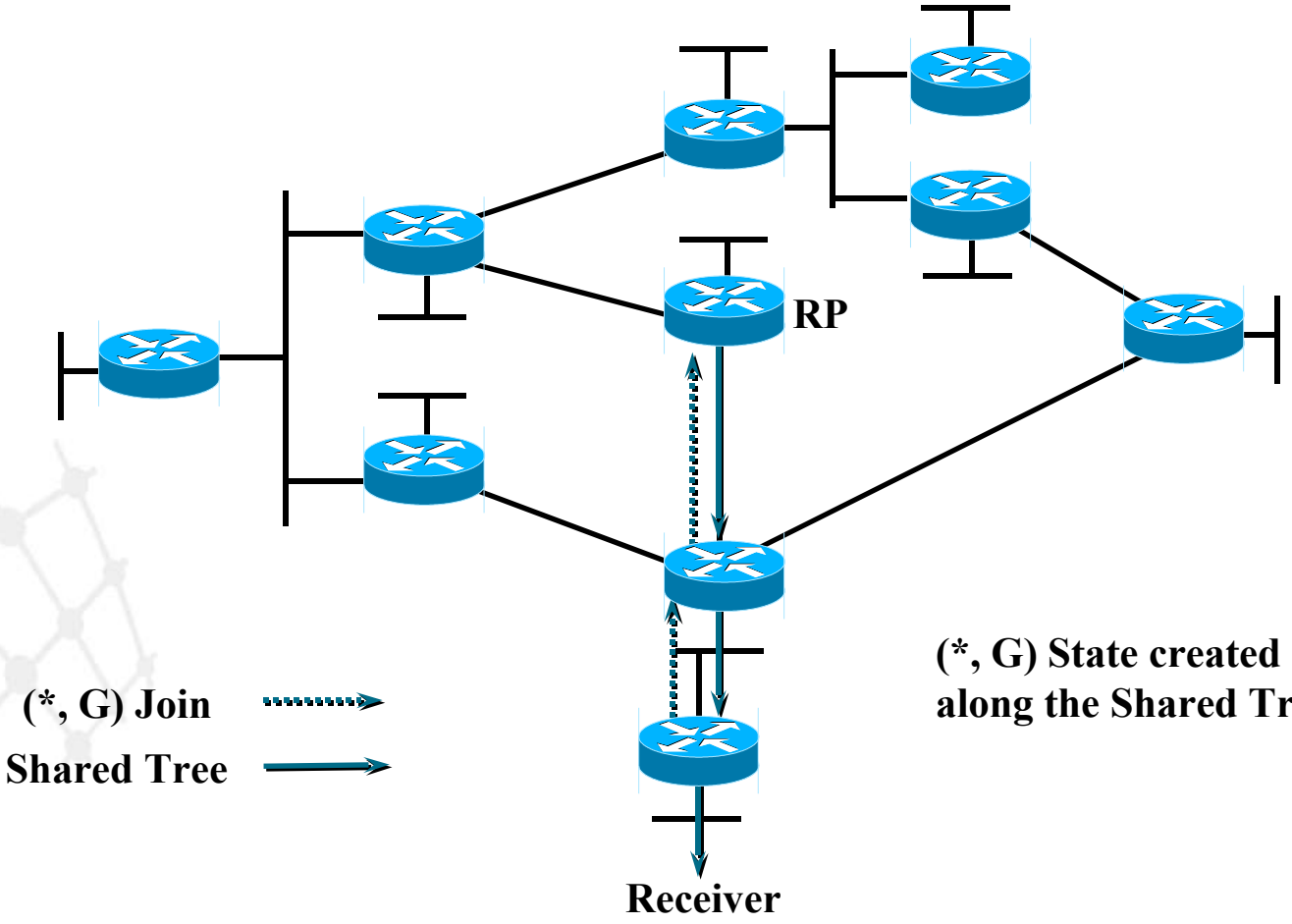
- Matches senders with receivers
- Provides network source discovery
- Root of shared tree

Typically use shared tree to bootstrap source tree

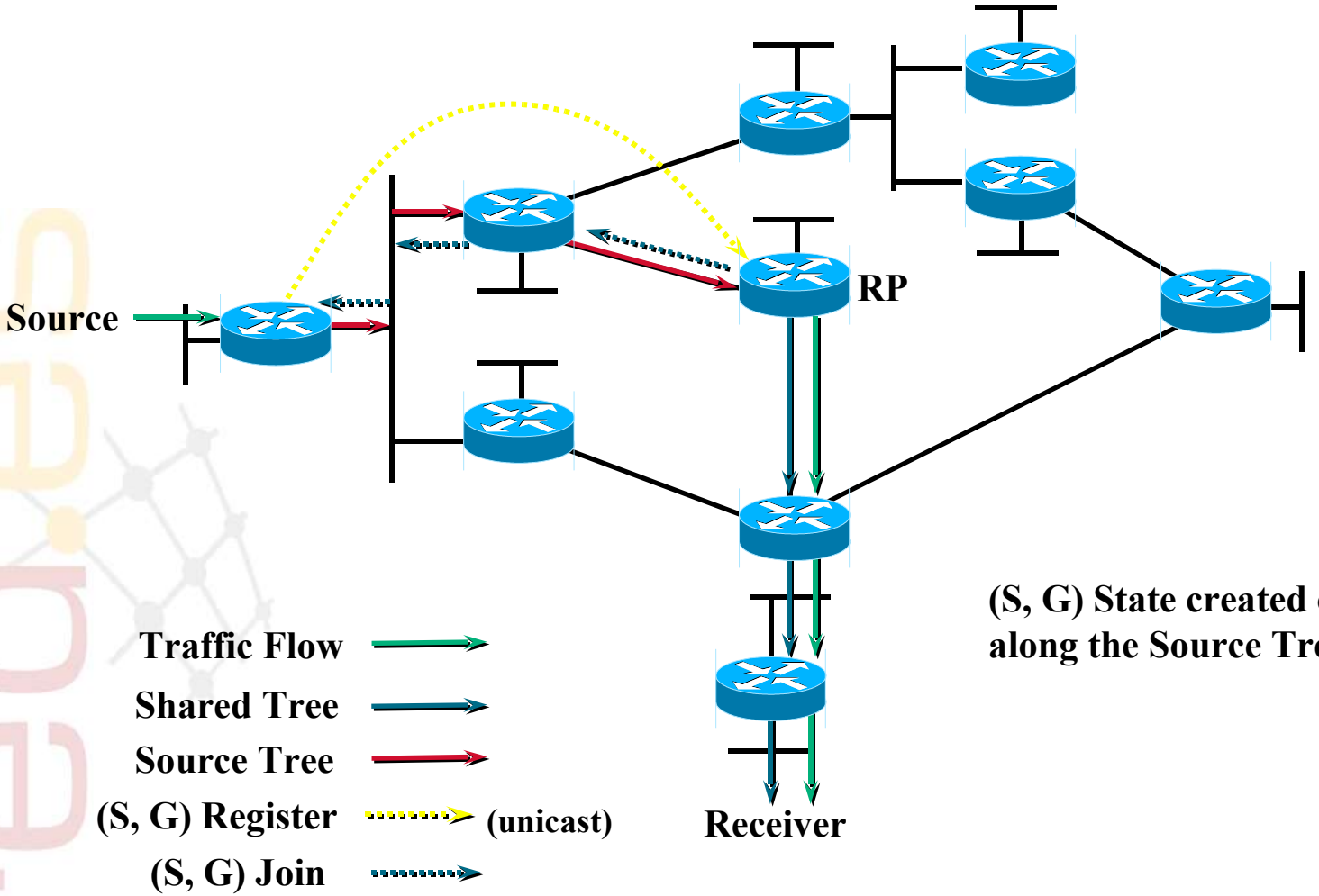
RP's can be learned via:

- Static configuration – RECOMMENDED
- Auto-RP (V1 & V2)
- Bootstrap Router (V2)

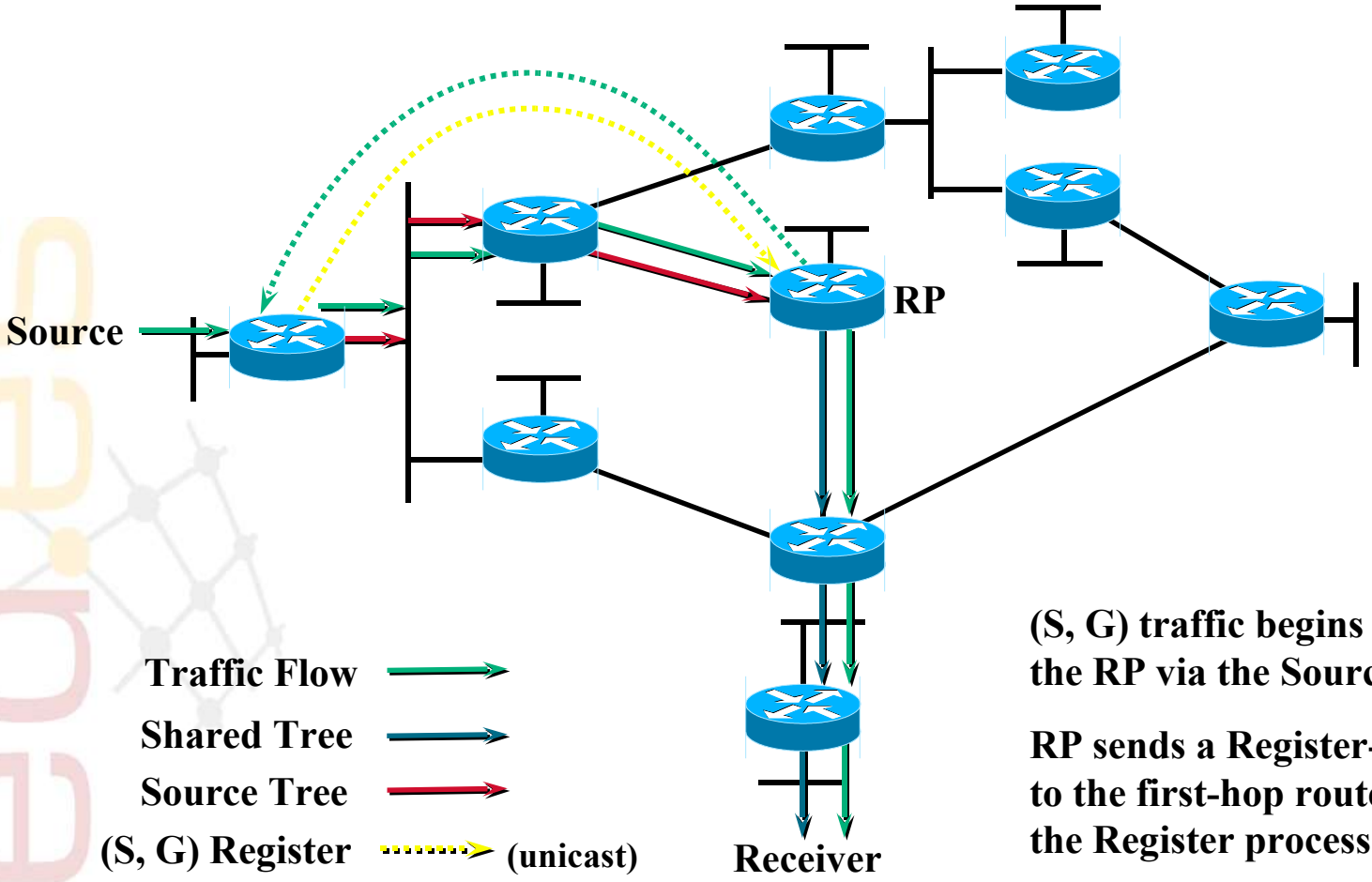
PIM-SM Shared Tree Join



PIM-SM Sender Registration



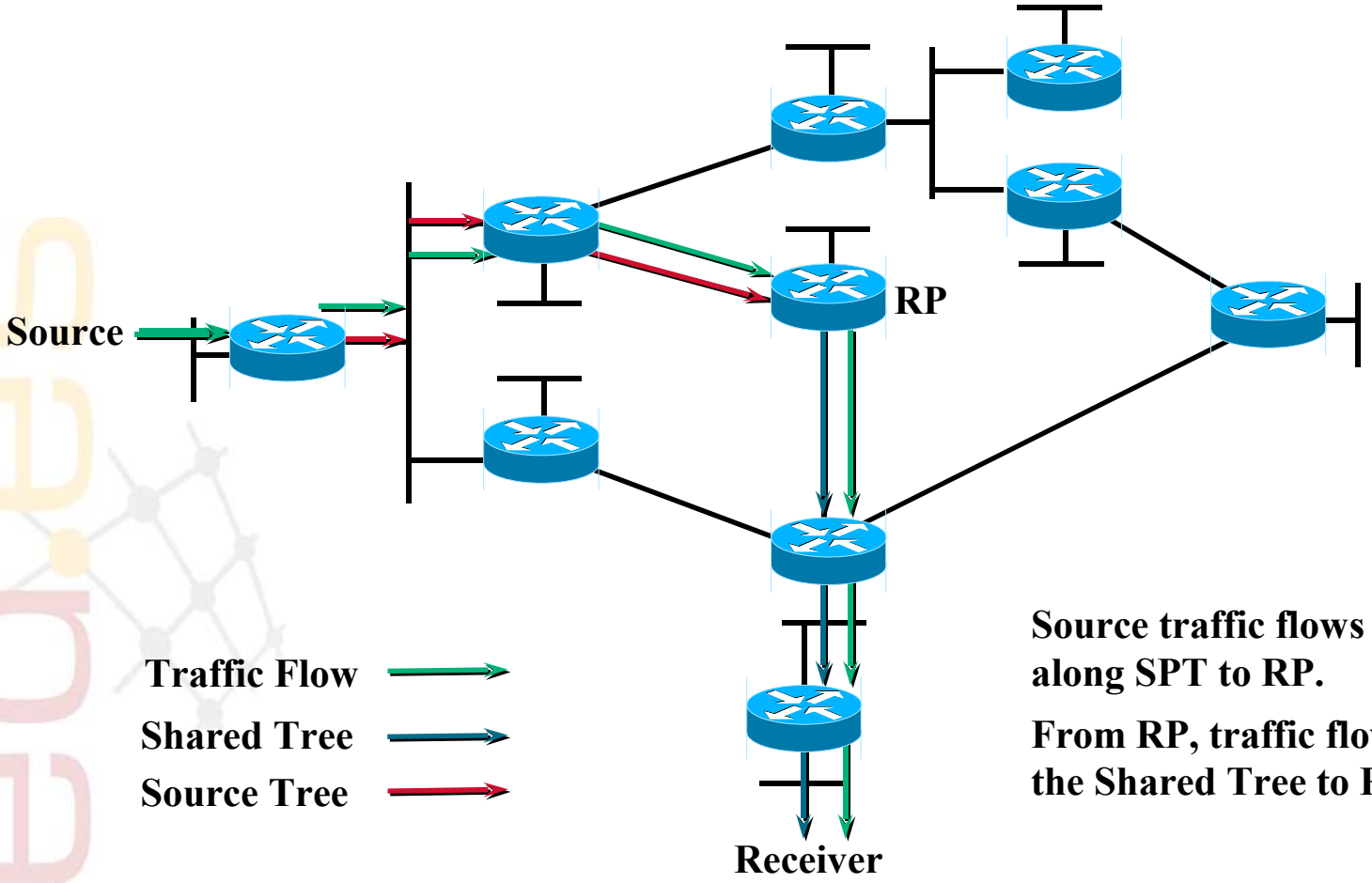
PIM-SM Sender Registration



(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.

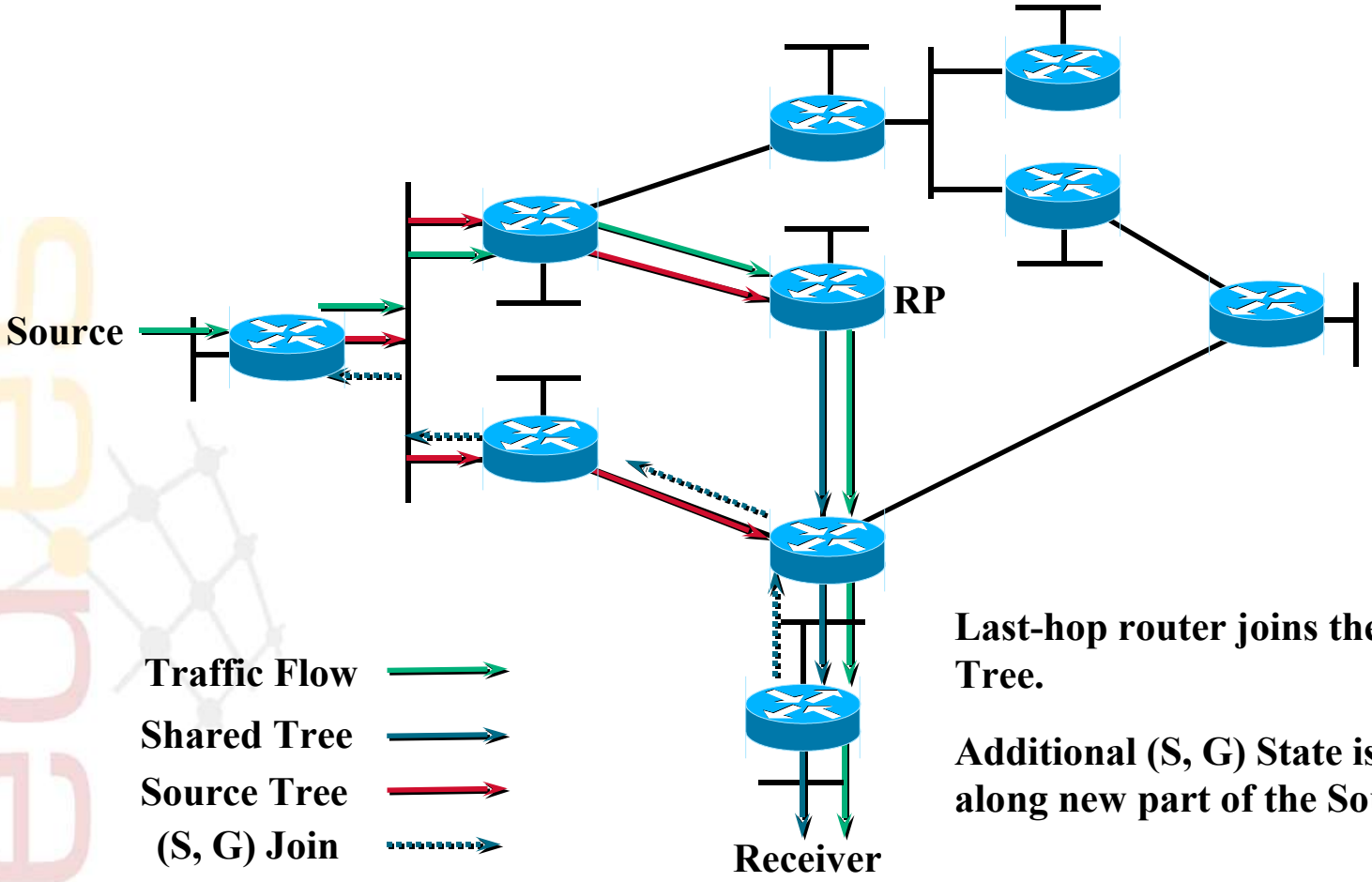
PIM-SM Sender Registration



Source traffic flows natively along SPT to RP.

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

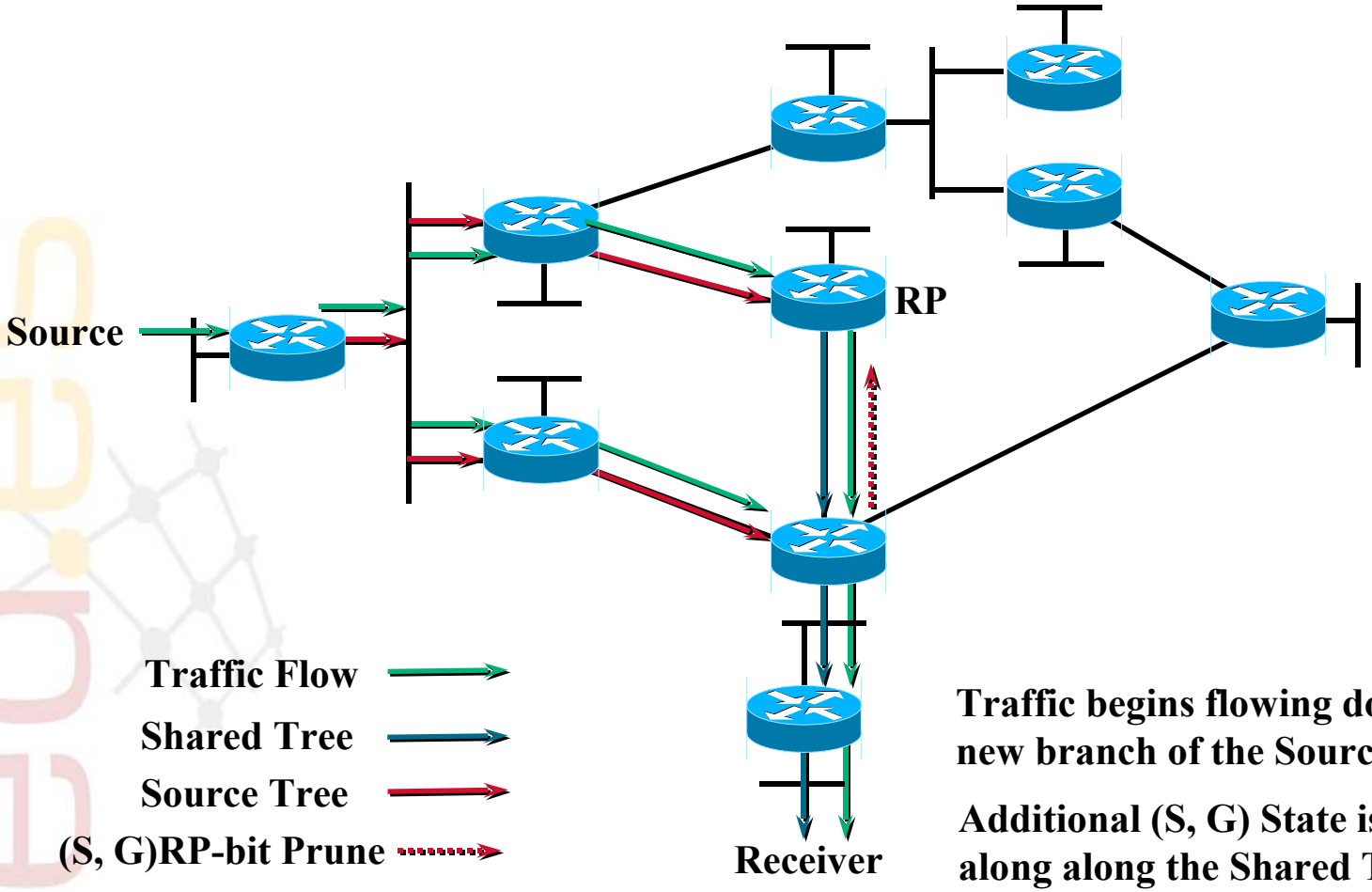
PIM-SM SPT Switchover



Last-hop router joins the Source Tree.

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

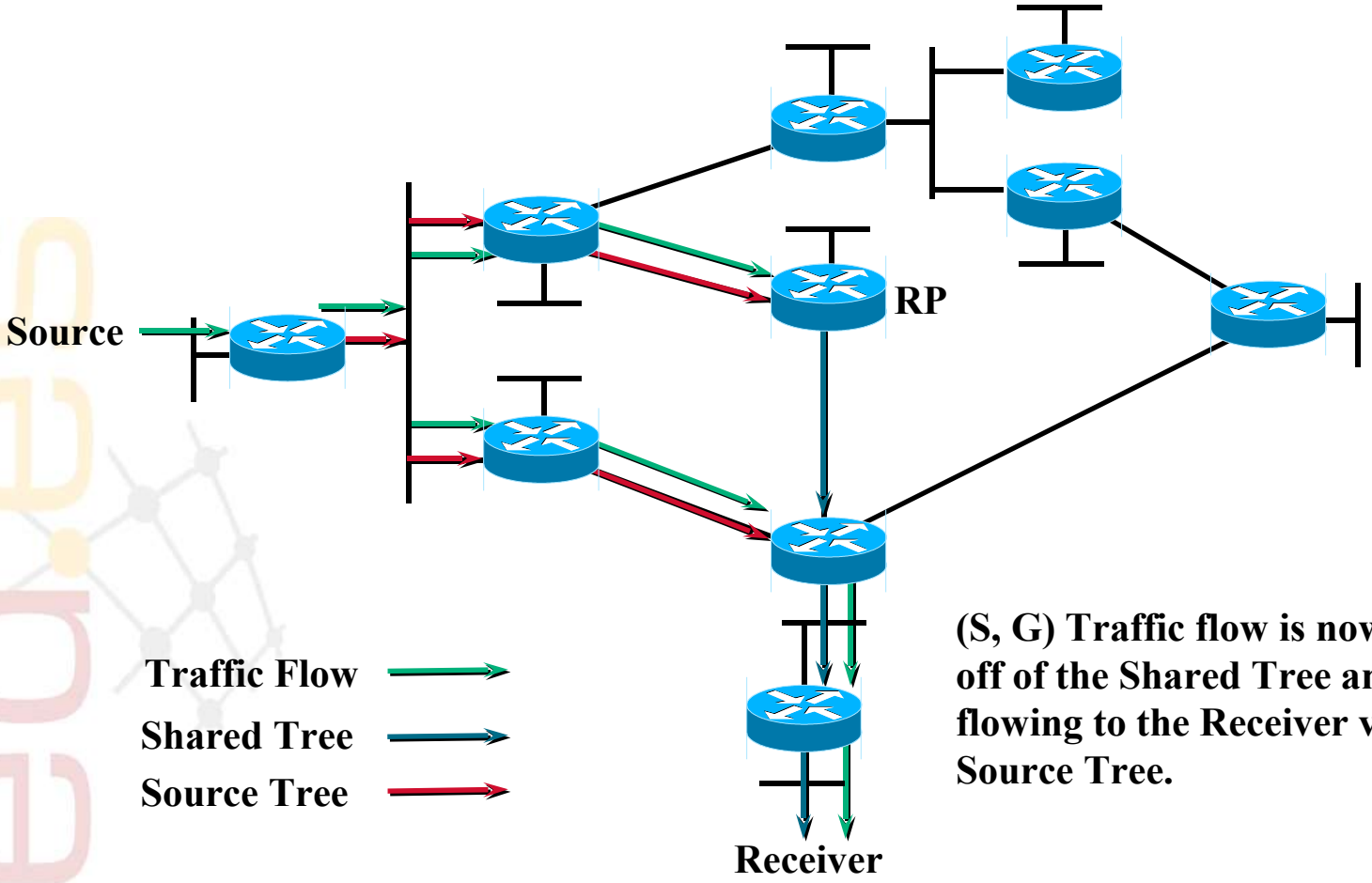
PIM-SM SPT Switchover



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

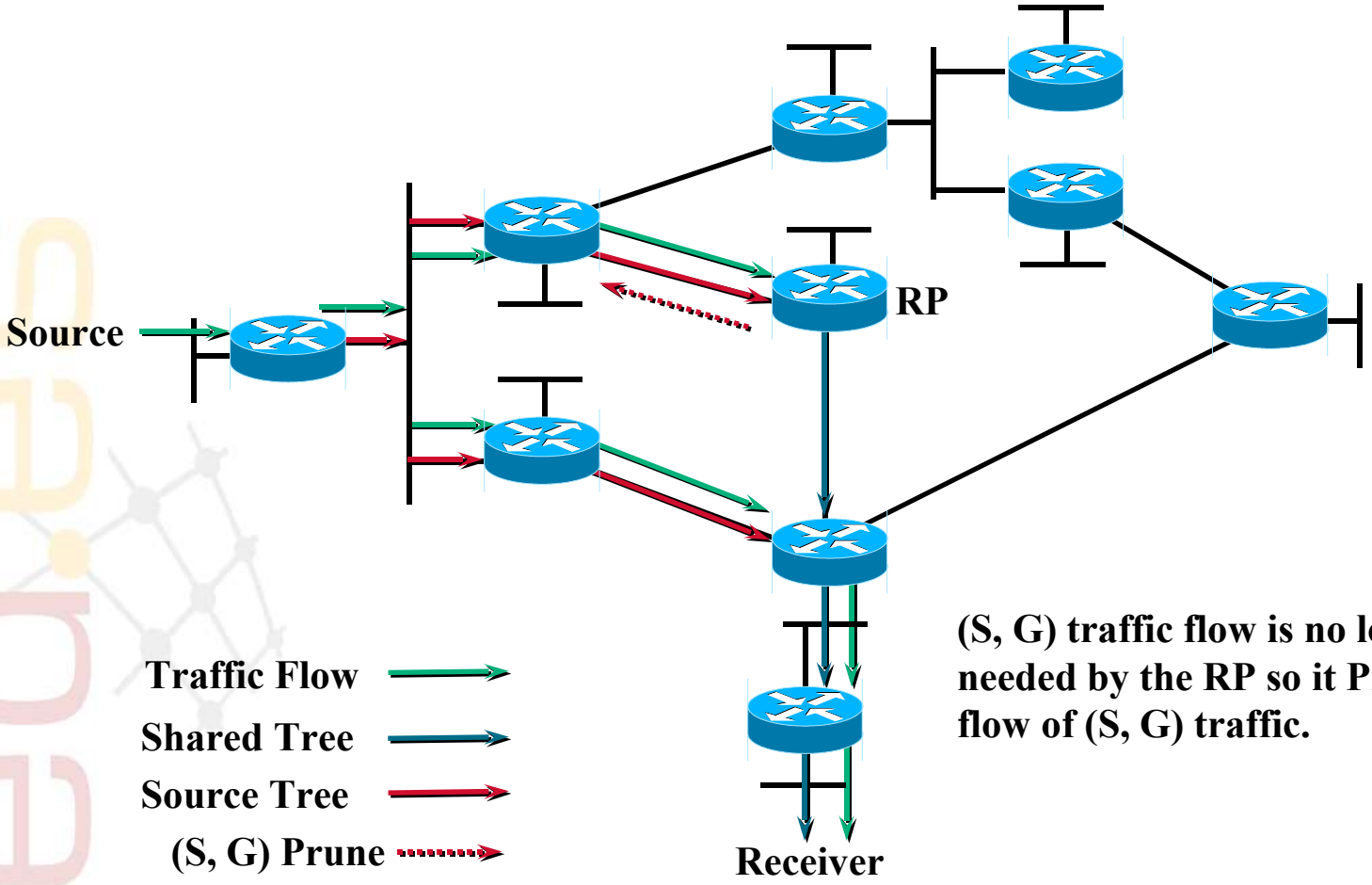
Additional (S, G) State is created along along the Shared Tree to prune off (S, G) traffic.

PIM-SM SPT Switchover



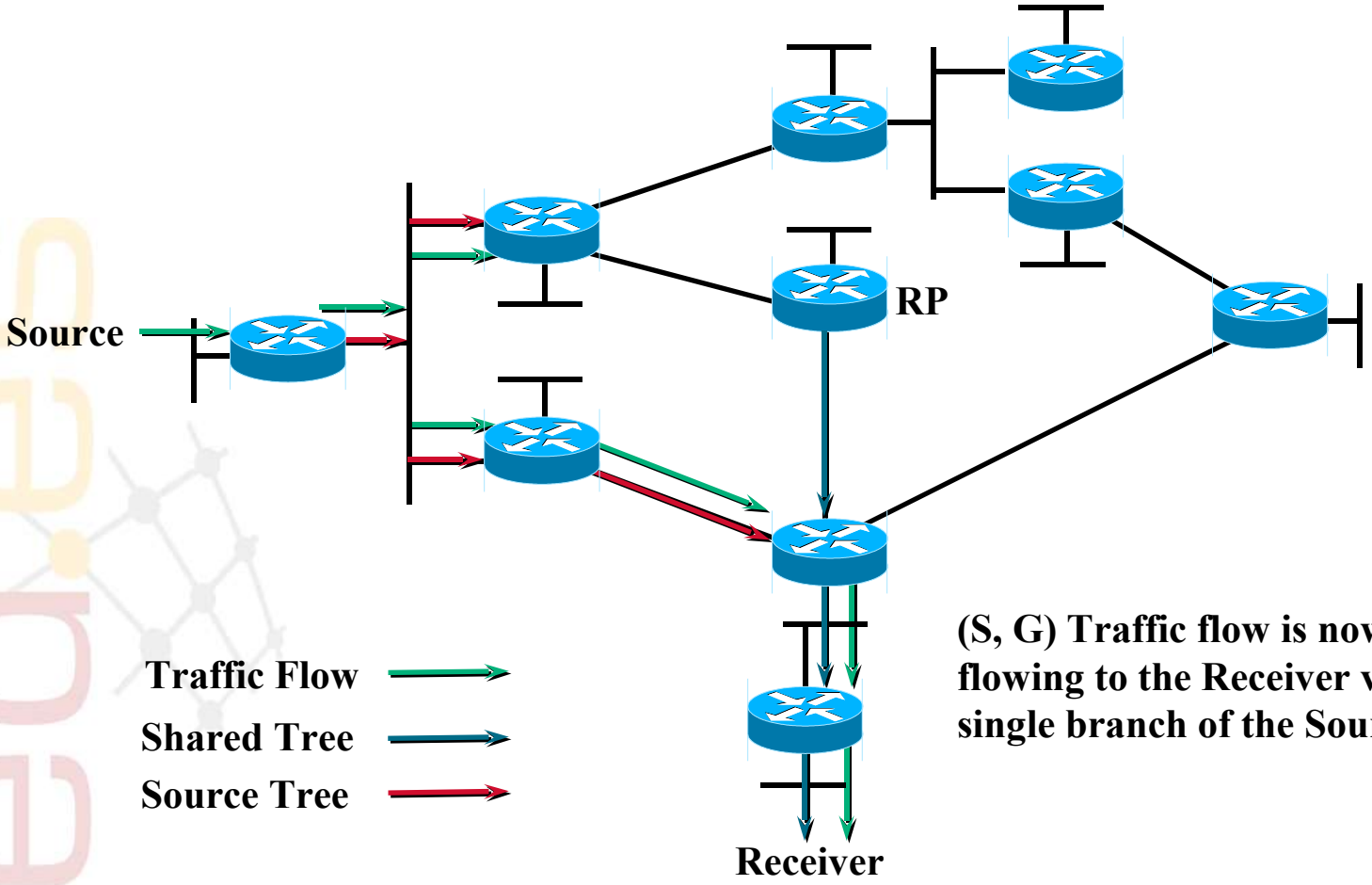
(S, G) Traffic flow is now pruned off of the Shared Tree and is flowing to the Receiver via the Source Tree.

PIM-SM SPT Switchover



(S, G) traffic flow is no longer needed by the RP so it Prunes the flow of (S, G) traffic.

PIM-SM SPT Switchover



(S, G) Traffic flow is now only flowing to the Receiver via a single branch of the Source Tree.

RP Mapping options

- Static RP
  - Recommended
  - Easy transition to Anycast-RP
  - Allows for a hierarchy of RPs
- Auto-RP
  - Fixed convergence timers (slow)
  - Must flood RP mapping traffic
- BSR
  - No longer in the PIM spec.
  - Fixed convergence timers (slow)
  - Allows for a hierarchy of RPs

PIM-SSM

No shared trees

No register packets

No RP required

No RP-to-RP source discovery (MSDP)

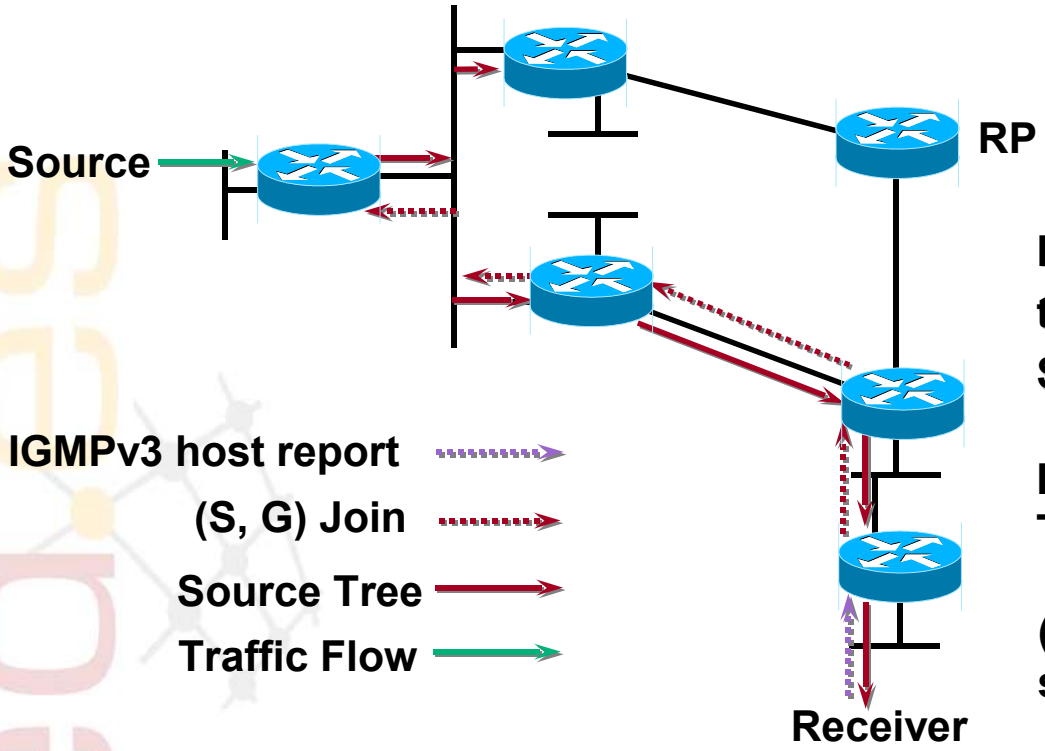
Requires IGMP include-source list – IGMPv3

- Host must learn of source address out-of-band (web page)
- Requires host-to-router source AND group request

Hard-coded behavior in 232/8

- Configurable to expand range

PIM-SSM

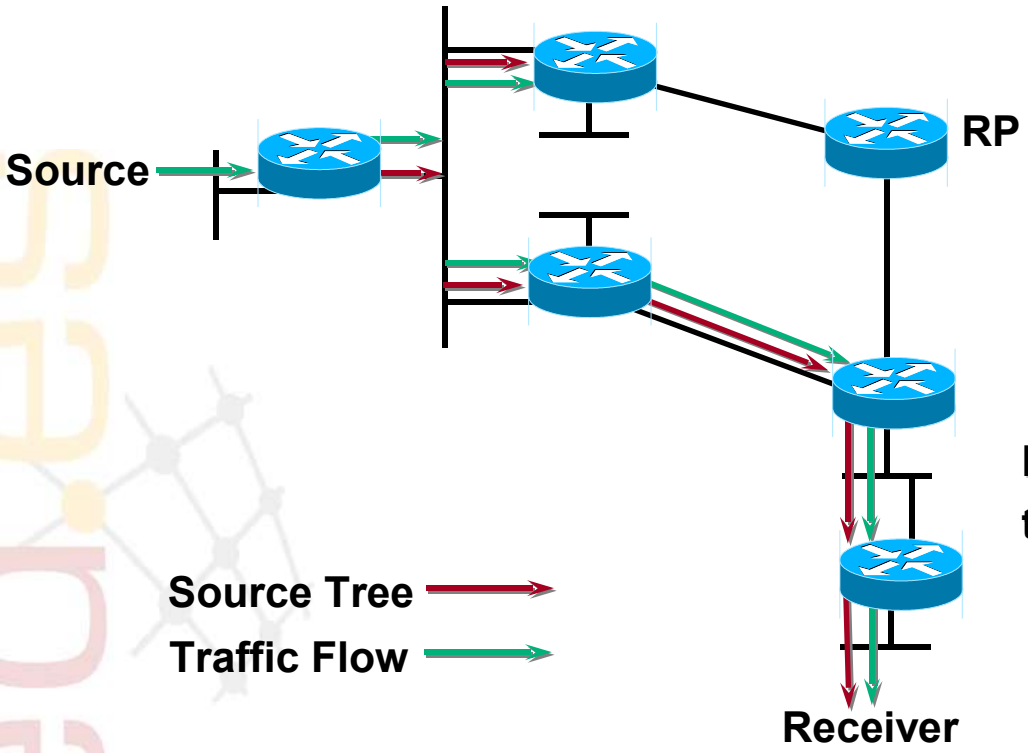


Receiver announces desire to join group G AND source S with an IGMPv3 include-list.

Last-hop router joins the Source Tree.

(S,G) state is built between the source and the receiver.

PIM-SSM



Data flows down the source tree to the receiver.

- Introduction
- Multicast addressing
- Group Membership Protocol
- PIM-SM / SSM
- MSDP
- MBGP

Multicast Source Discovery Protocol

- RFC 3618
- Allows each domain to control its own RP(s)
- Interconnect RPs between domains with TCP connections to pass source active messages (SAs)
- Can also be used within a domain to provide RP redundancy (Anycast-RP)
- RPs send SA messages for internal sources to MSDP peers
- SAs are Peer-RPF checked before accepting or forwarding
- RPs learn about external sources via SA messages and may trigger (S,G) joins on behalf of local receivers
- MSDP connections typically parallel MBGP connections

MSDP Operation

MSDP peers (inter or intra domain)

- (TCP port 639 with higher IP addr LISTENS)

“FLOOD & join”

- SA (source active) packets periodically sent to MSDP peers indicating:
  - source address of active streams
  - group address of active streams
  - IP address of RP originating the SA
- only originate SA’s for its sources within its domain
- interested parties can send PIM JOIN’s towards source (creates inter-domain source trees)

MSDP Source Active Messages

Initial SA message sent when source first registers

- May optionally encapsulate first data packet

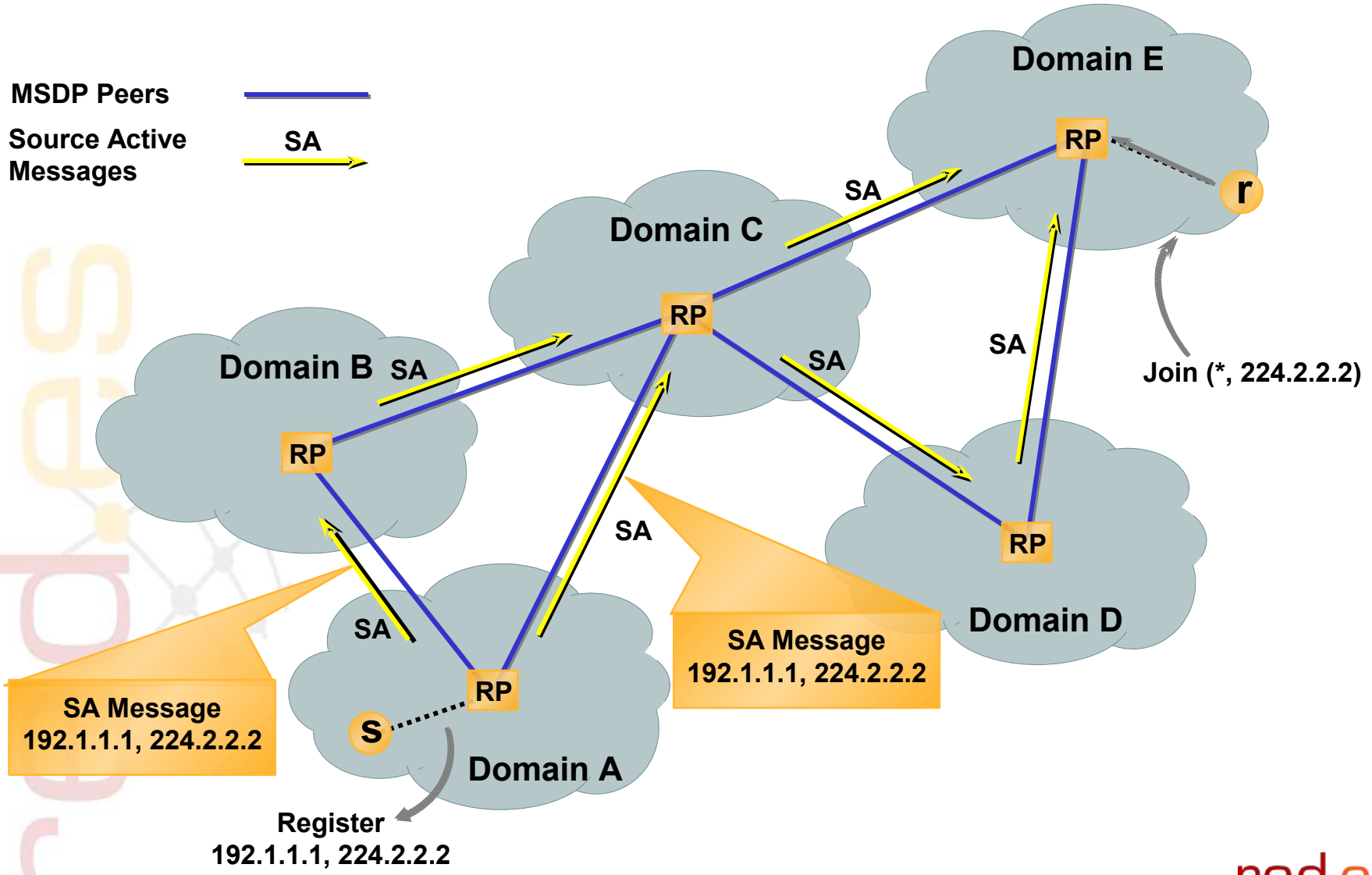
Subsequent SA messages periodically refreshed every 30 seconds as long as source still active by originating RP

Other MSDP peers don't originate this SA but only forward it if received

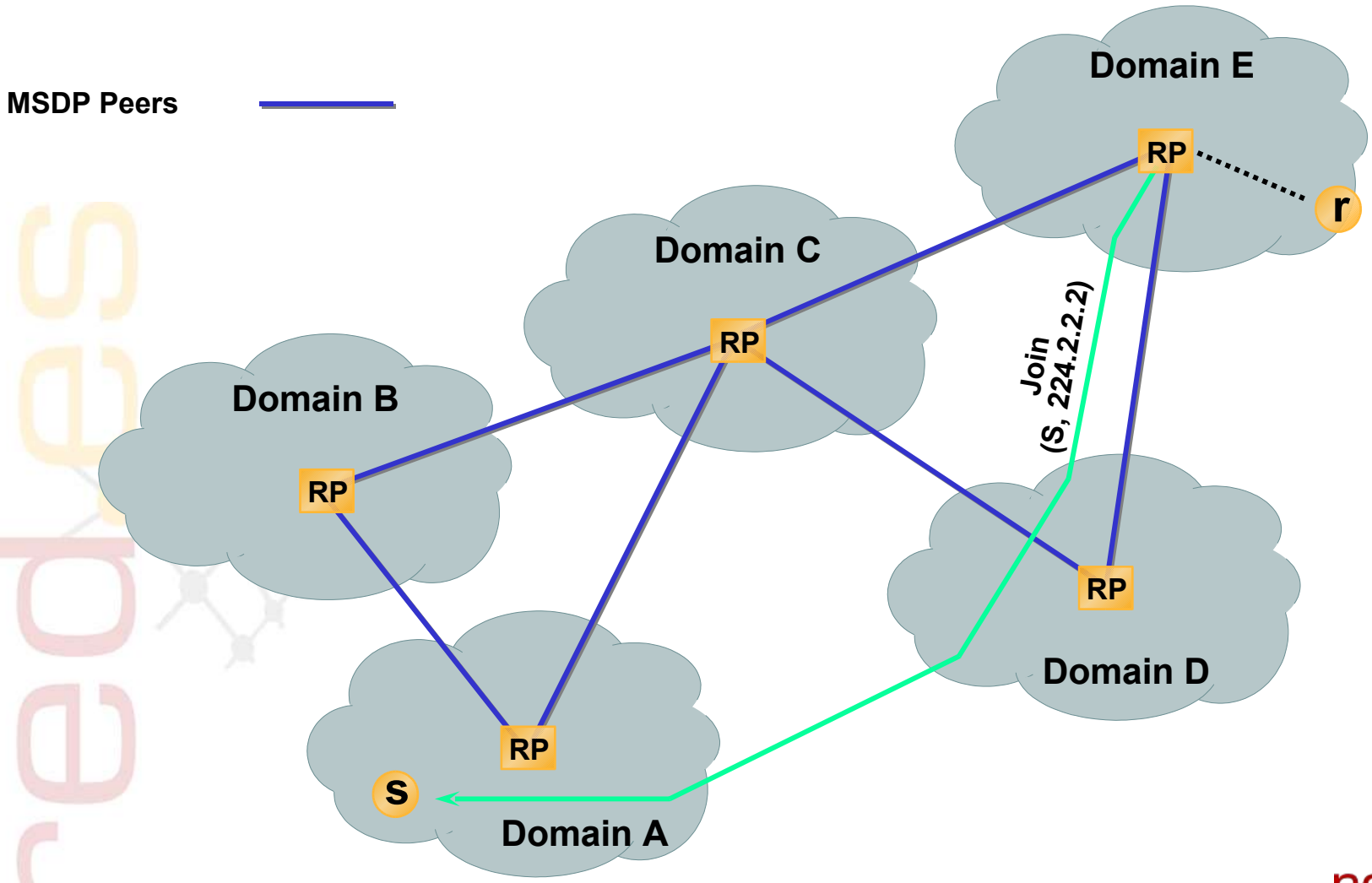
SA messages cached on router for new group members that may join

- Reduced join latency
- Prevent SA storm propagation

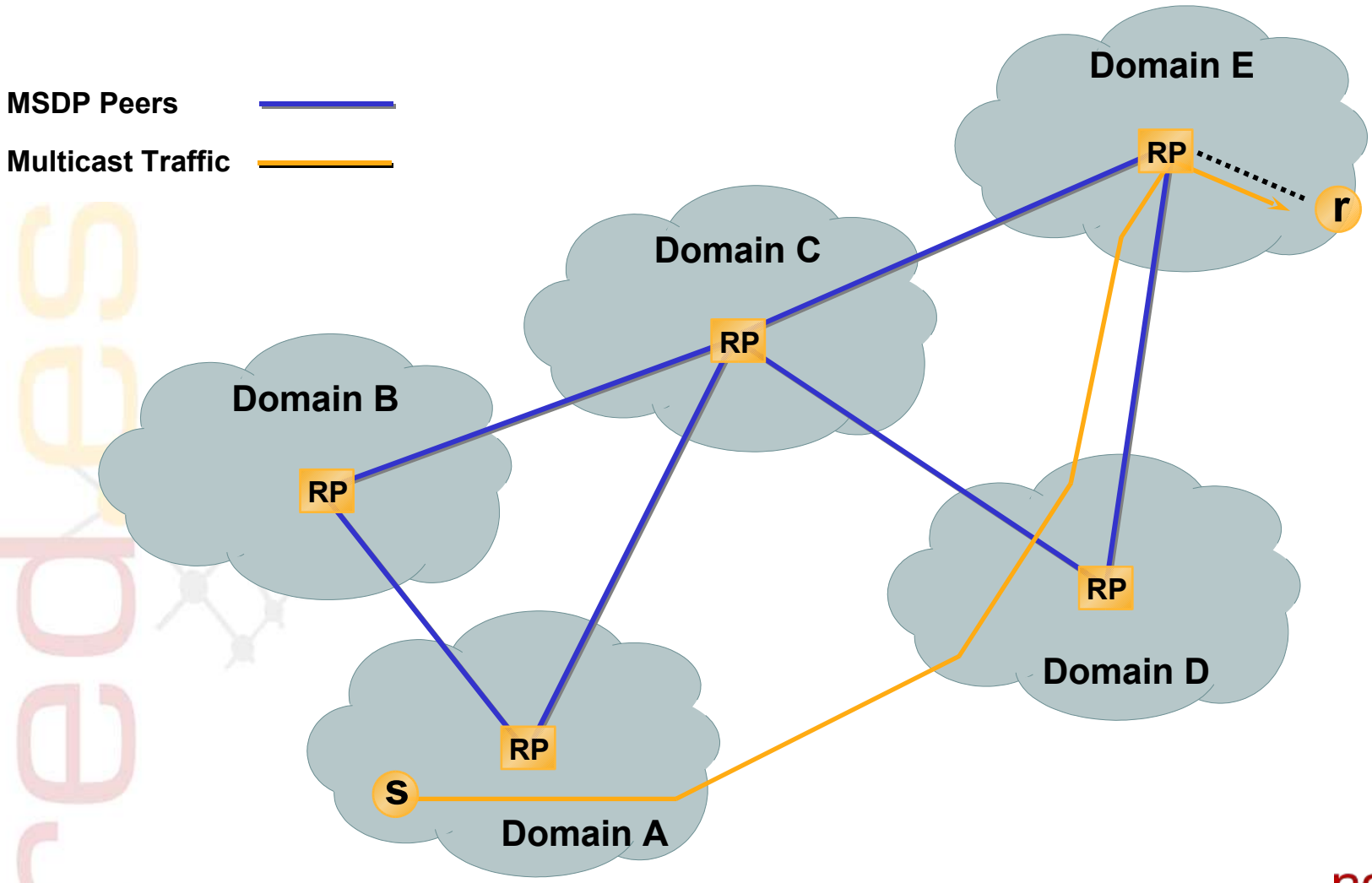
MSDP Overview



MSDP Overview



MSDP Overview



MSDP Peers

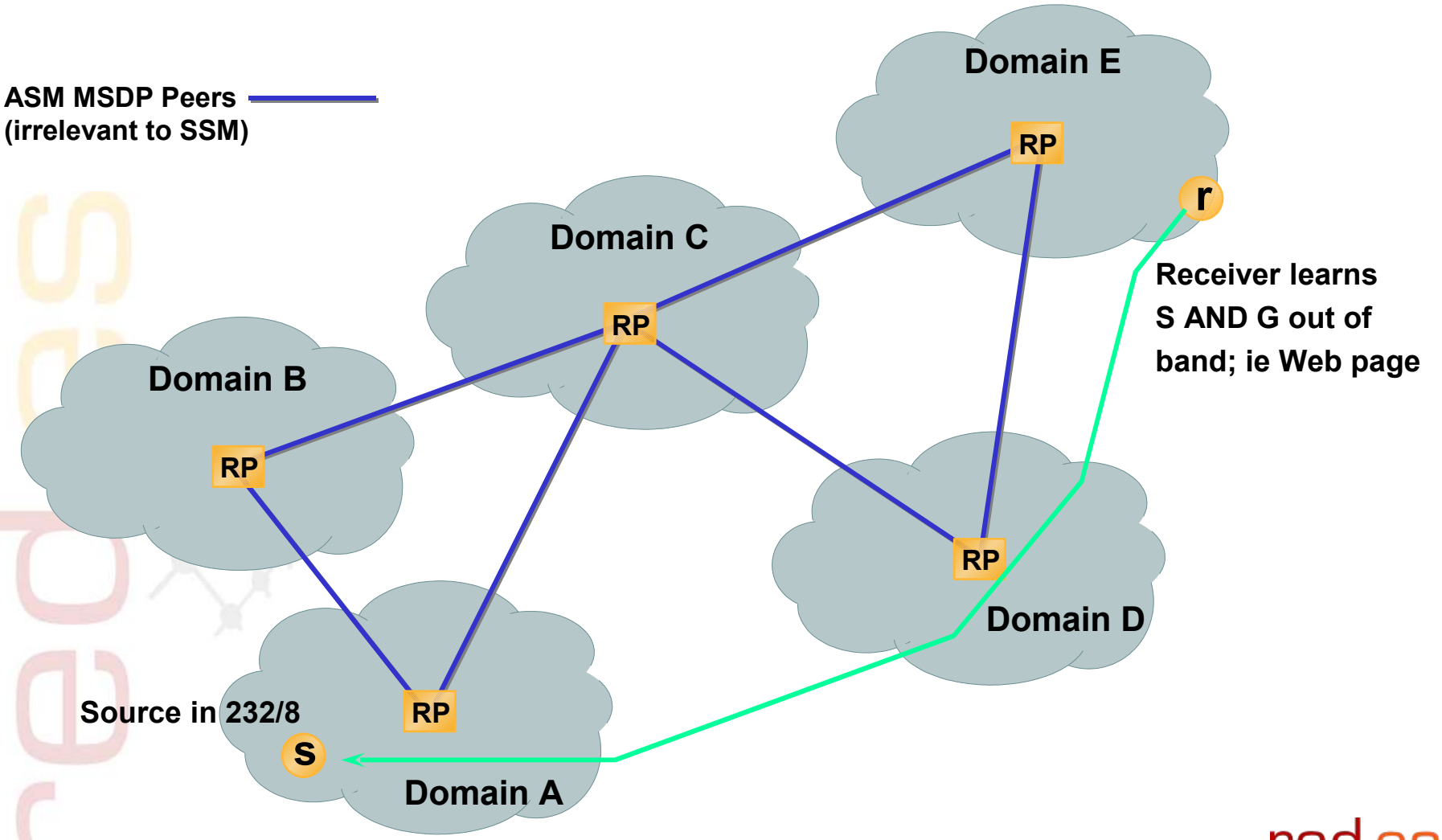
- **MSDP establishes a neighbor relationship between MSDP peers**
  - Peers connect using TCP port 639
- **MSDP peers may run mBGP**
  - May be an MBGP peer, a BGP peer or both
  - Required for peer-RPF checking of the RP address in the SA to prevent SA looping
  - Exception: BGP is unnecessary when peering with only a single MSDP peer (default-peer)

RPF-peer Rules

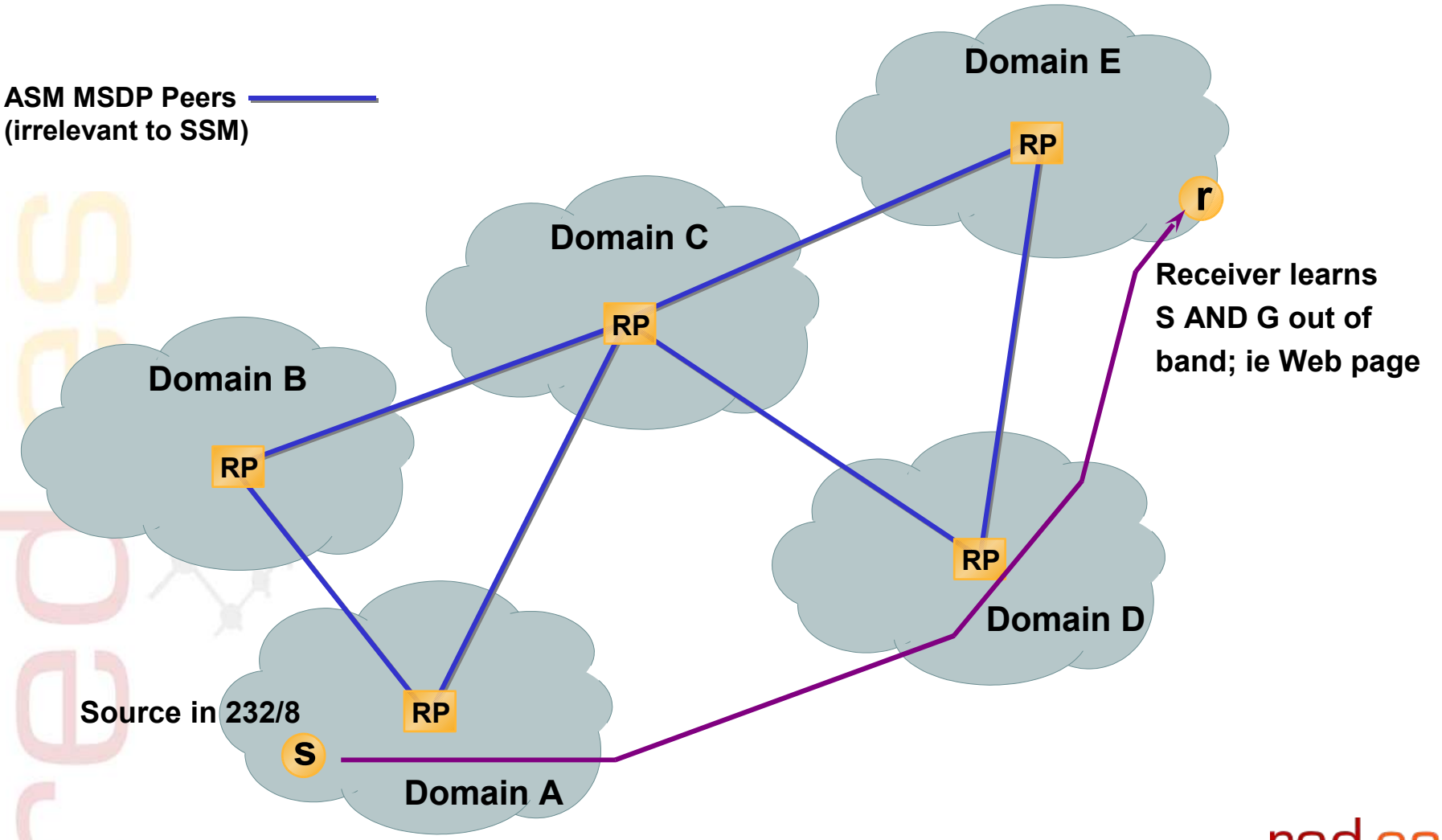
- **Skip RPF Check and accept SA if:**
  - Sending MSDP peer is default-peer
  - Sending MSDP peer = Mesh-Group peer

- **Otherwise, being a MSDP peer, the RPF-peer will be:**
  - The originating RP.
  - The eBGP next-hop toward the originating RP.
  - The iBGP peer that advertise the route or is the IGP next-hop toward the originating RP.
  - The one with the highest IP address of all the MSDP peers in the AS path toward the originating RP.
  - The static RPF-peer.

MSDP with SSM – Unnecessary!



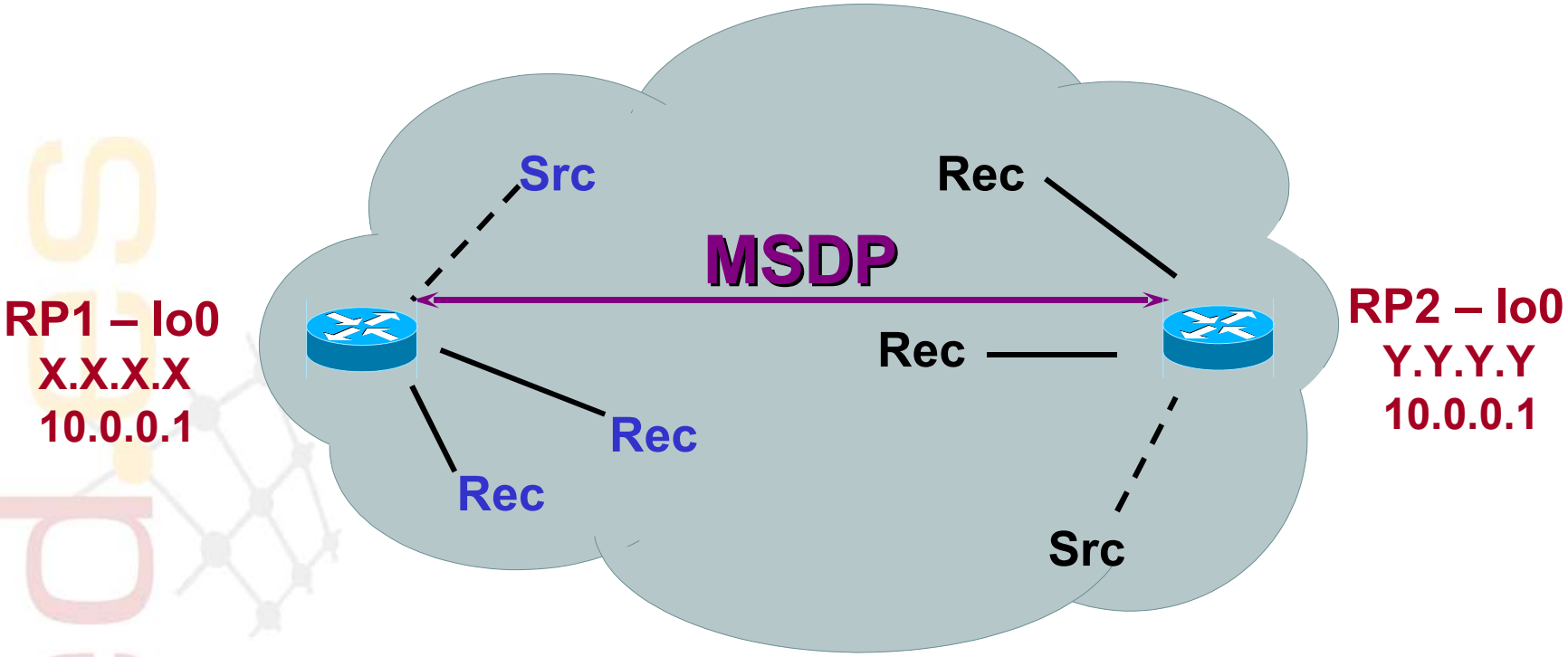
MSDP with SSM – Unnecessary!



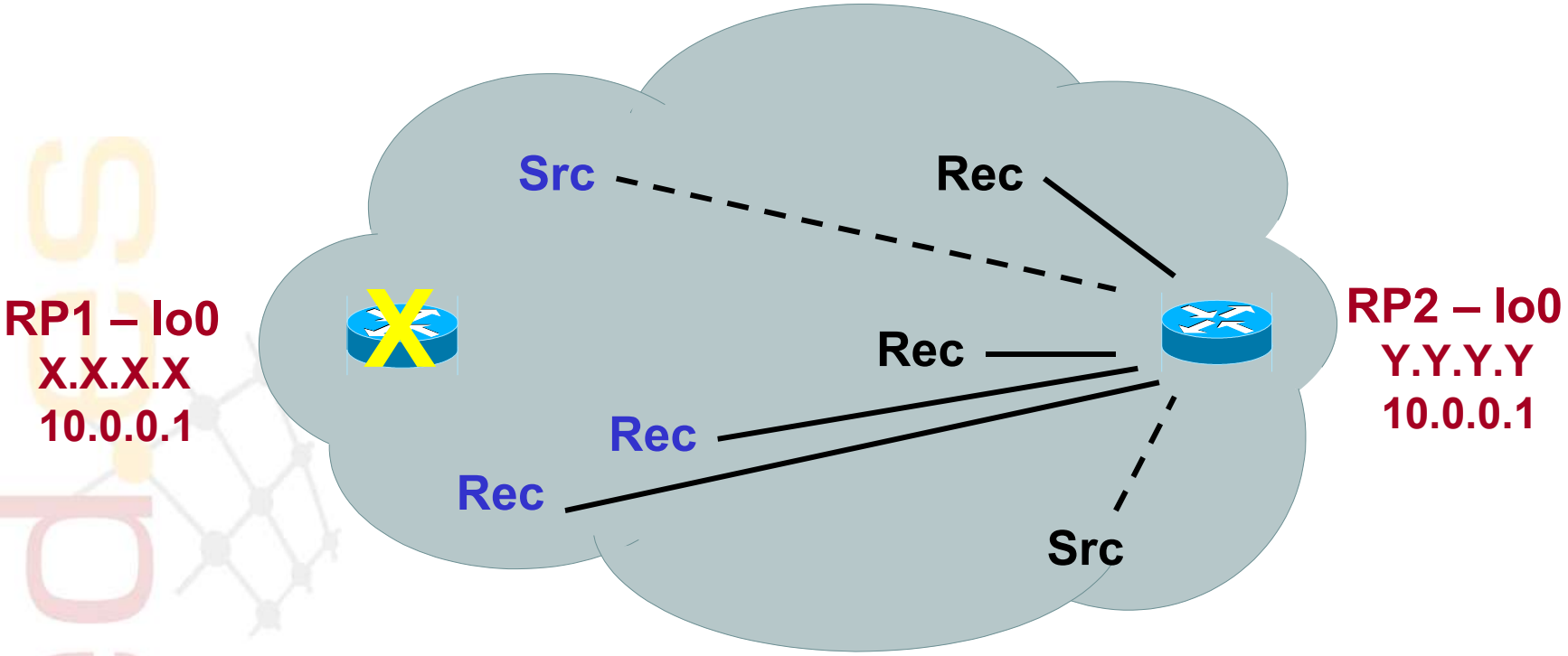
MSDP Application: Anycast-RP

- RFC 3446
- Within a domain, deploy more than one RP for the same group range
- Sources from one RP are known to other RPs using MSDP
- Give each RP the same /32 IP address
- Sources and receivers use closest RP, as determined by the IGP
- Used intra-domain to provide redundancy and RP load sharing, when an RP goes down, sources and receivers are taken to new RP via unicast routing
  - Fast convergence!

Anycast-RP


RP1 – lo0
X.X.X.X
10.0.0.1
Src
Rec
MSDP
Rec
Rec
Rec
Src
RP2 – lo0
Y.Y.Y.Y
10.0.0.1

Anycast-RP



- Introduction
- Multicast addressing
- Group Membership Protocol
- PIM-SM / SSM
- MSDP
- MBGP

MBGP Overview

- **Multiprotocol Extensions to BGP (RFC 2858).**
- **Tag unicast prefixes as multicast source prefixes for intra-domain mcast routing protocols to do RPF checks.**
- **WHY? Allows for interdomain RPF checking where unicast and multicast paths are non-congruent.**
- **DO I REALLY NEED IT?**
  - YES, if:
    - ISP to ISP peering
    - Multiple-homed networks
  - NO, if:
    - You are single-homed

MBGP Overview

- **MBGP: Multiprotocol BGP (multicast BGP in multicast networks)**
  - Defined in RFC 2858 (extensions to BGP)
  - Can carry different route types for different purposes
    - Unicast
    - Multicast
  - Both route types carried in same BGP session
  - Does not propagate multicast state information
  - Same path selection and validation rules
    - AS-Path, LocalPref, MED, …

MBGP Overview

- **New multiprotocol attributes:**
  - MP_REACH_NLRI
    - Used to advertise one or more routes to a peer that shares the same path attribute
  - MP_UNREACH_NLRI
    - Used to indicate a previously route is no longer reachable
- **They include the next information:**
  - Address Family Information (AFI) = 1 (IPv4)
    - Sub-AFI = 1 (NLRI is used for unicast)
    - Sub-AFI = 2 (NLRI is used for multicast RPF check)
    - Sub-AFI = 3 (NLRI is used for both unicast and multicast RPF check)
- **This information is used to build routing tables**
- **Allows different policies and topologies between multicast and unicast**

MBGP—Capability Negotiation

- **RFC 2842**

- **BGP routers establish BGP sessions through the OPEN message**
  - OPEN message contains optional parameters
  - BGP session is terminated if OPEN parameters are not recognised

- **MBGP peers use this procedure to determine if they support MBGP and which AFIs and SAFIs support each one**
  - If there is no match, notification is sent and peering doesn't come up
  - If neighbor doesn't include the capability parameters in open, session backs off and reopens with no capability parameters
    - Peering comes up in unicast-only mode

Summary

- **IGMP** - Internet Group Management Protocol is used by hosts and routers to tell each other about group membership.

- **PIM-SM** - Protocol Independent Multicast-Sparse Mode is used to propagate forwarding state between routers.

- **SSM** - Source Specific Multicast utilizes a subset of PIM?s functionality to guaranty source-only trees in the 232/8 range.

- **MBGP** - Multiprotocol Border Gateway Protocol is used to exchange routing information for interdomain RPF checking.

- **MSDP** - Multicast Source Discovery Protocol is used to exchange ASM active source information between RPs.

ISP Requirements

- **Current solution: MBGP + PIM-SM + MSDP**
  - Environment
    - ISPs run iMBGP and PIM-SM (internally)
    - ISPs multicast peer at a public interconnect
  - Deployment
    - Border routers run eMBGP
    - The interfaces on interconnect run PIM-SM
    - RPs' MSDP peering must be consistant with eMBGP peering
    - All peers set a common distance for eMBGP