

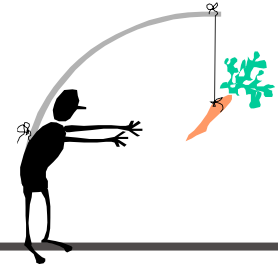
# Multicast

Tópicos Avançados em Redes  
2023/2024

# Sources

- Kevin Almeroth, *Introduction to Multicast Protocols and Applications*
- CS 640 , University of Wisconsin
- Cisco presentation *Introduction to IP Multicast* and others by M. McBride, D. Meyer, T. Bryant
- *Computer Networking: A Top Down Approach*, 5th edition. Jim Kurose, Keith Ross Addison-Wesley, April 2009
- Others...

# Why multicast?

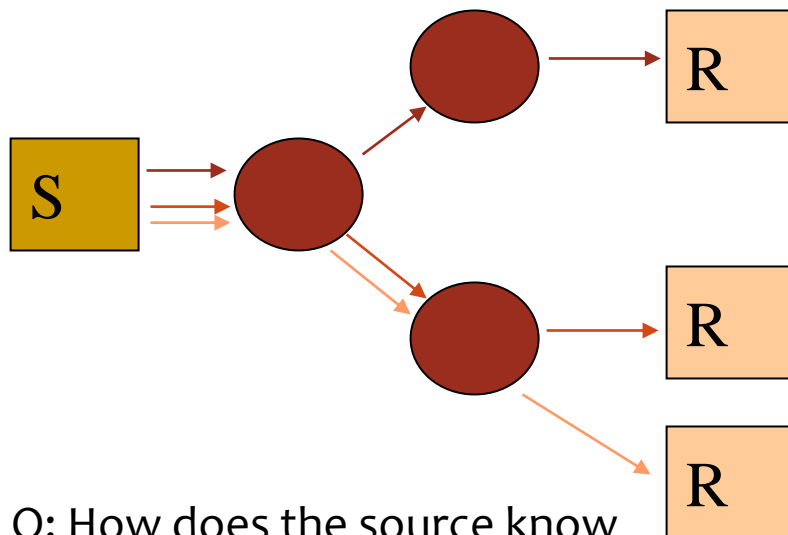


- Need to send the same information to multiple receivers
  - Better use of network capacity
  - Less processing at routers and sources
  - Quicker participation
- Multicast applications
  - Audio and video diffusion
  - Software distribution
  - Web caches updates (e.g., CDN nodes)
  - Teleconference (audio, video, whiteboard, etc.)
  - Games
  - Service or server discovery
  - Other distributed apps

# One to many communication

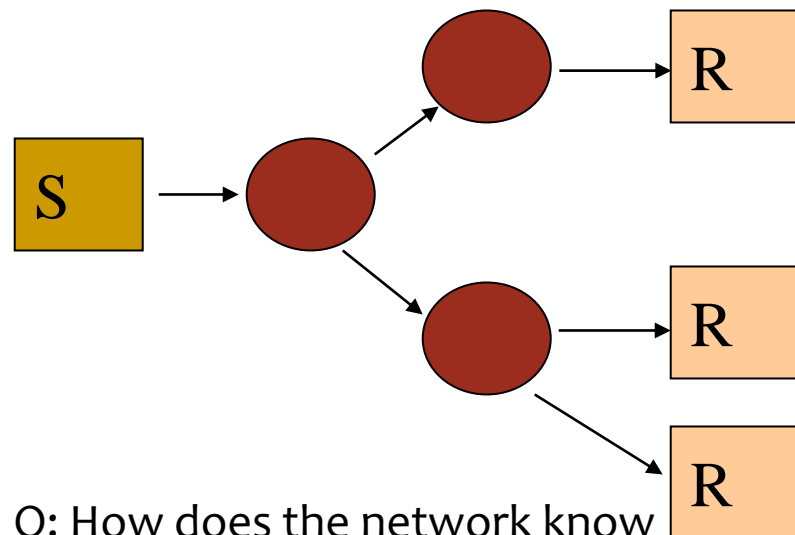
Source [2]

- Multiple unicasts



Q: How does the source know about receivers' addresses?

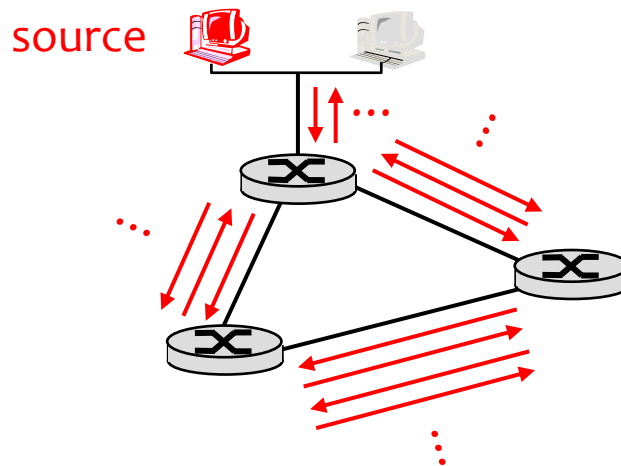
- In-network duplication: multicast/broadcast



Q: How does the network know where to forward the packets?

# In-network duplication

- Simple flooding: when node receives broadcast packet, sends copy to all neighbours
  - Problem: network cycles create broadcast storms

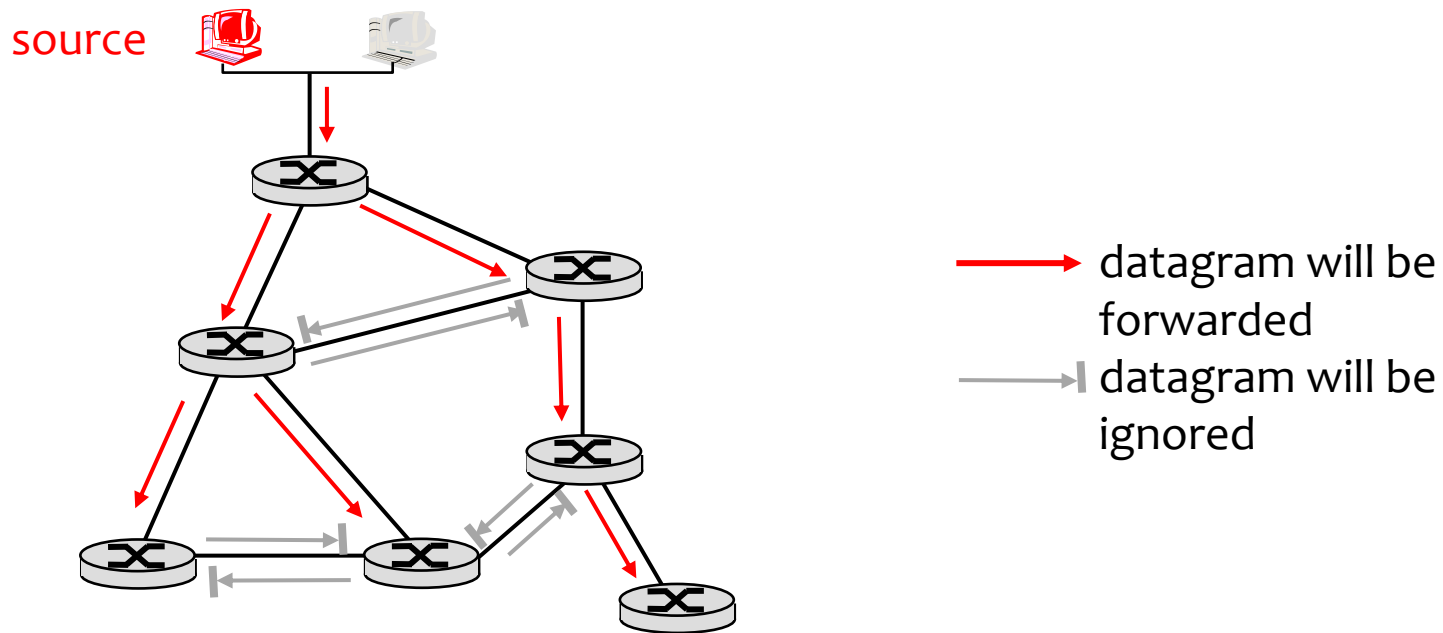


# In-network duplication

- Controlled flooding: node only broadcasts packets it hasn't broadcast before
  - Nodes keeps track of packets already broadcast (e.g., using per-source sequence numbers/packet IDs)
  - or
  - Reverse path forwarding (RPF) — only forward packets that arrived on shortest path between node and source
- Nodes may still receive redundant copies

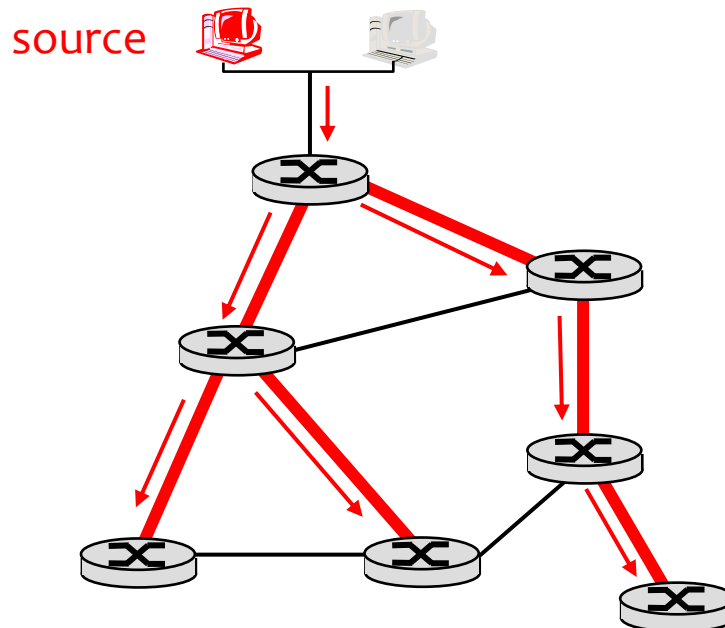
# In-network duplication

- Reverse Path Forwarding example:



# In-network duplication

- Spanning tree
  - Retransmit only on links belonging to the tree
  - Single path between source and any node:  
no redundant packets received by anyone



Typically **not** minimum spanning tree (MST)



# Hardware Multicast

- Broadcast
  - Packets delivered to every host
  - Ethernet: MAC=FF:FF:FF:FF:FF:FF
  - Problem: CPU use in every host
    - Even those not interested





# Hardware Multicast II

- Multicast
  - Each host decides whether or not to receive
  - Network technologies have reserved addresses for multicast
    - Ethernet has 1 bit to indicate if MAC address is unicast or multicast – 01:00:00:00:00:00
    - See [ethernet multicast addresses](#):
      - Ex.: 01:80:C2:00:00:00 – Spanning Tree Protocol
  - Every host that configures the address, receives the packets destined to that address
  - Efficient support needs appropriate hardware
    - Chips without multicast filtering use promiscuous mode

# IP Multicast

- Abstraction of hardware multicast
  - Reach beyond local links
- Multicast groups identified by IP address
  - IPv4: Class D, up to  $2^{28}$  different groups
  - IPv6: 112 bits → up to  $2^{112}$  groups, with defined scope
- Dynamic attachment: nodes can
  - Join/leave groups at any moment
  - Join an arbitrary number of groups
- Hardware support
  - If hardware supports multicast, make use of it
  - Otherwise, simulate with unicast or broadcast

# IP multicast IPv4 addresses

- Class D addresses
  - range: 224.0.0.0 — 239.255.255.255



- Two types
  - Permanent (well-known)
    - Defined by IANA
    - Usually used for control protocols
  - Temporary
    - Dynamic management

# Well known addresses (IANA)

## Local Network Control Block

224.0.0.0 - 224.0.0.255 (224.0.0/24)

-----

224.0.0.0	Base Address (Reserved)
224.0.0.1	All Systems on this Subnet
224.0.0.2	All Routers on this Subnet
224.0.0.3	Unassigned
224.0.0.4	DVMRP Routers
224.0.0.5	OSPF All Routers
224.0.0.6	OSPF Designated Routers
224.0.0.7	ST Routers
224.0.0.8	ST Hosts
224.0.0.9	RIP2 Routers
224.0.0.10	IGRP Routers
224.0.0.11	Mobile-Agents
224.0.0.12	DHCP Server / Relay Agent
224.0.0.13	All PIM Routers
224.0.0.14	RSVP-ENCAPSULATION
224.0.0.15	all-cbt-routers
224.0.0.16	designated-sbm
224.0.0.17	all-sbms
224.0.0.18	VRRP
224.0.0.19	IPAllL1ISs
224.0.0.20	IPAllL2ISs

...

Non routable



RFC 5771

## Internetwork Control Block

224.0.1.0 - 224.0.1.255 (224.0.1/24)

-----

224.0.1.0	VMTP Managers Group
224.0.1.1	NTP Network Time Protocol
224.0.1.2	SGI-Dogfight
224.0.1.3	Rwhod
224.0.1.4	VNP
224.0.1.5	Artificial Horizons - Aviator
224.0.1.6	NSS - Name Service Server
224.0.1.7	AUDIONEWS - Audio News Multi
224.0.1.8	SUN NIS+ Information Service
224.0.1.9	MTP Multicast Transport Prot
224.0.1.10	IETF-1-LOW-AUDIO
224.0.1.11	IETF-1-AUDIO

...

# Well known addresses (IANA)

## Administratively Scoped ([RFC-2365](#))

239.0.0.0 - 239.255.255.255 (239/8)

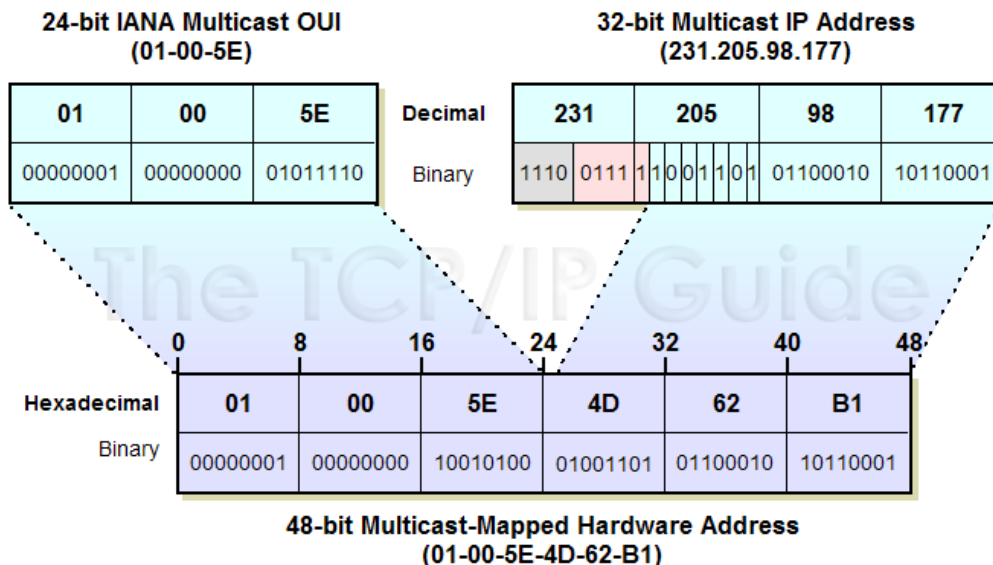
-----  
 239.192.0.0-239.251.255.255 Organization-Local Scope  
 239.255.0.0-239.255.255.255 Site-Local Scope  
 ...

## Others

224.0.2.0-224.0.255.255	Ad-Hoc Block I
224.3.0.0-224.4.255.255	Ad-Hoc Block II
224.233.252.0.0-233.255.255.255	Ad-Hoc Block III
224.2.0.0-224.2.255.255 (224.2/16)	SDP/SAP Block
233.0.0.0-233.255.255.255 (233/8)	GLOP Block (middle octets contain AS number)
...	

# IP multicast **IPv4** addresses

- Semantics
  - Only destination addresses are relevant
  - No ICMP errors for multicast packets
- Mapping to MAC addresses



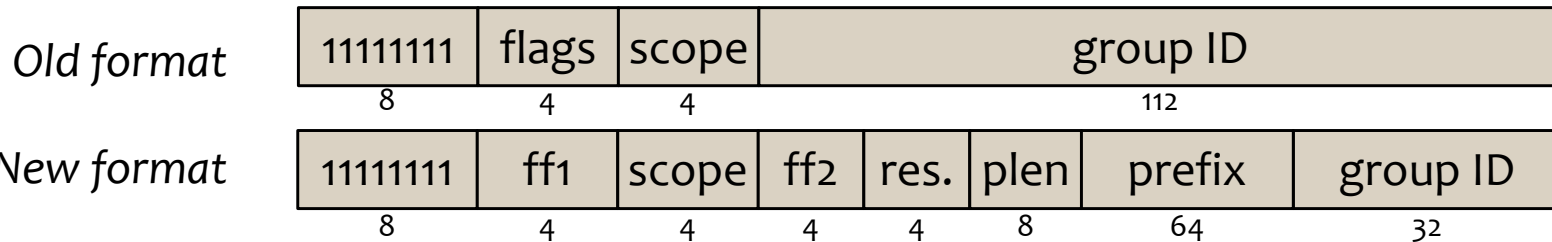
**Non-unique mapping:**  
need to check at IP layer  
if subscribed to that  
multicast address

# Multicast IPv4 Scope (range)

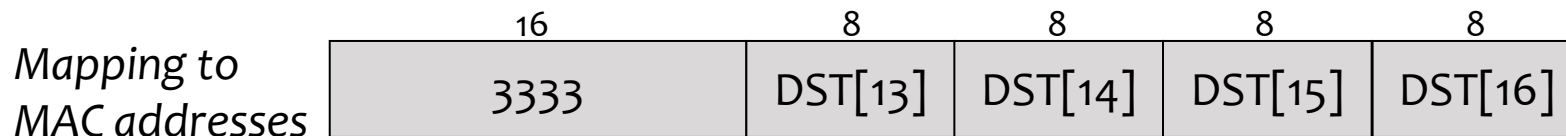
- Local (link)
- Global
- Administratively scoped (239.xxx.xxx.xxx)
  - Organization-local, site-local
- Range restricted using TTL
  - 0 – host; 1 – local link; 2 – past 1<sup>st</sup> router; etc.



# IPv6 multicast addresses ([RFC4291-2.7](#))



- Flgs (oRPT) — T: temporary (1) or ‘well-known’ (0) address; P: prefix-based multicast address; R: RP address embedded
- Scope: 1 node-local scope, 2 link-local scope, 5 site-local scope, 8 organization-local scope, E global scope
- Reserved addresses: [RFC 2375](#)
- Mapping to Ethernet uses last 32 bits ([RFC2464#7](#)):



- Uses last 4 octets of IPv6 multicast destination address (13, 14, 15 and 16)

# IPv6 Addresses – Reserved RFC 2375

- Node-Local Scope

- FF01:0:0:0:0:0:1 All Nodes Address
- FF01:0:0:0:0:0:2 All Routers Address

- Link-Local Scope

- FF02:0:0:0:0:0:1 All Nodes Address
- FF02:0:0:0:0:0:2 All Routers Address
- FF02:0:0:0:0:0:4 DVMRP Routers
- FF02:0:0:0:0:0:6 OSPF IGP Designated Routers
- FF02:0:0:0:0:0:9 RIP Routers
- FF02:0:0:0:0:0:A EIGRP Routers
- FF02:0:0:0:0:0:B Mobile-Agents
- FF02:0:0:0:0:0:D All PIM Routers
- FF02:0:0:0:0:0:E RSVP-ENCAPSULATION
- FF02:0:0:0:0:1:2 All-dhcp-agents
- FF02:0:0:0:0:1:FFYY:YYYY Solicited-Node MC Address

- Site-Local Scope

- FF05:0:0:0:0:0:2 All Routers Address
- FF05:0:0:0:0:0:1:3 All-dhcp-servers
- FF05:0:0:0:0:0:1:4 All-dhcp-relays
- FF05:0:0:0:0:0:1:1000 Service Location

- All Scope Multicast Addresses

- FF0X:0:0:0:0:0:101 Network Time Protocol (NTP)
- FF0X:0:0:0:0:0:107 AUDIONEWS - Audio News Multicast
- FF0X:0:0:0:0:0:108 SUN NIS+ Information Service
- FF0X:0:0:0:0:0:109 MTP Multicast Transport Protocol
- FF0X:0:0:0:0:0:10A IETF-1-LOW-AUDIO
- FF0X:0:0:0:0:0:10B IETF-1-AUDIO
- FF0X:0:0:0:0:0:10C IETF-1-VIDEO
- FF0X:0:0:0:0:0:10D IETF-2-LOW-AUDIO
- FF0X:0:0:0:0:0:10E IETF-2-AUDIO
- FF0X:0:0:0:0:0:10F IETF-2-VIDEO
- FF0X:0:0:0:0:0:110 MUSIC-SERVICE

X in the addresses is any legal scope value

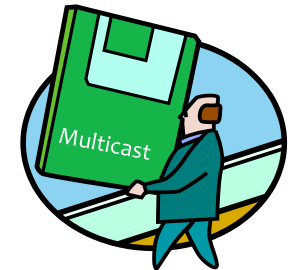
# Multicast Host Software extensions

- Sending:

- Hosts do not need routes
  - Not even a default route (!)
- Mapping to MAC address
- Specify TTL

- Receiving:

- API to:
  - Join / leave a group (on a specific interface)
  - Activate/deactivate loopback of packets sent to subscribed group
- Delivering packets to every process/app that joined the group
  - Maintain info on subscribed groups

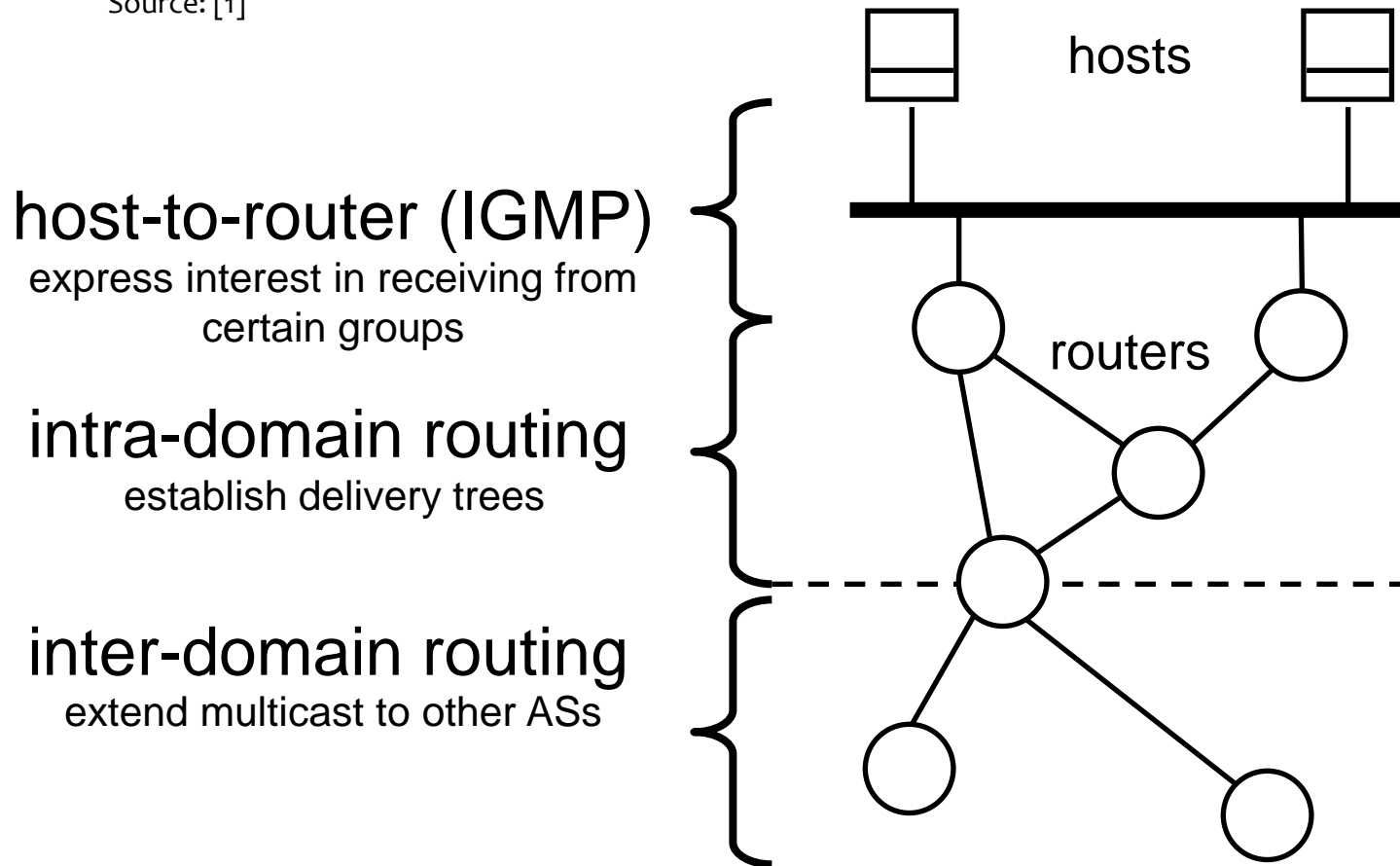


# Multicast IP

- Routing across different subnetworks
  - Routers multicast
    - Shortest path
    - Do not send through path that leads to no receivers
    - Joining/leaving groups at any moment
- Best effort semantics
  - Losses, duplicates, out-of-order delivery
- Any host can send to the multicast address
  - Joining group only required for receiving

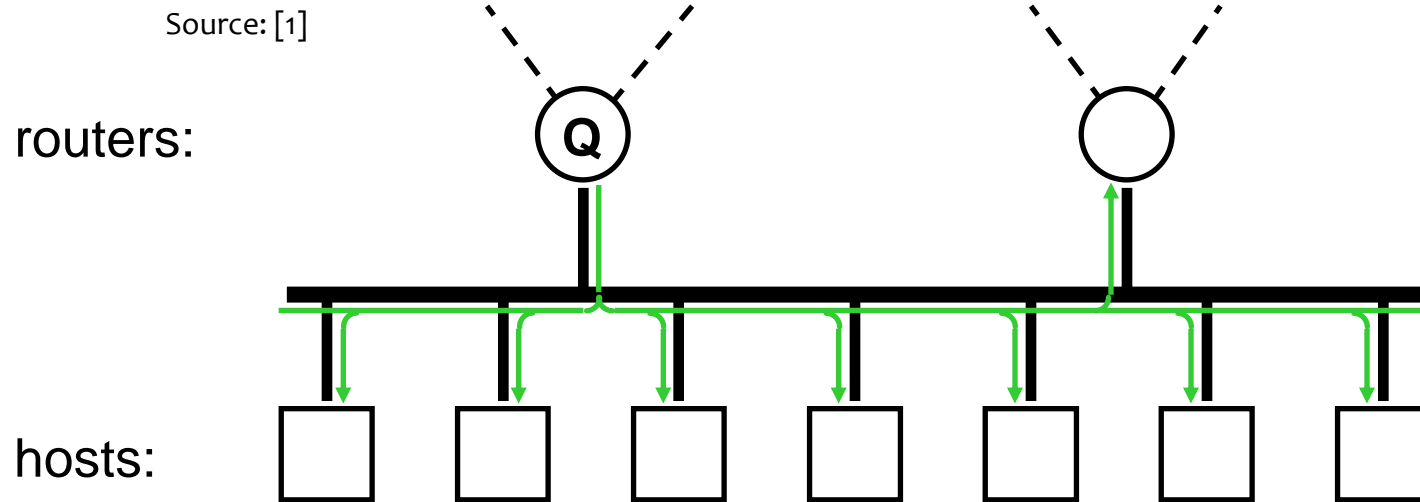
# Components

Source: [1]



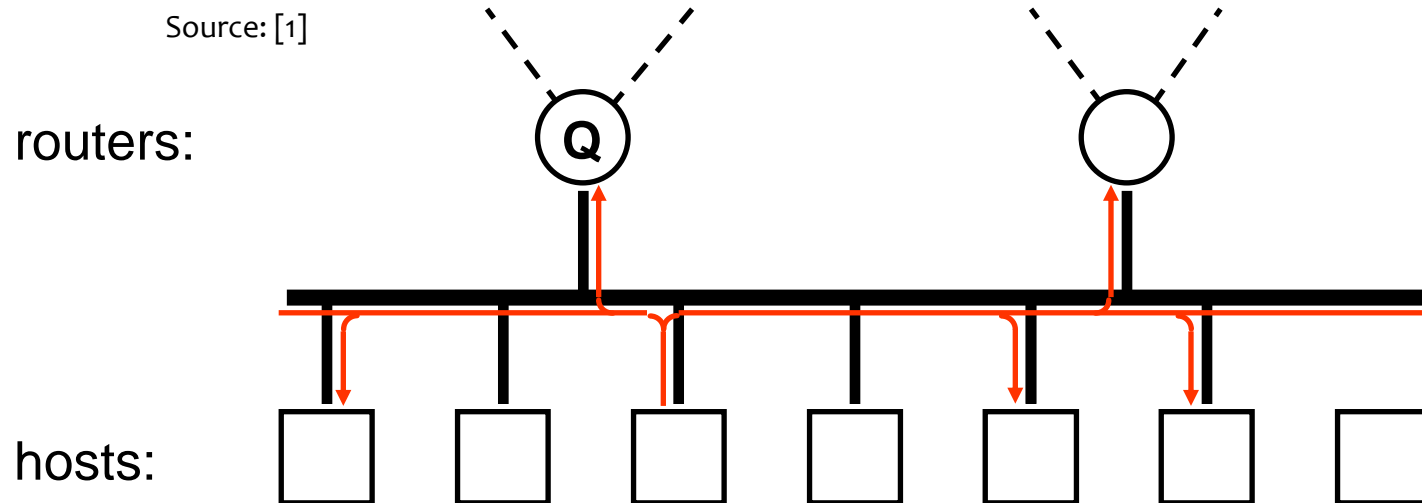
# IPv4 IGMP – RFC3376

## Internet Group Management Protocol



- A Querier is elected among MC routers on link
- Periodically (usually every 125 s) sends Membership Query message to all-systems group (224.0.0.1)
- Hosts start random timer for each group they belong to

# IPv4 IGMP



- When timer expires host sends Membership Report
- Other hosts cancel timer for that group
- Router listens to every Report and removes groups where no report is received
- When a host joins a group it sends a Report without waiting for a Query

# IPv4 IGMP: messages (v3)

24

TYPE	Max Resp Time	CHECKSUM
Group address (ZERO in QUERY)		

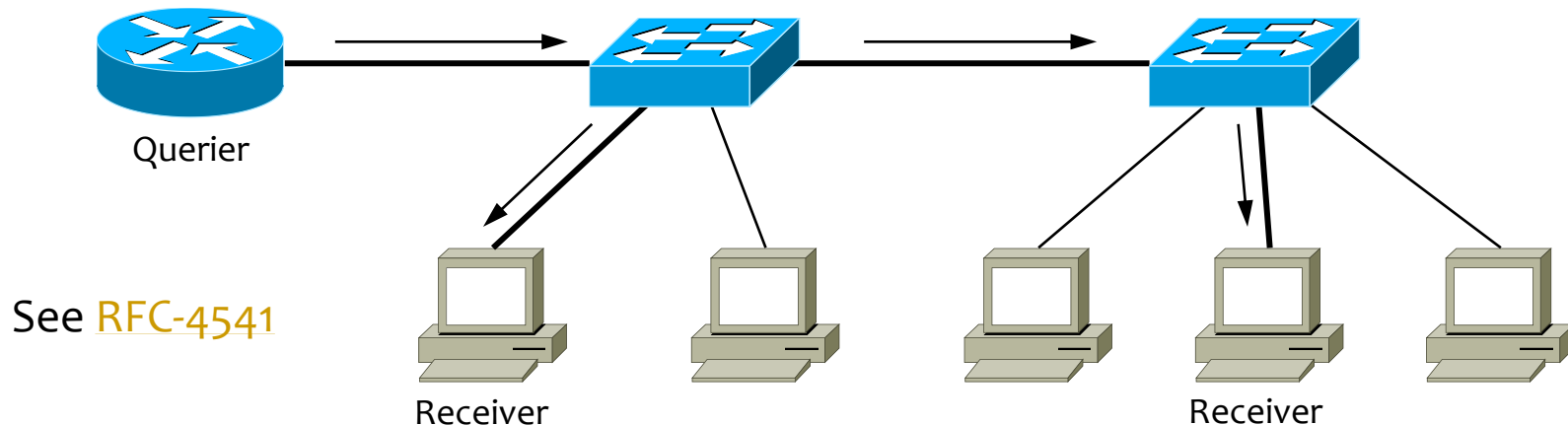
Type	Group Address	Meaning
0x11	N/U (zero)	Membership Query (general)
0x11	Used	Membership Query (specific)
0x12	Used	Membership Report (v1)
0x16	Used	Membership Report (v2)
0x17	Used	Leave Group
0x22	Used	Membership Report (v3)

- IGMP v1: RFC 1112
- IGMP v2: RFC 2236 (see [Appendix I](#))
  - Leave message, querier election algorithm
- IGMP v3: [RFC3376](#) (see [Appendix B](#))
  - Allows indication of specific source to subscribe to



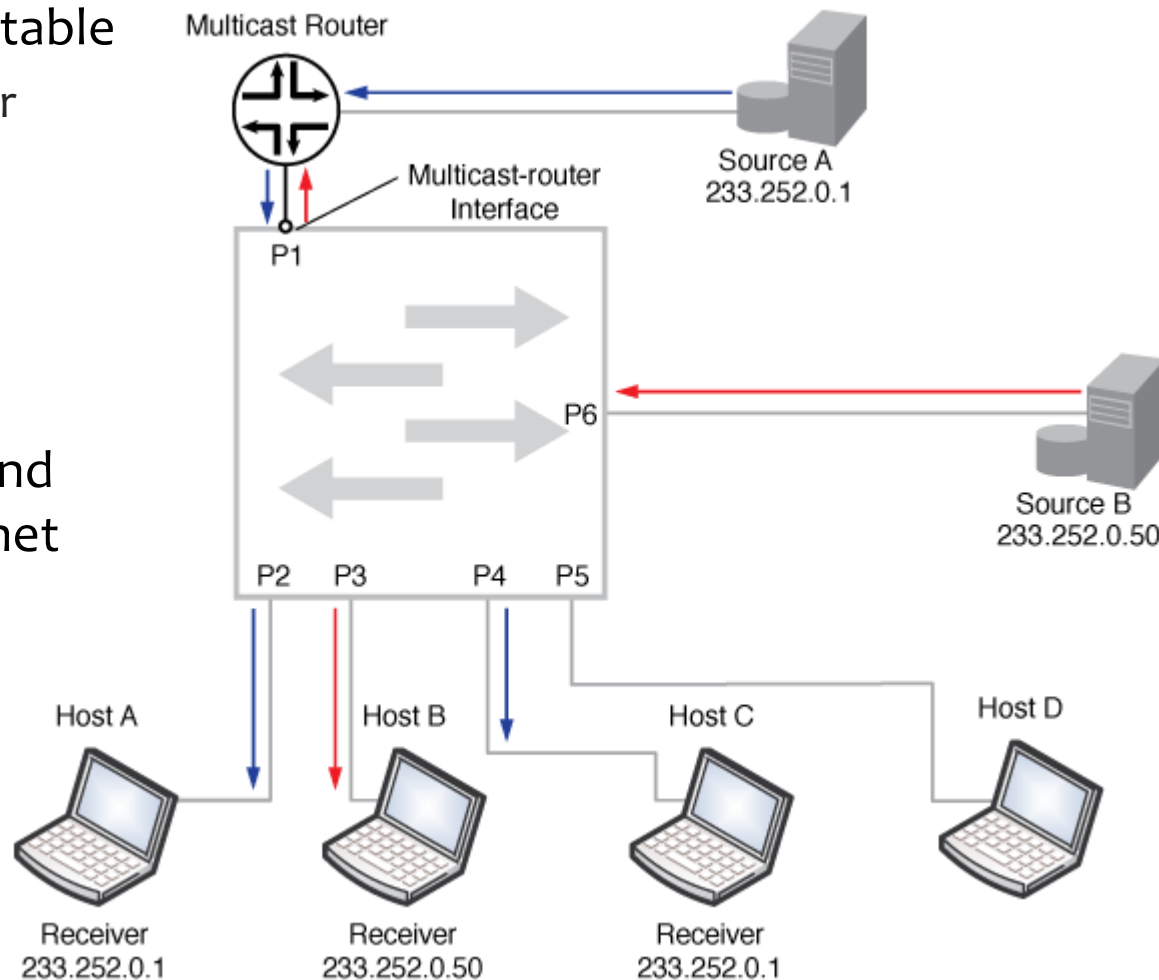
# IGMP Snooping

- IGMP runs between routers and hosts
- In switched Ethernet, efficient use of capacity requires sending only to ports with listeners
- Ethernet switches snoop IGMP packets to know whether or not to forward on each port
- Optional: proxy reporting / report suppression
  - To reduce load at the multicast router



# IGMP Snooping — Example 1

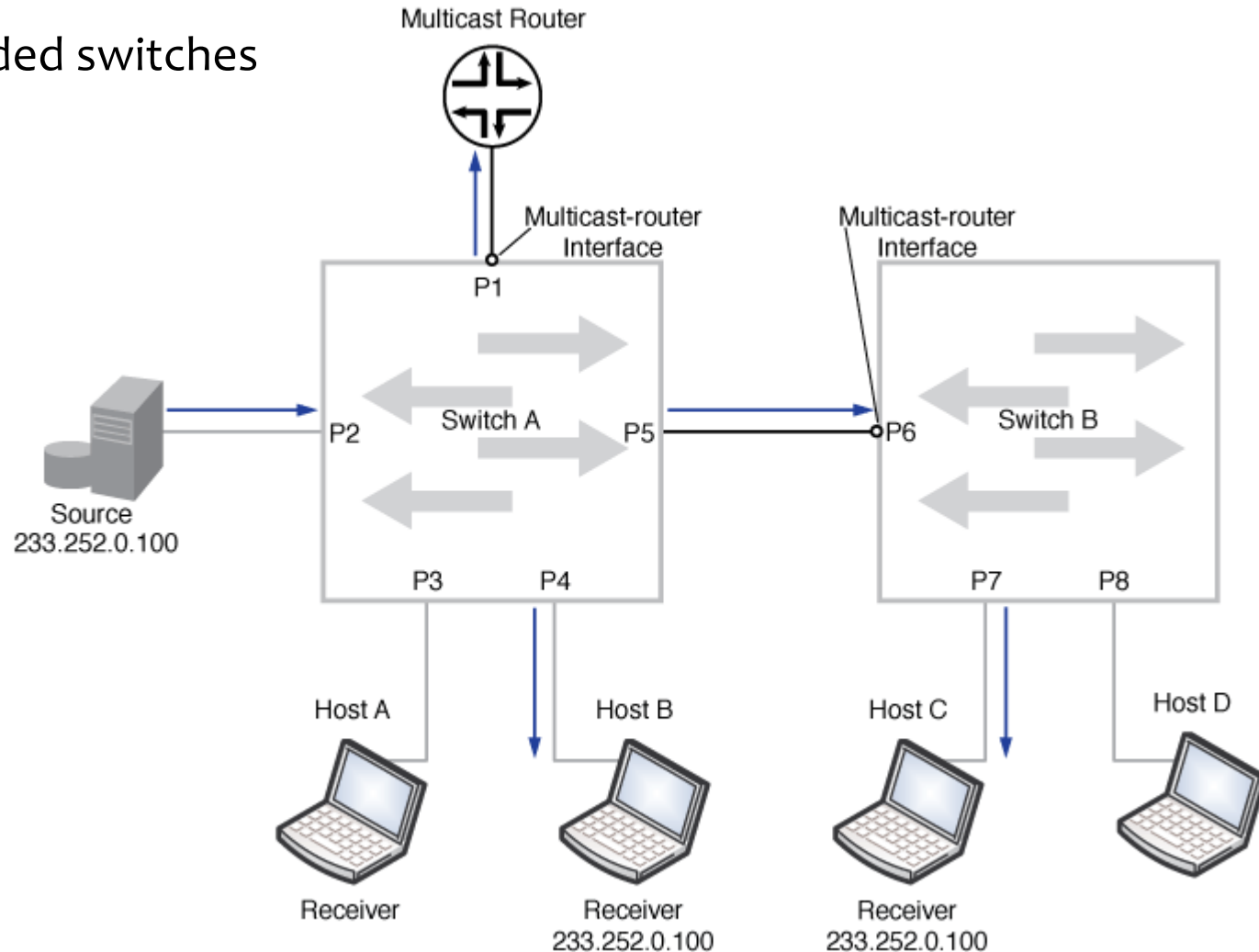
- Multicast cache table
  - Multicast router interfaces
  - Group member interfaces
- Sources inside and outside the subnet



Source: Understanding IGMP Snooping and Multicast Forwarding, Juniper TechLibrary

# IGMP Snooping — Example 2

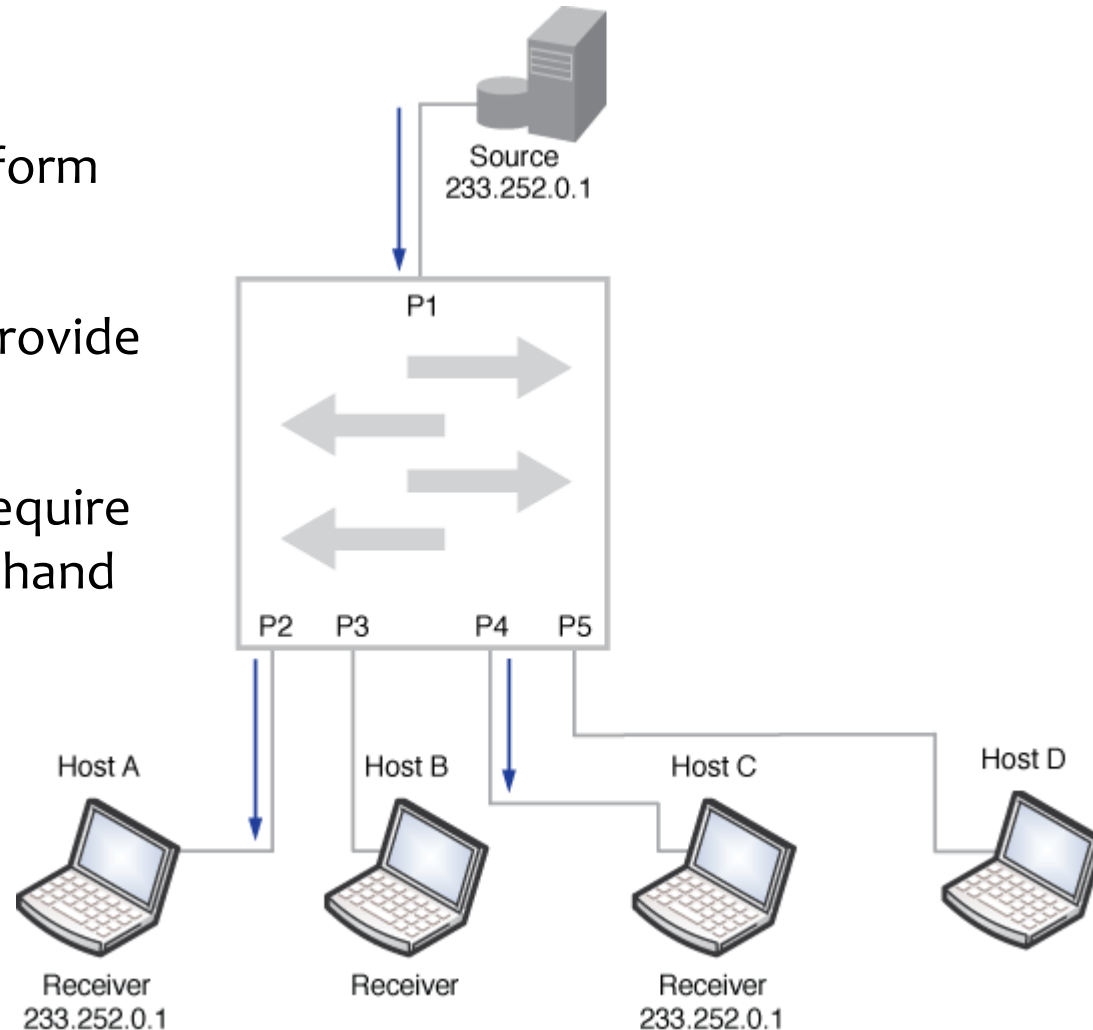
- Cascaded switches



Source: Understanding IGMP Snooping and Multicast Forwarding, Juniper TechLibrary

# IGMP Snooping — Example 3

- Isolated subnet
- No router to perform IGMP queries
- Some switches provide an IGMP querier
- Some switches require configuration by hand for this to work



Source: Understanding IGMP Snooping and Multicast Forwarding, Juniper TechLibrary

# IPv6 Multicast Listener Discovery

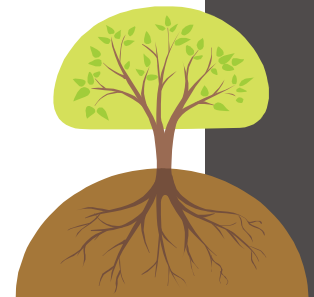
- [RFC3810](#)
- “MLDv2 is a translation of the IGMPv3 protocol for IPv6 semantics.”
- “Used by an IPv6 router to discover the presence of multicast listeners on directly attached links, and to discover which multicast addresses are of interest to those neighbouring nodes.”
- [RFC4604](#) updates MLD and IGMP for Source Specific Multicast (SSM).

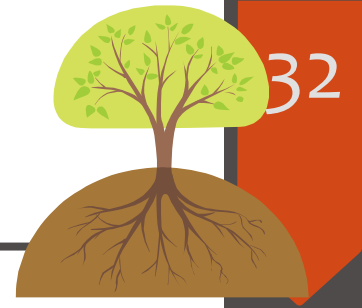
# Multicast routing: requirements

- Dynamic routing
  - Changes if host join/leaves group
- Need to check more than destination address
  - Routing can also use the source address
- Multicast packets can originate on a host not belonging to the group

# Routing characteristics

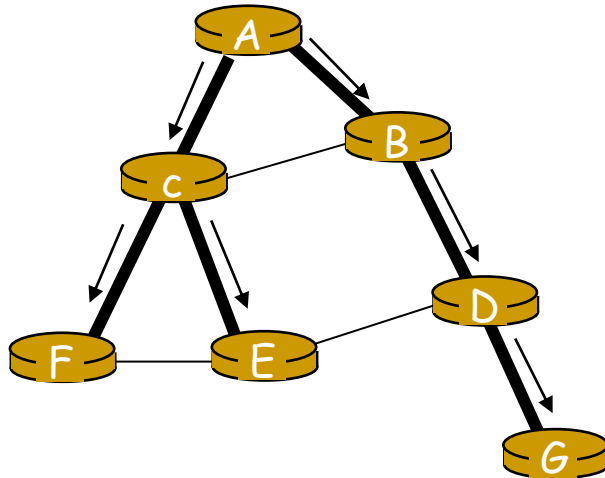
- Data policy:
  - Receiver-driven / opt-in: send only to hosts that joined
  - Data-driven / opt-out: broadcast, then prune paths without receivers
- Distribution trees:
  - Source tree: from each source to receivers
  - Shared tree: several sources share (part of) the path



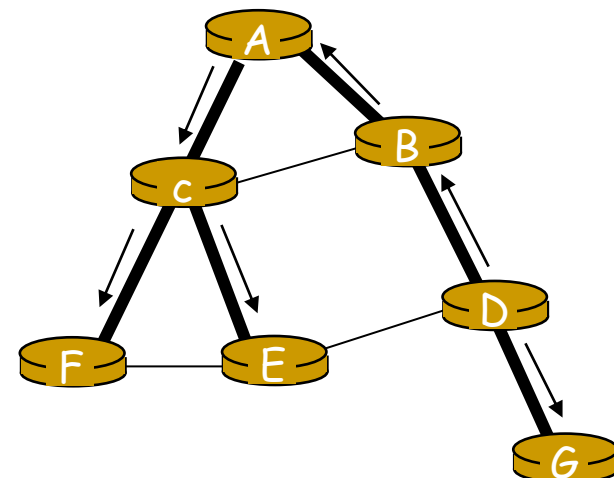


# Spanning Tree

- First construct a spanning tree
- Nodes forward copies only along spanning tree
  - Packets can be sent in different directions on a link depending on the source



(a) Multicast from A

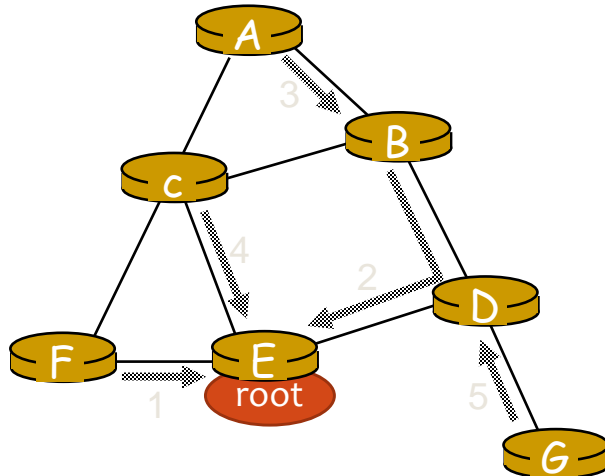


(b) Multicast from D

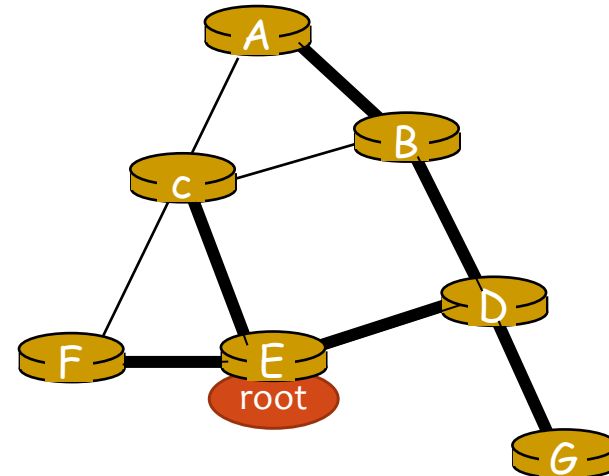


# Spanning Tree: Creation

- Select a centre/root node: in the example, node E
- Each node sends unicast join message to centre/root
  - Message is forwarded until it arrives at a node already in the spanning tree

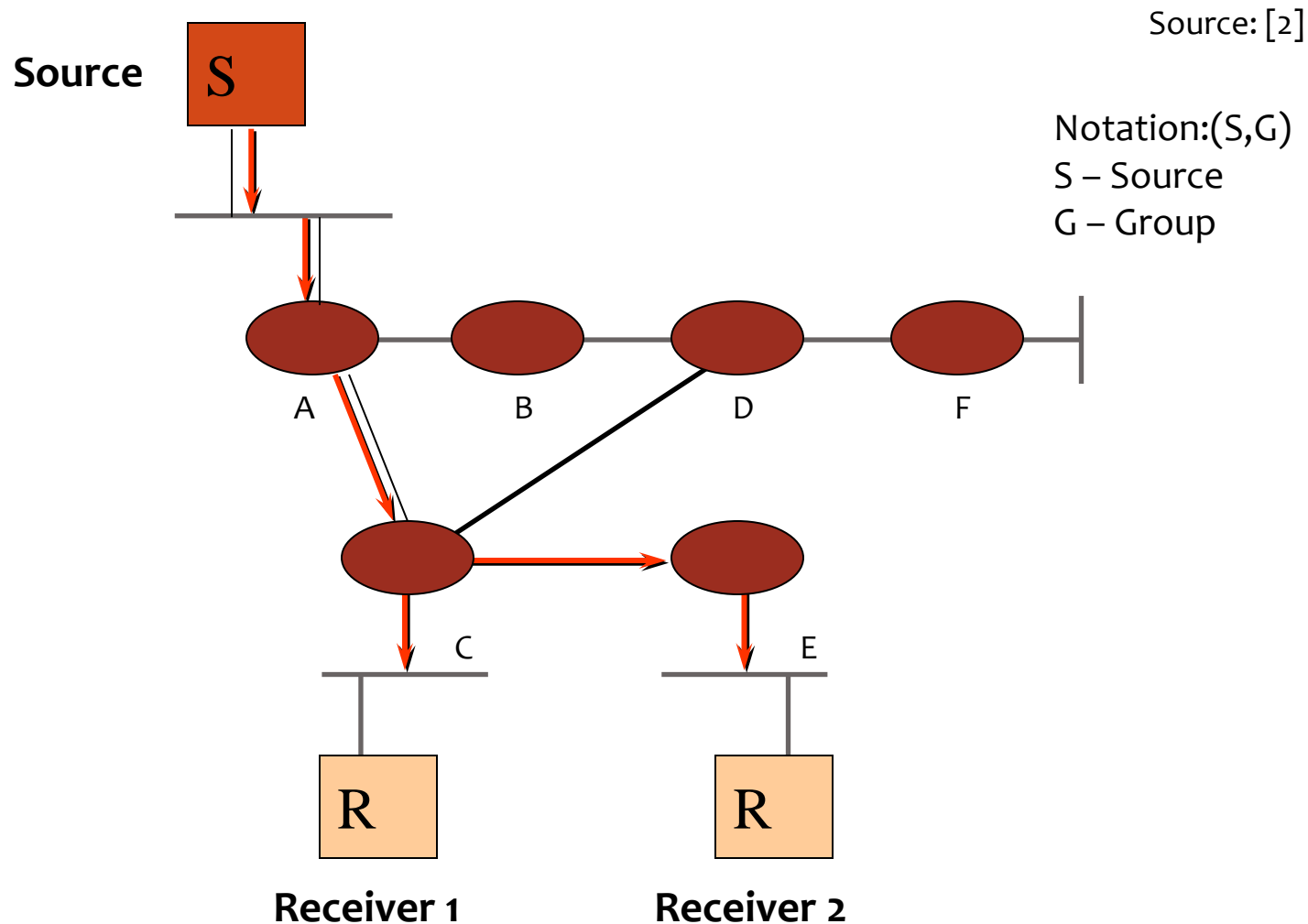


(a) Stepwise construction of spanning tree

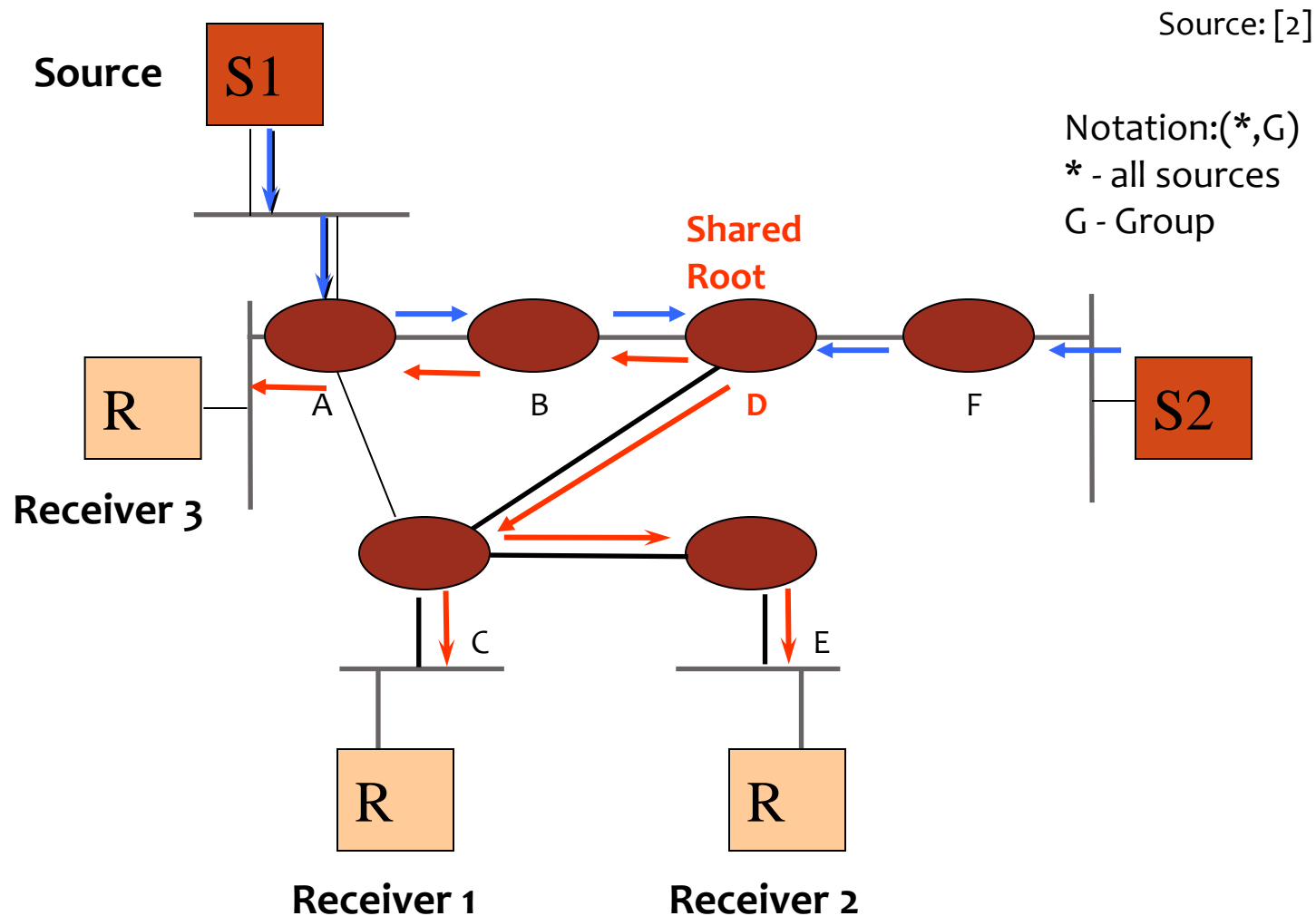


(b) Constructed spanning tree

# Source-specific tree

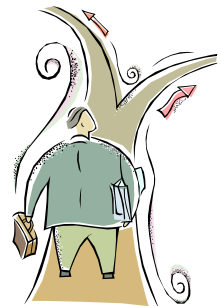


# Shared tree



# Multicast routing

- Reverse Path Forwarding (RPF)
  - Only retransmit packets that arrive on the interface that leads to the source (would be the next hop)
    - Retransmit on all other interfaces
- Truncated RPF (TRPF)
  - Additional restriction: only send through interfaces that lead to group members
    - Prune branches leading to no members

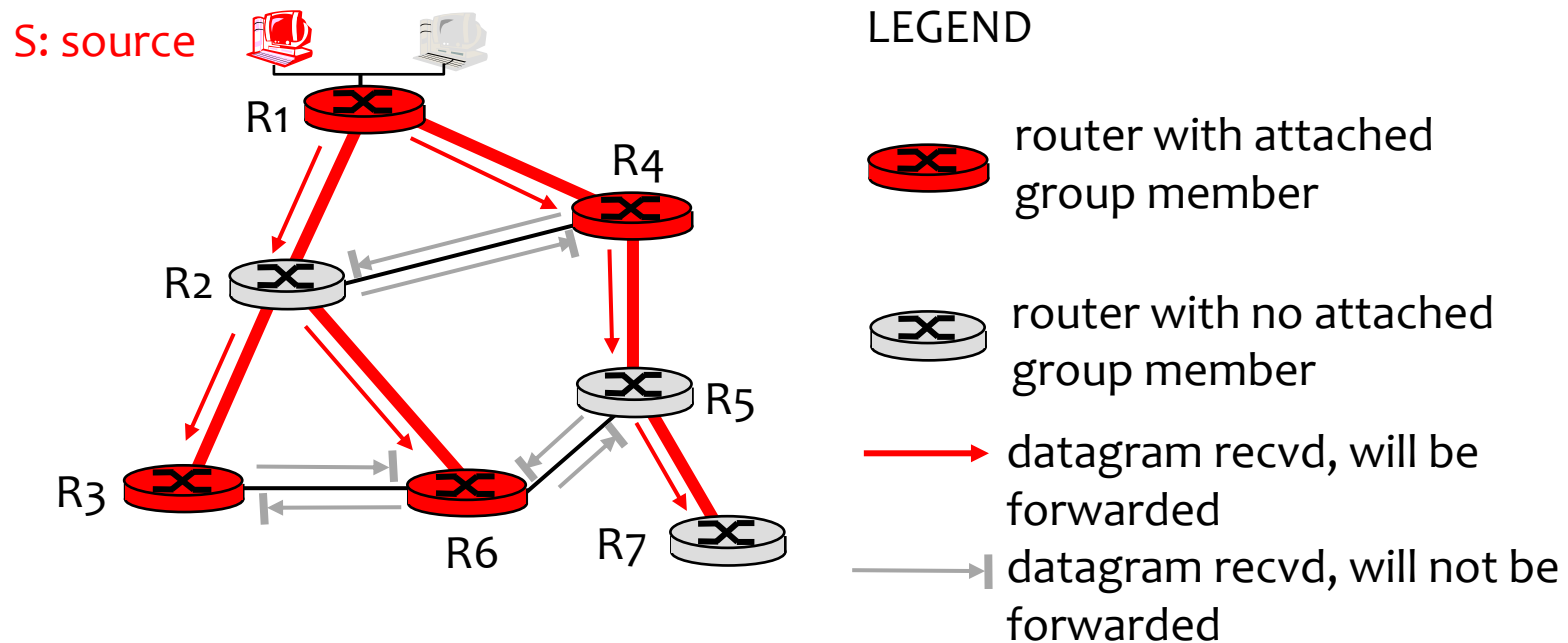


# Multicast routing

- Problems
  - Packet duplication
  - Delivery path depends on source
    - Tree rooted at source
- Multicast routers know which hosts belong to which group
  - Need to propagate group membership information through the network/Internet
- Routes may change according to receivers
  - Conflict between overhead and routing efficiency



# Reverse Path Forwarding: example



- Result is a source-specific reverse Shortest Path Tree
  - May be a bad choice with asymmetric links

# SSM and ASM

- Source-Specific Multicast
  - Specify which source to subscribe to
    - Requires host (API) and host-to-router (IGMP/MLD) support
  - [RFC3569](#), “An Overview of Source-Specific Multicast (SSM)”
- Any Source Multicast
  - Does not specify source
    - Every transmission to the group is received, irrespective of the sender



# Routing

IPv4 and IPv6



# PIM: Protocol Independent Multicast

- not dependent on any specific underlying unicast routing algorithm (works with all)
- two different multicast distribution scenarios

## Dense:

- group members densely packed, in “close” proximity
- bandwidth plentiful

## Sparse:

- # networks with group members small compared to # interconnected networks
- group members “widely dispersed”
- bandwidth scarce

# Consequences of Sparse-Dense Dichotomy

## Dense:

- group membership assumed until routers explicitly prune
- data-driven construction of multicast tree (e.g., RPF)
- bandwidth and non-group-router processing profligate

## Sparse:

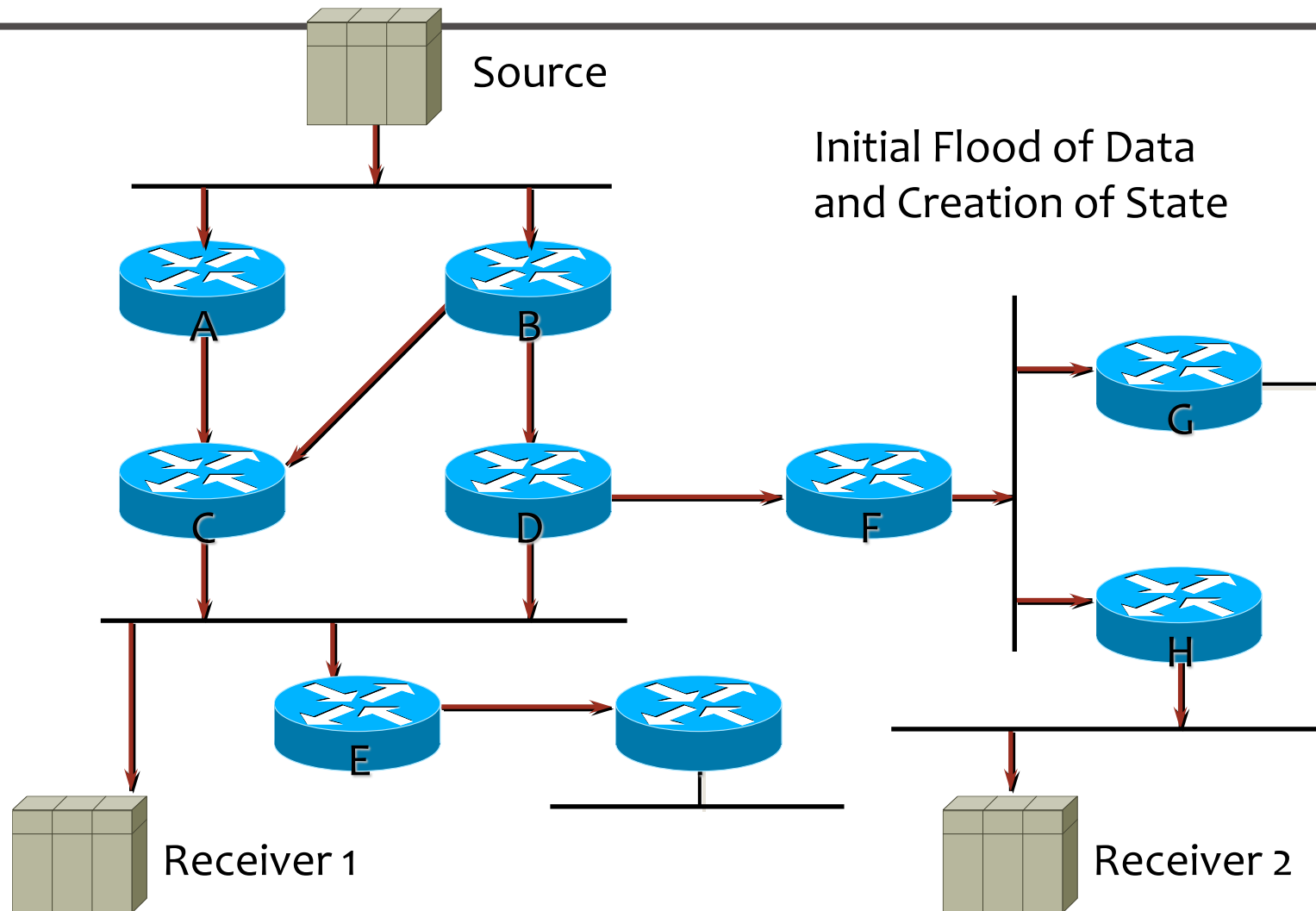
- no membership until routers explicitly join
- receiver-driven construction of multicast tree (centre-based)
- bandwidth and non-group-router processing conservative

# PIM-DM (Dense Mode)

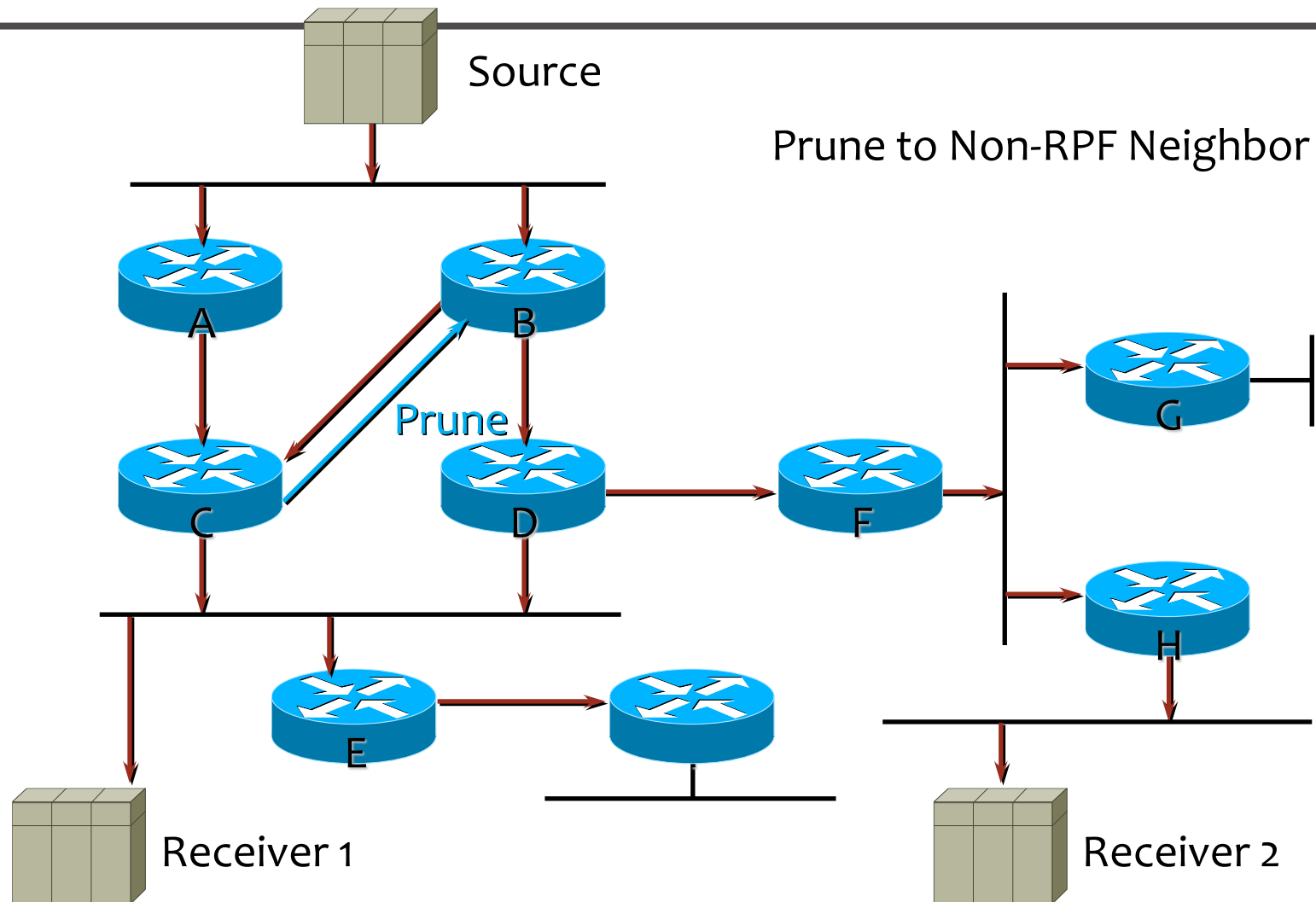
- Flood-and-prune
- Uses existing unicast routing tables for RPF
  - Protocol independent...
- Simpler (and less efficient) downstream flood than in DVMRP
  - To reduce reliance on underlying routing protocol
- Creates (S,G) state in every router
  - RPF → sender dependent
  - Even without receivers!



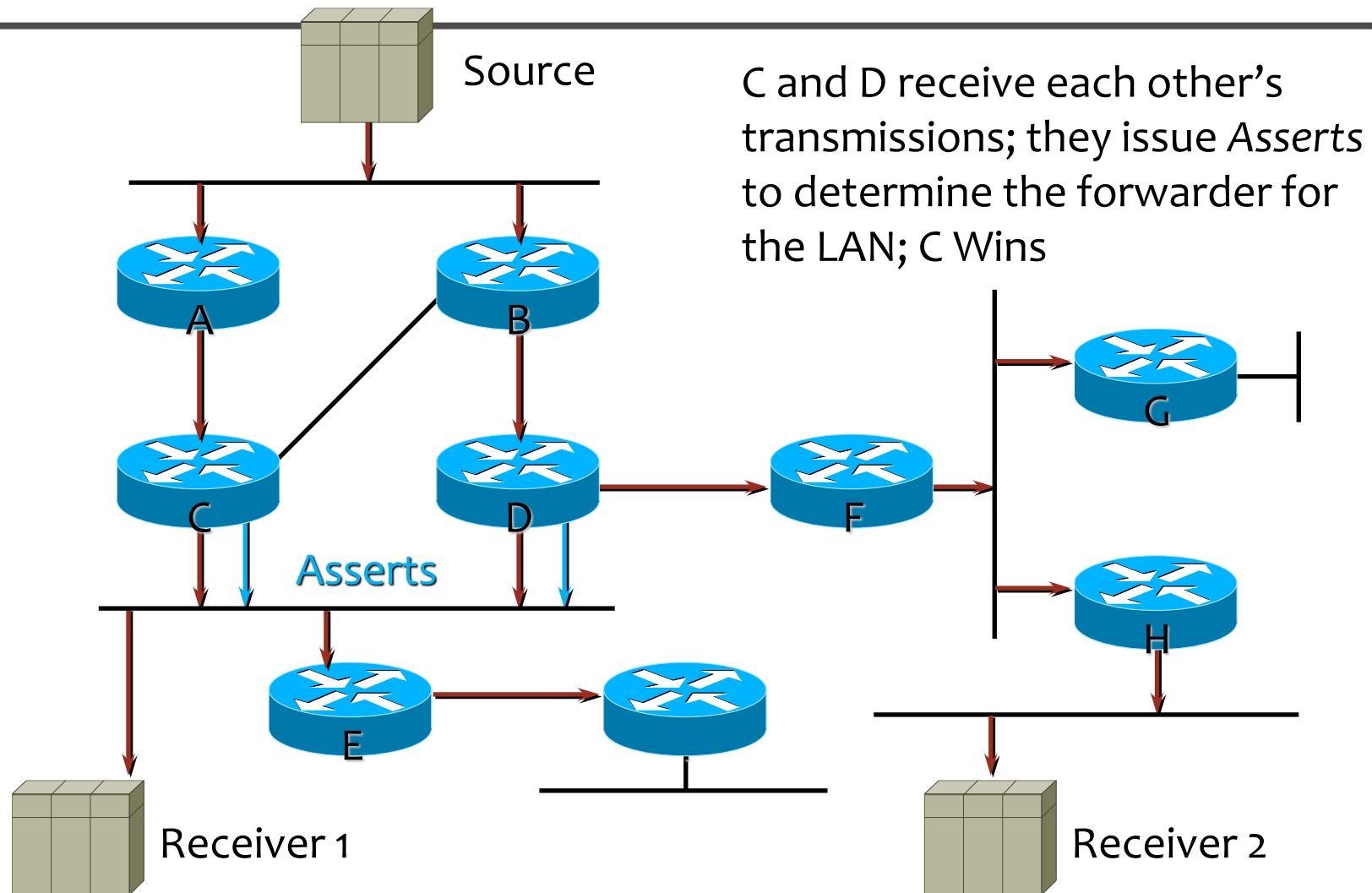
# PIM-DM



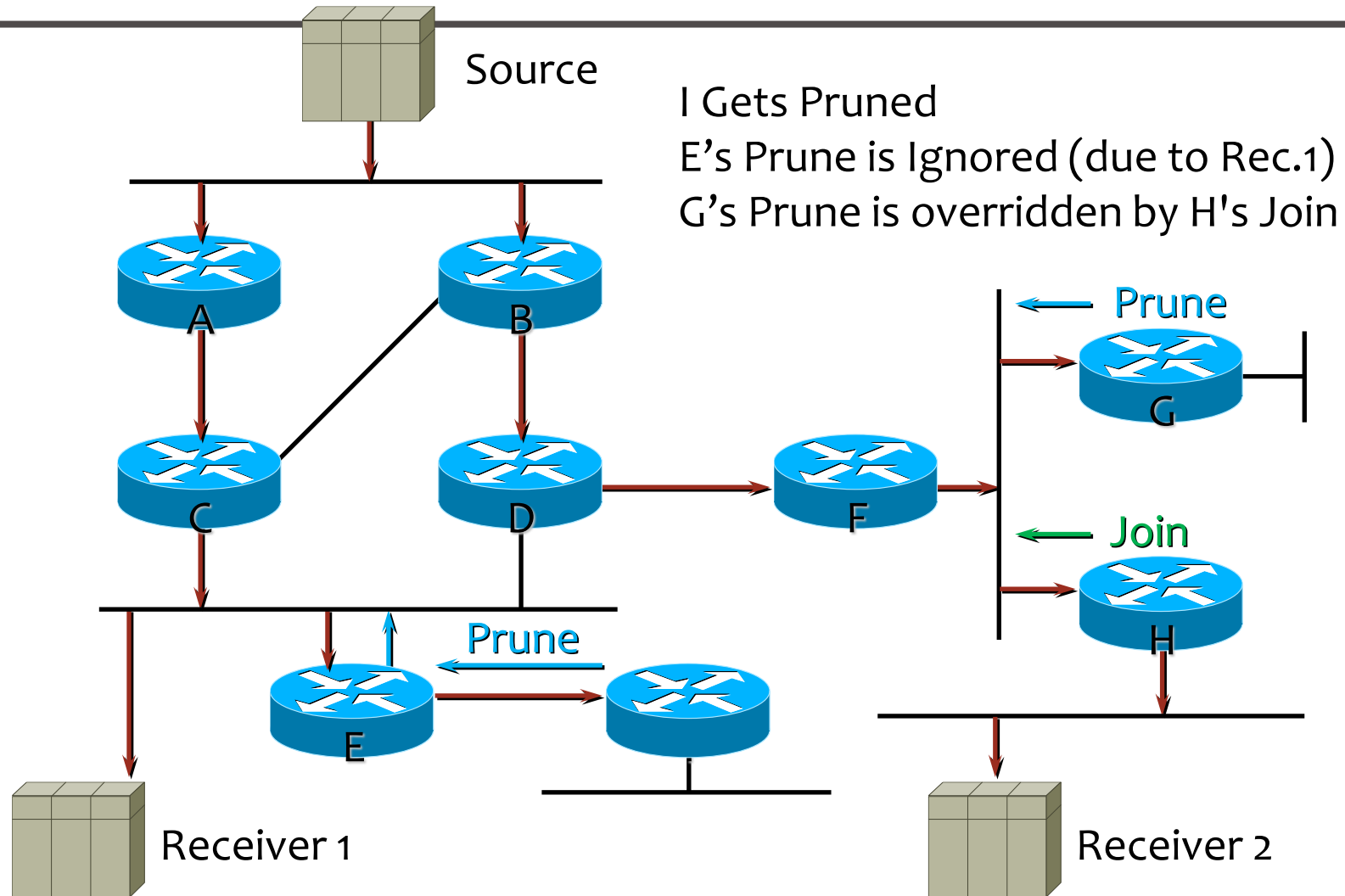
# PIM-DM



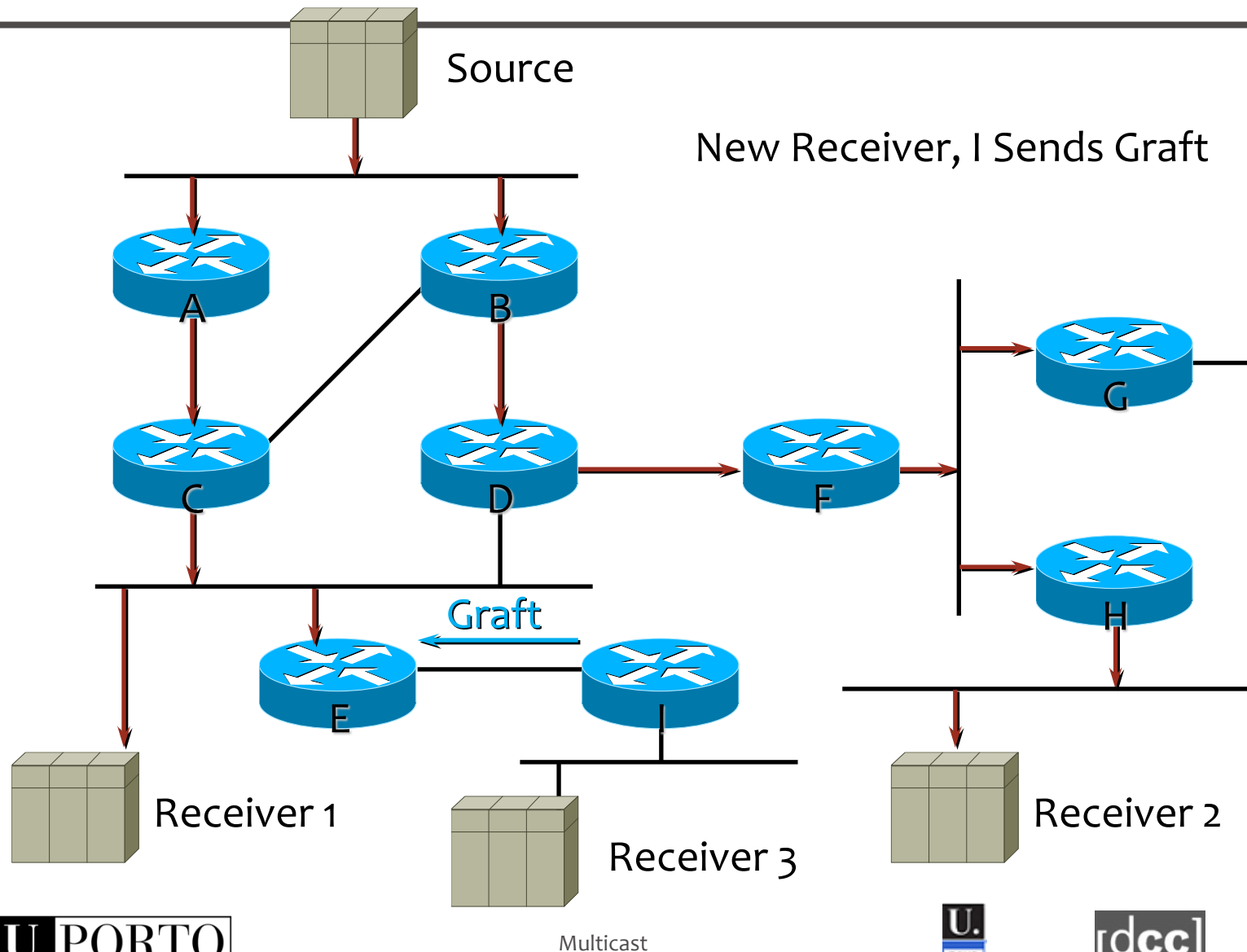
# PIM-DM



# PIM-DM



# PIM-DM

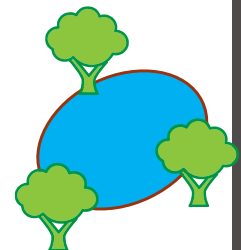




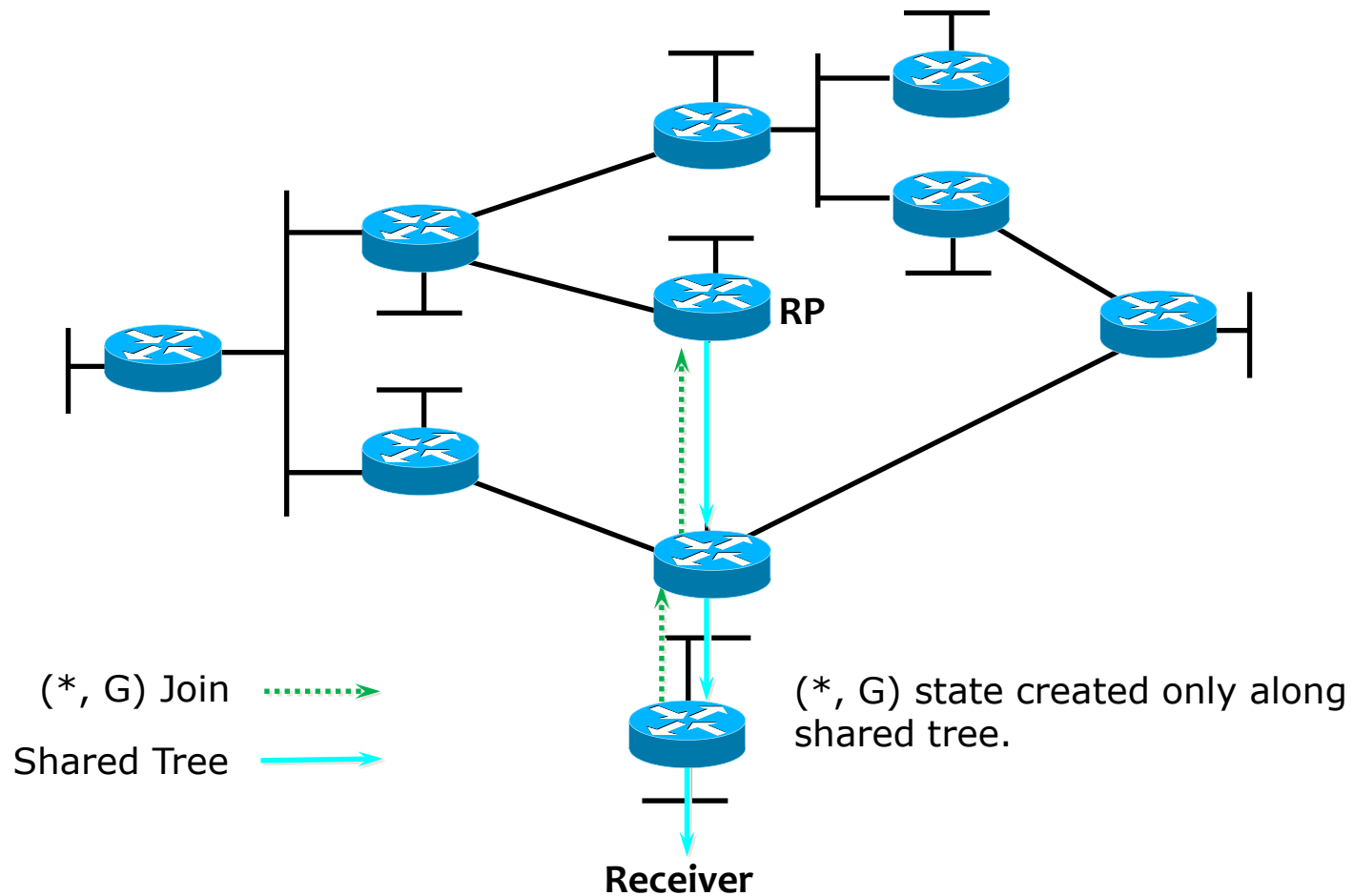


# PIM-SM (Sparse Mode)

- Receiver-driven (opt-in)
- Uses a unidirectional shared tree
- Rendezvous Point (RP) is the root of the tree
  - Group of potential RP
    - Only one active
    - Dynamic reconfiguration
- Data flows first through the shared tree
- Then source trees are created and data flows through them (switchover)
  - More efficient

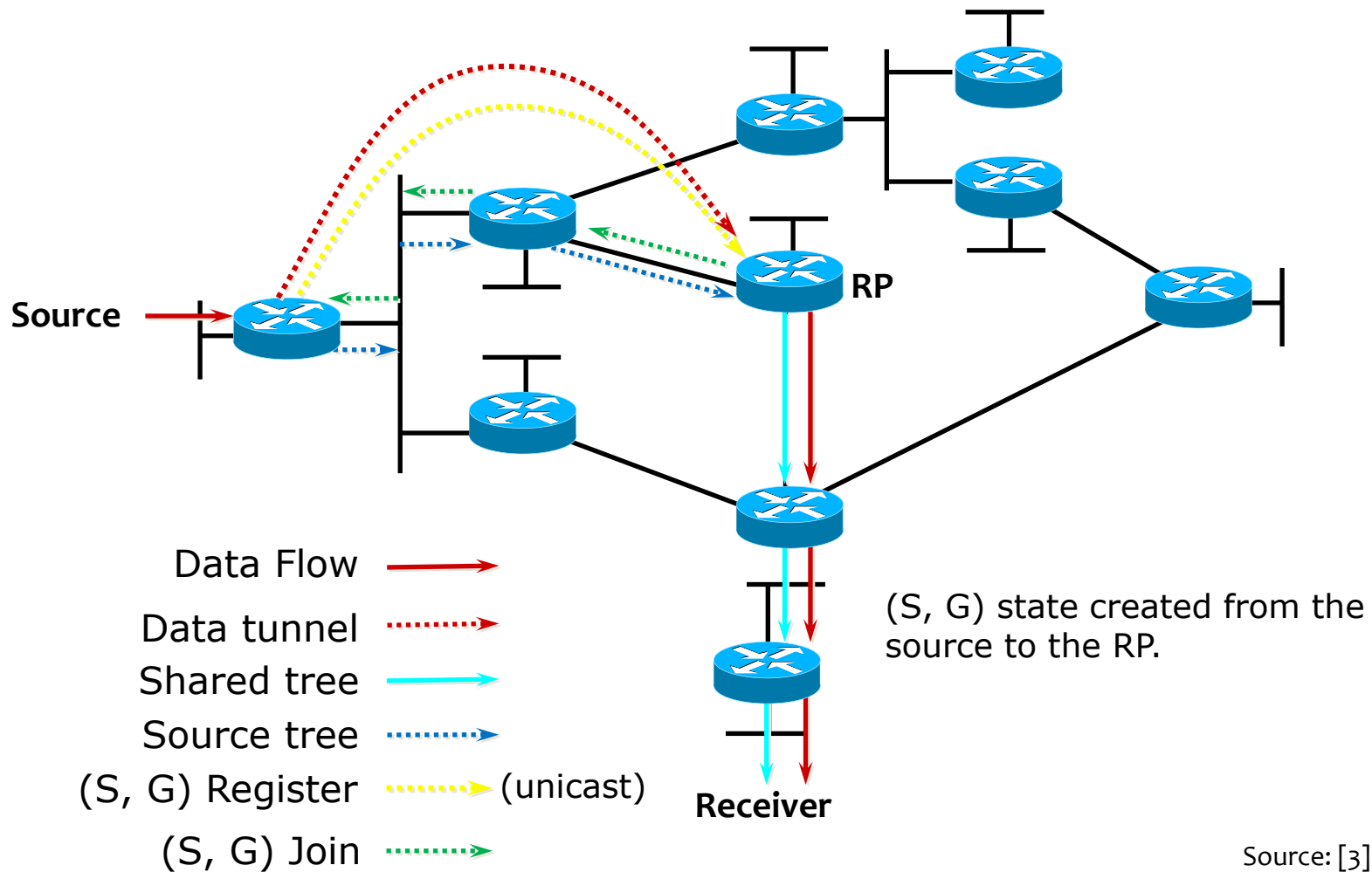


# PIM-SM: Joining shared tree

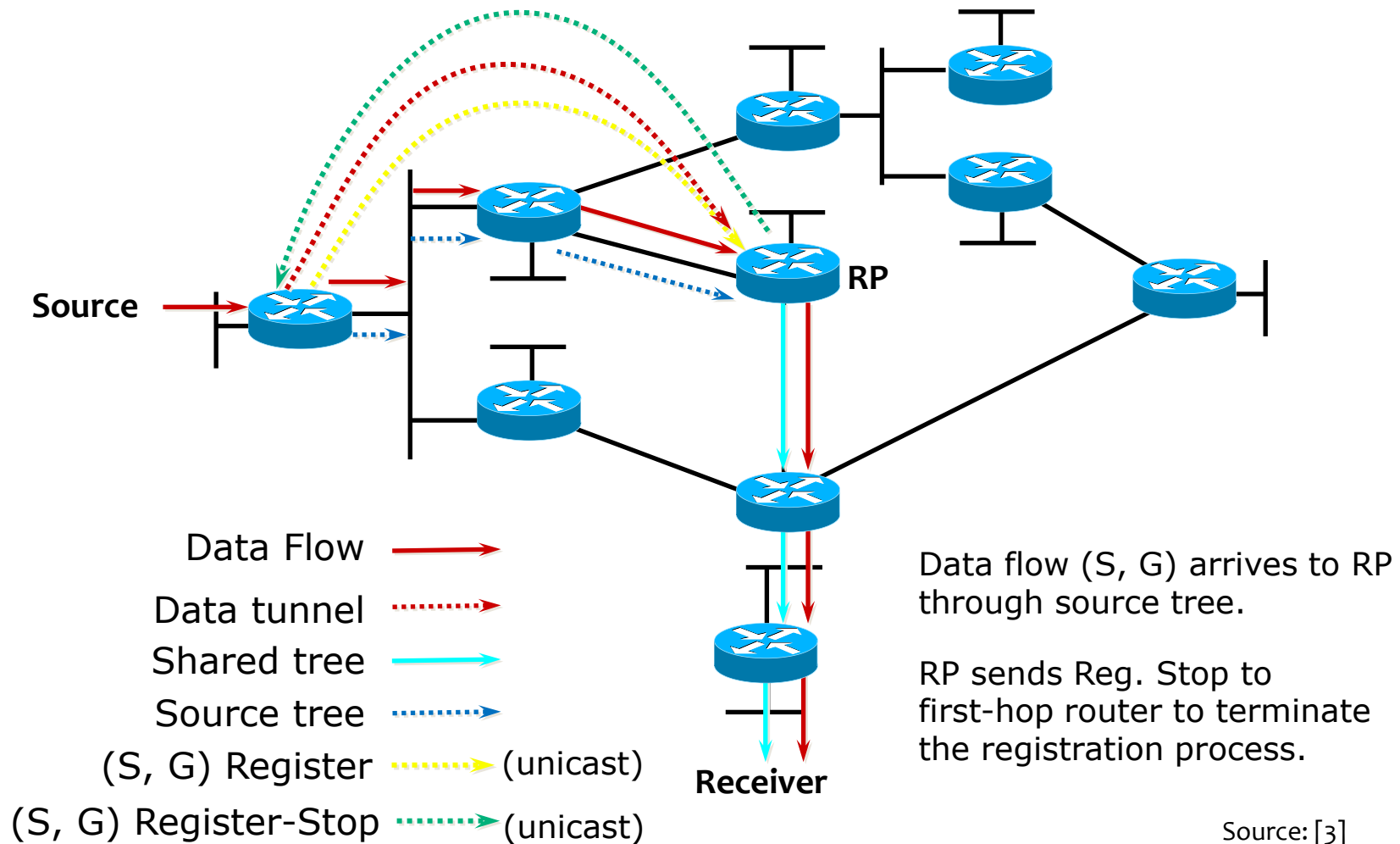


Source: [3]

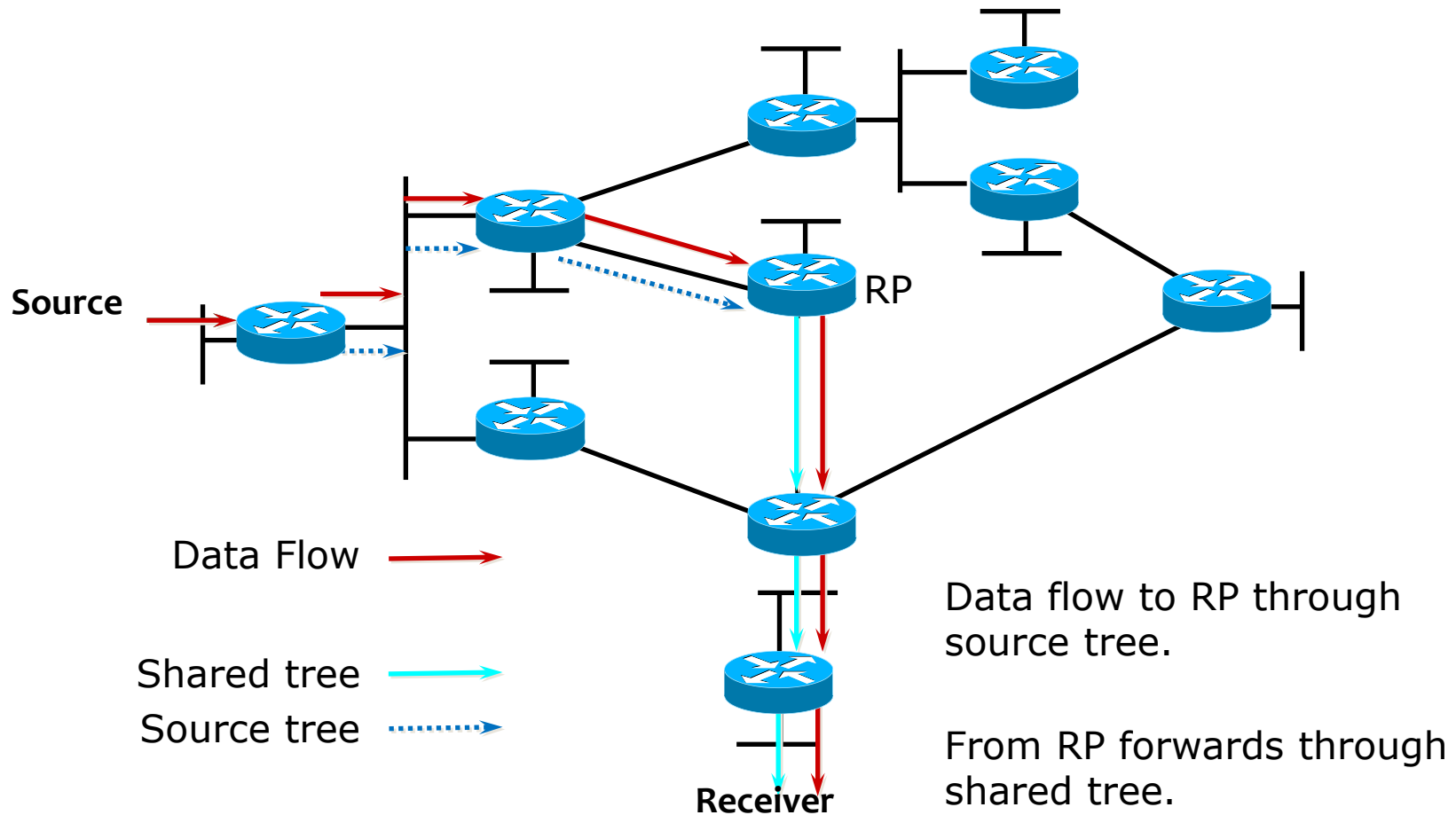
# PIM-SM: Source registration I



# PIM-SM: Source registration II



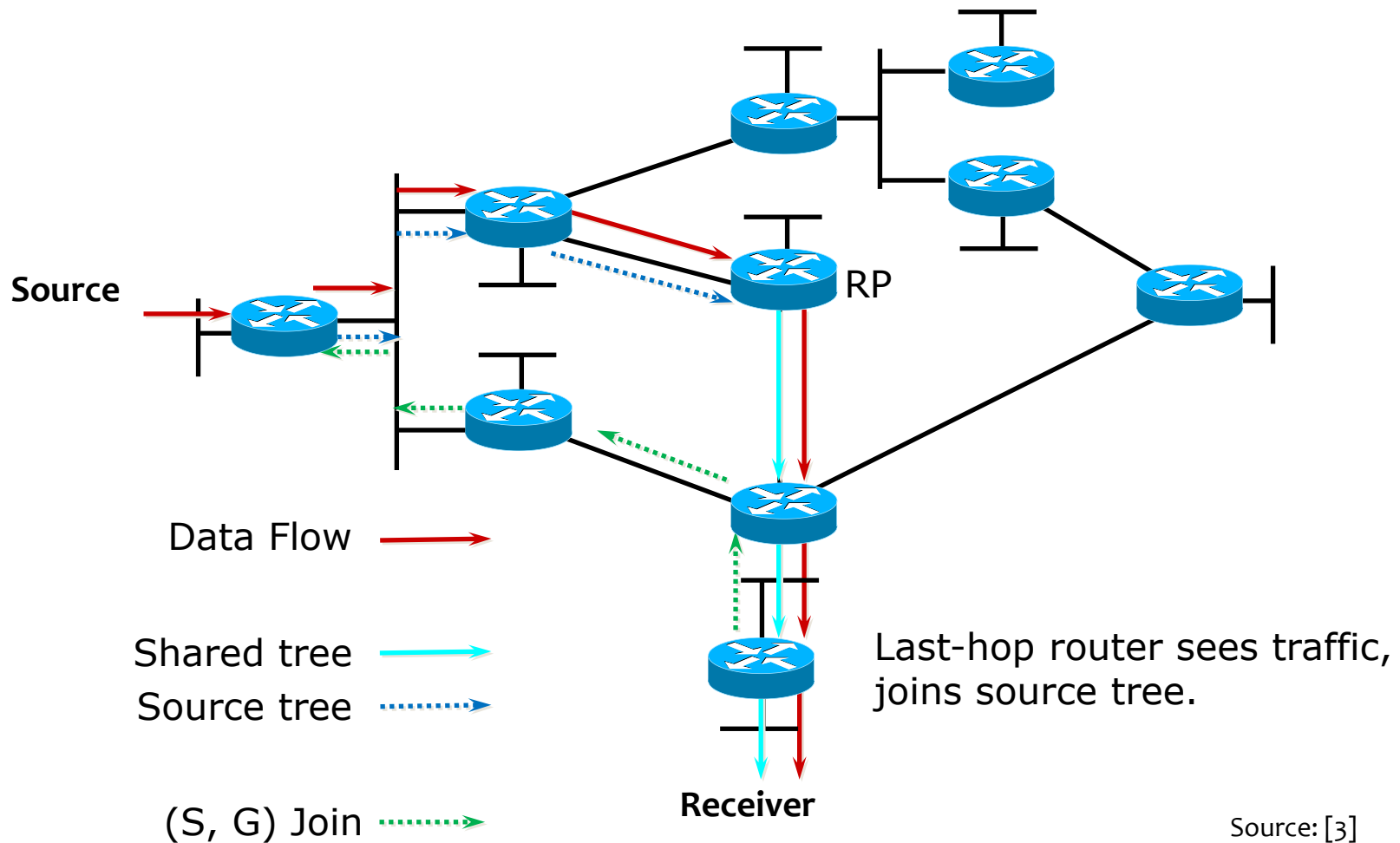
# PIM-SM: Source registration III



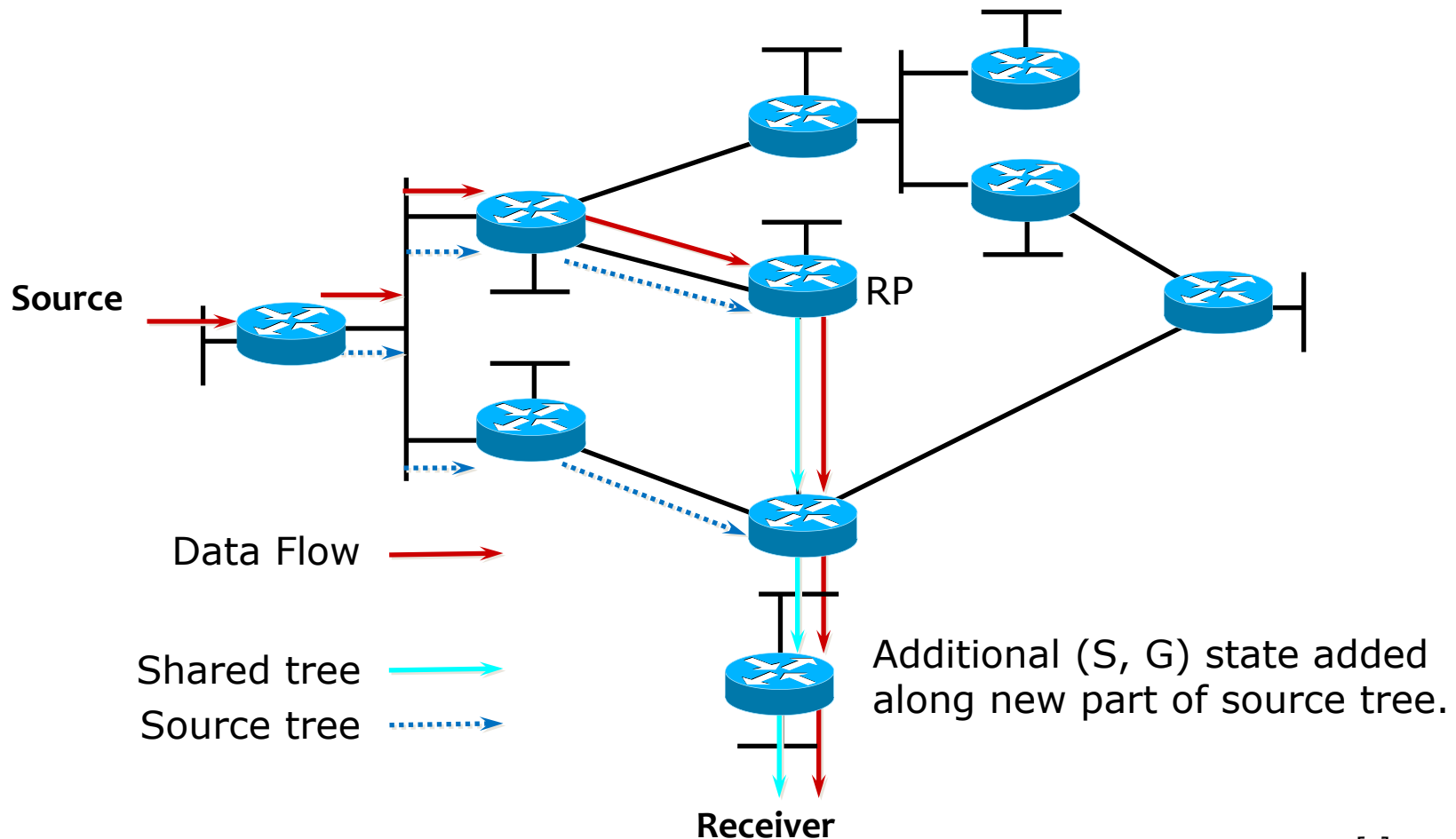
Source: [3]

# PIM-SM:

## Commuting to source tree



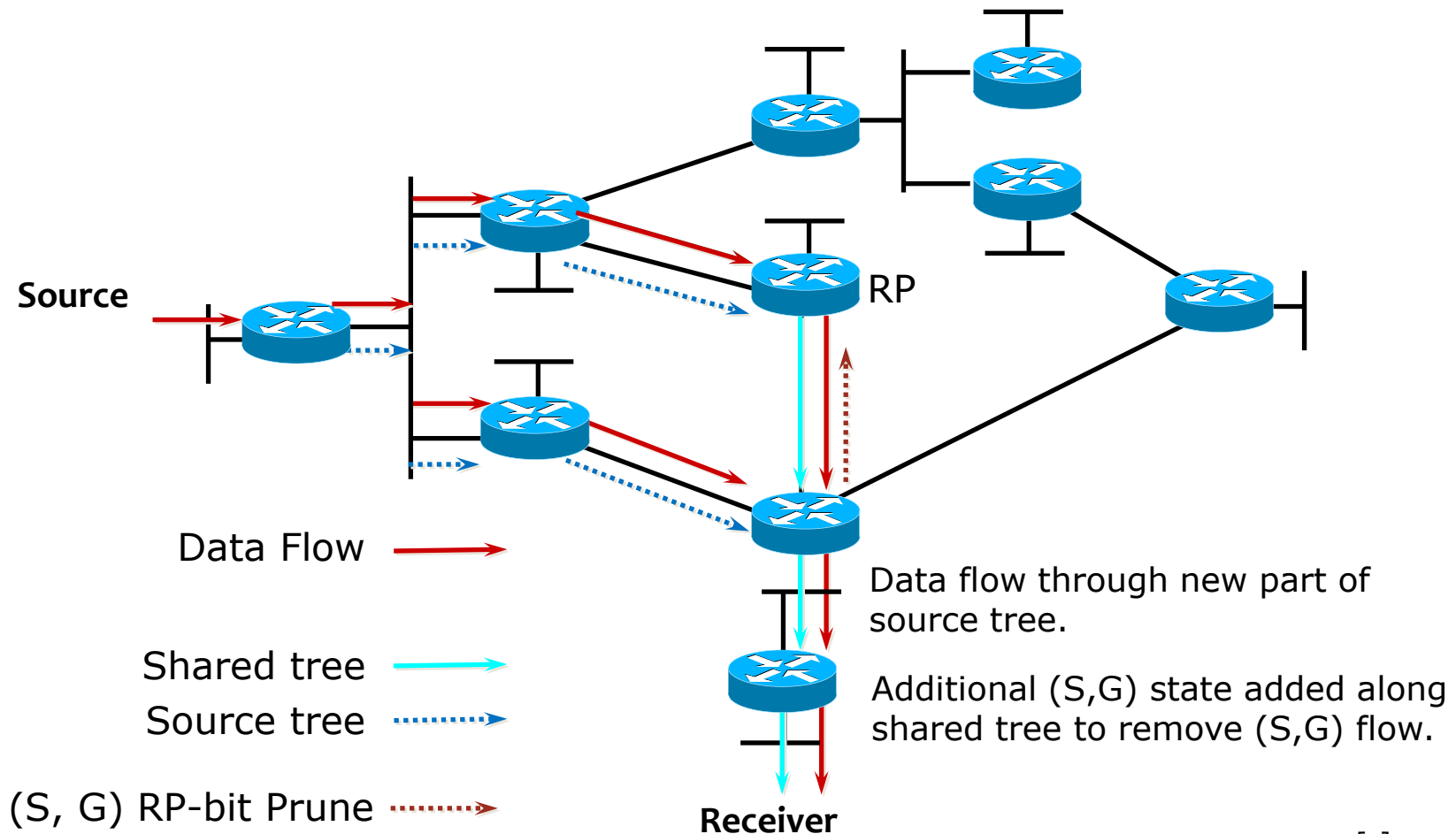
# PIM-SM: Commuting to source tree



Source: [3]

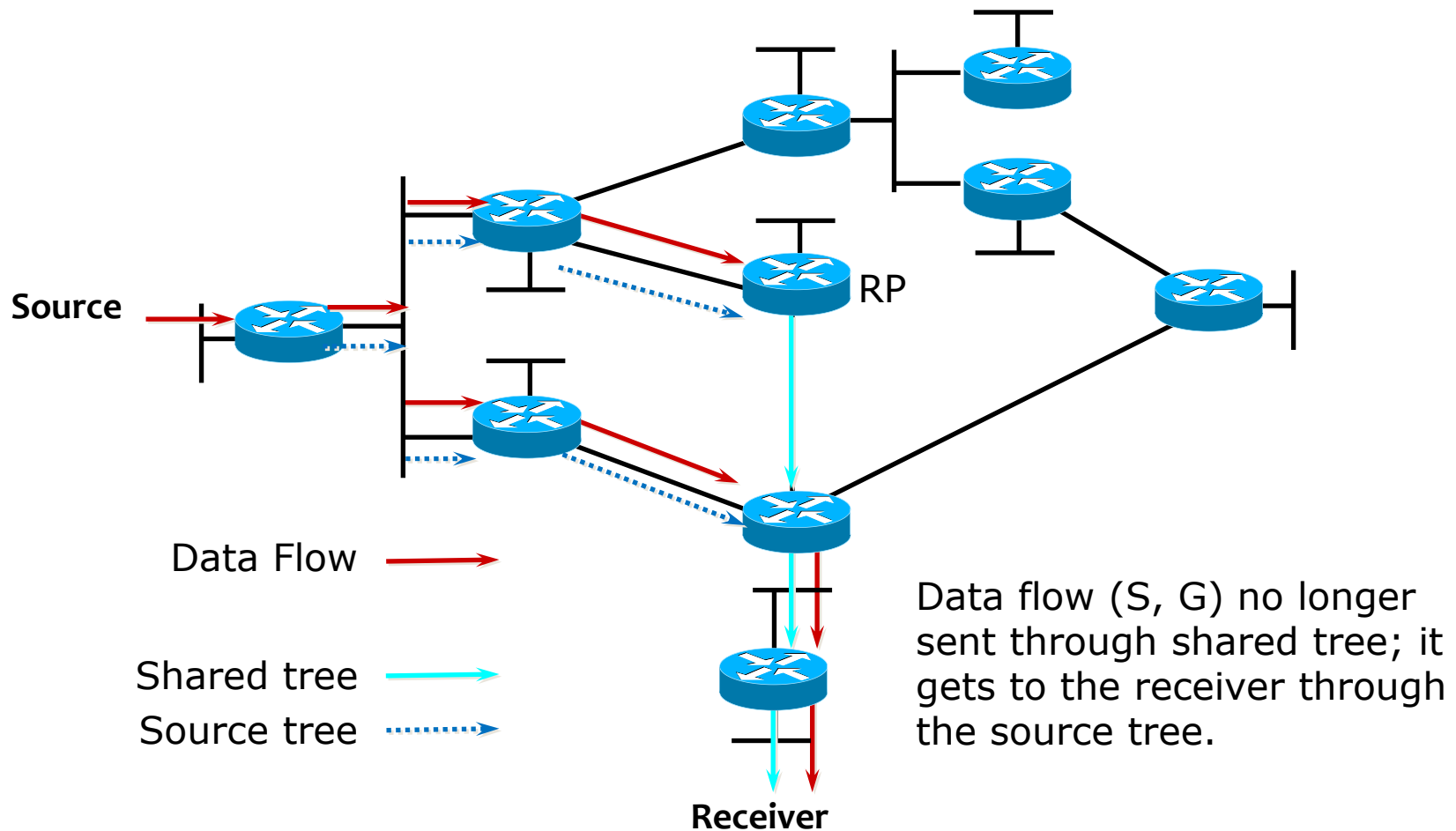


# PIM-SM: Commuting to source tree



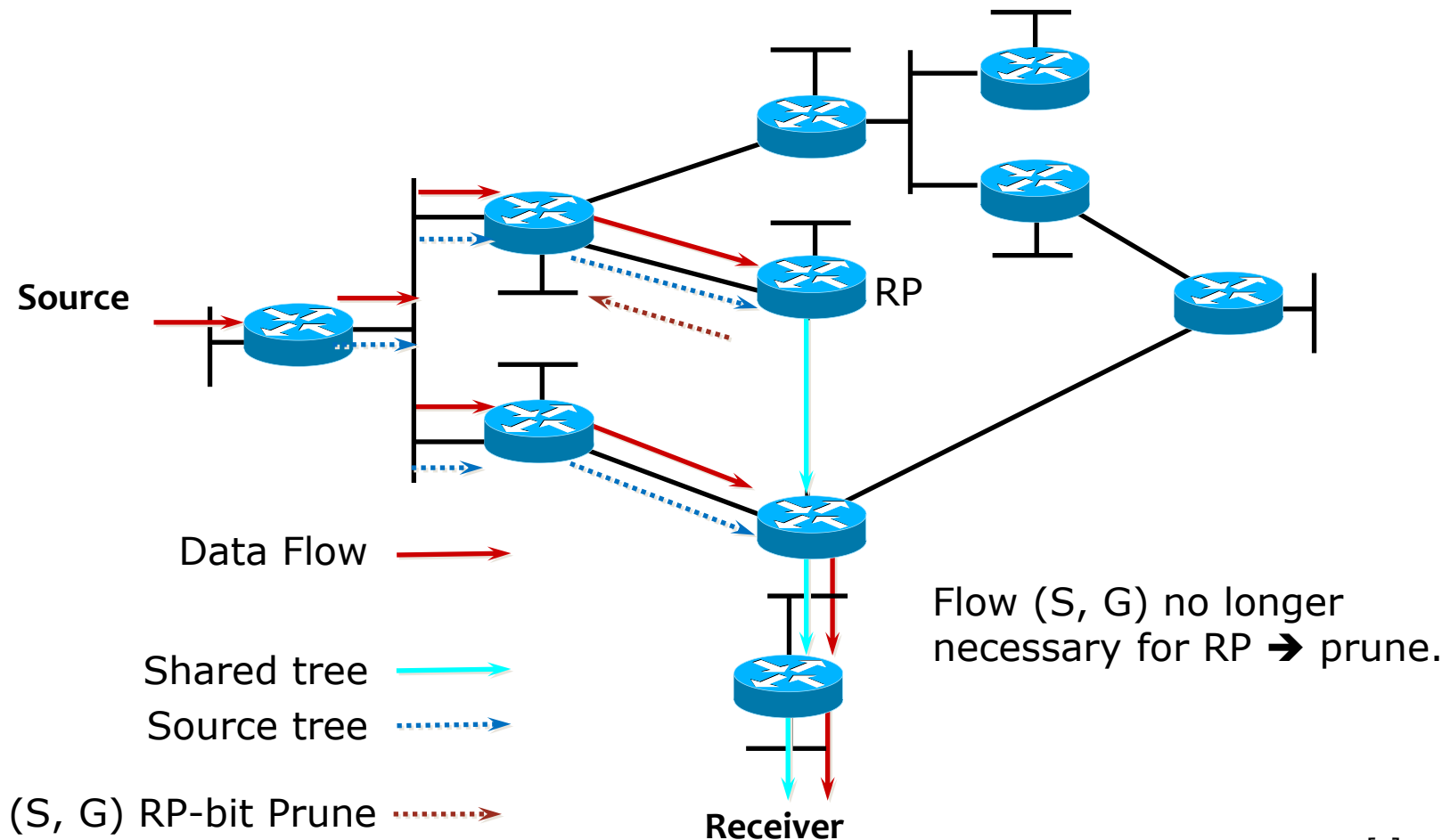
Source: [3]

# PIM-SM: Commuting to source tree



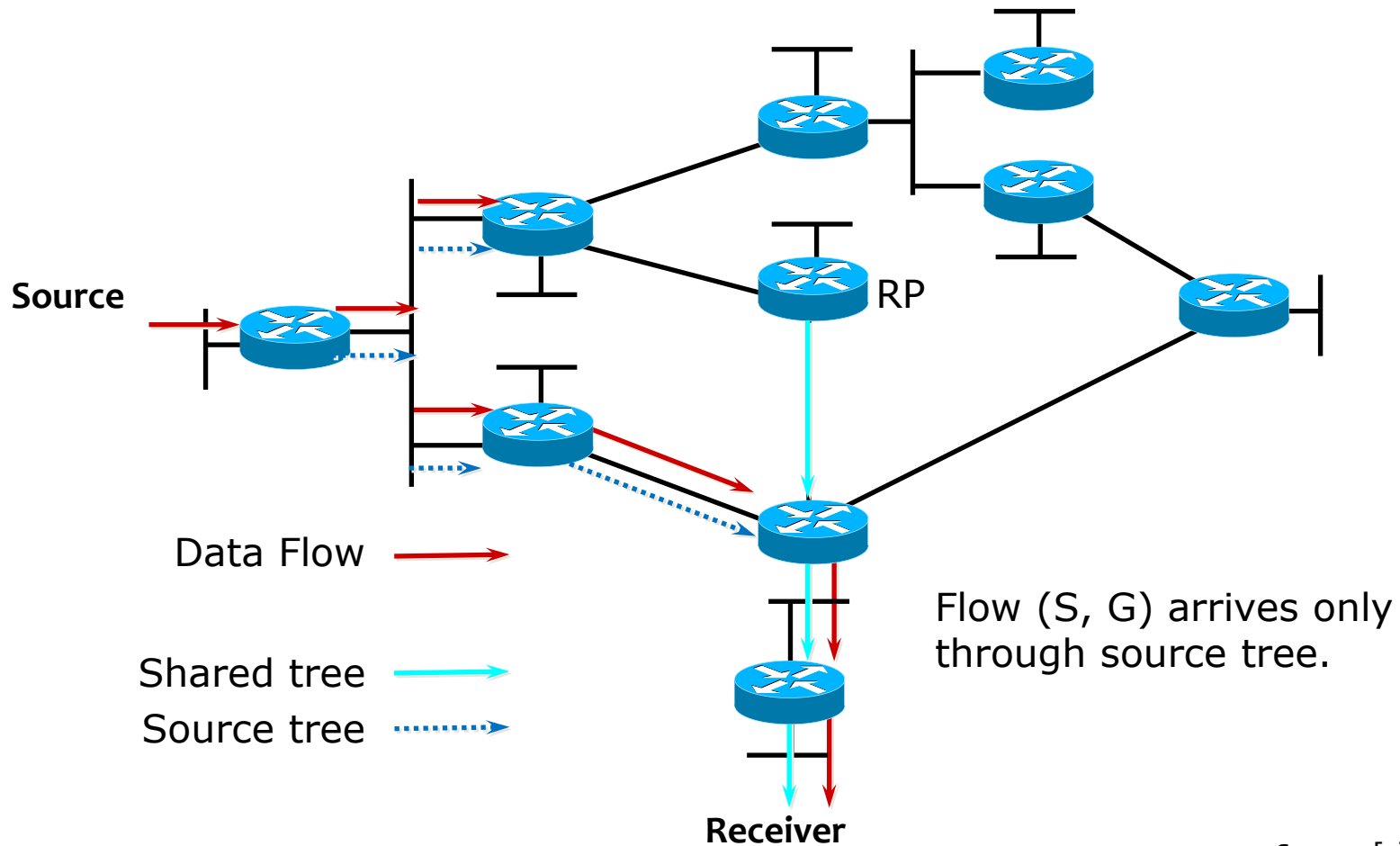
Source: [3]

# PIM-SM: Commuting to source tree



Source: [3]

# PIM-SM: Commuting to source tree



Source: [3]

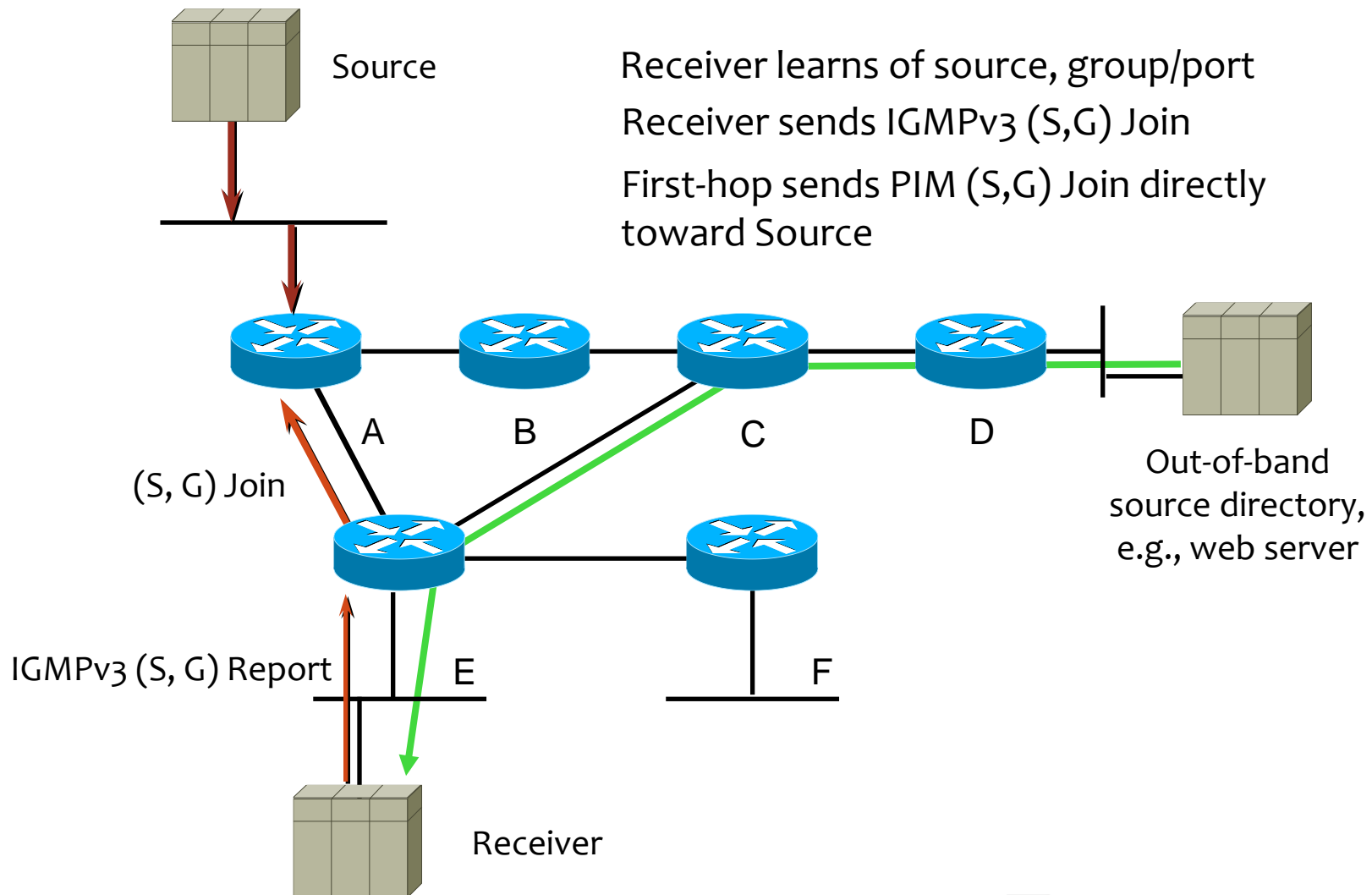
# Source Specific Multicast (SSM)

- Uses Source Trees only — **simple!**
- Assumes One-to-Many model
  - Most Internet multicast fits this model
- Hosts responsible for source discovery
  - Typically, via some out-of-band mechanism
    - E.g., web page
  - Eliminates need for RP and Shared Trees
  - Eliminates need for additional inter-domain protocol ☺
- 232.0.0.0/8 reserved exclusively for SSM
  - Building shared trees for these addresses disallowed
  - Can use other addresses through configuration

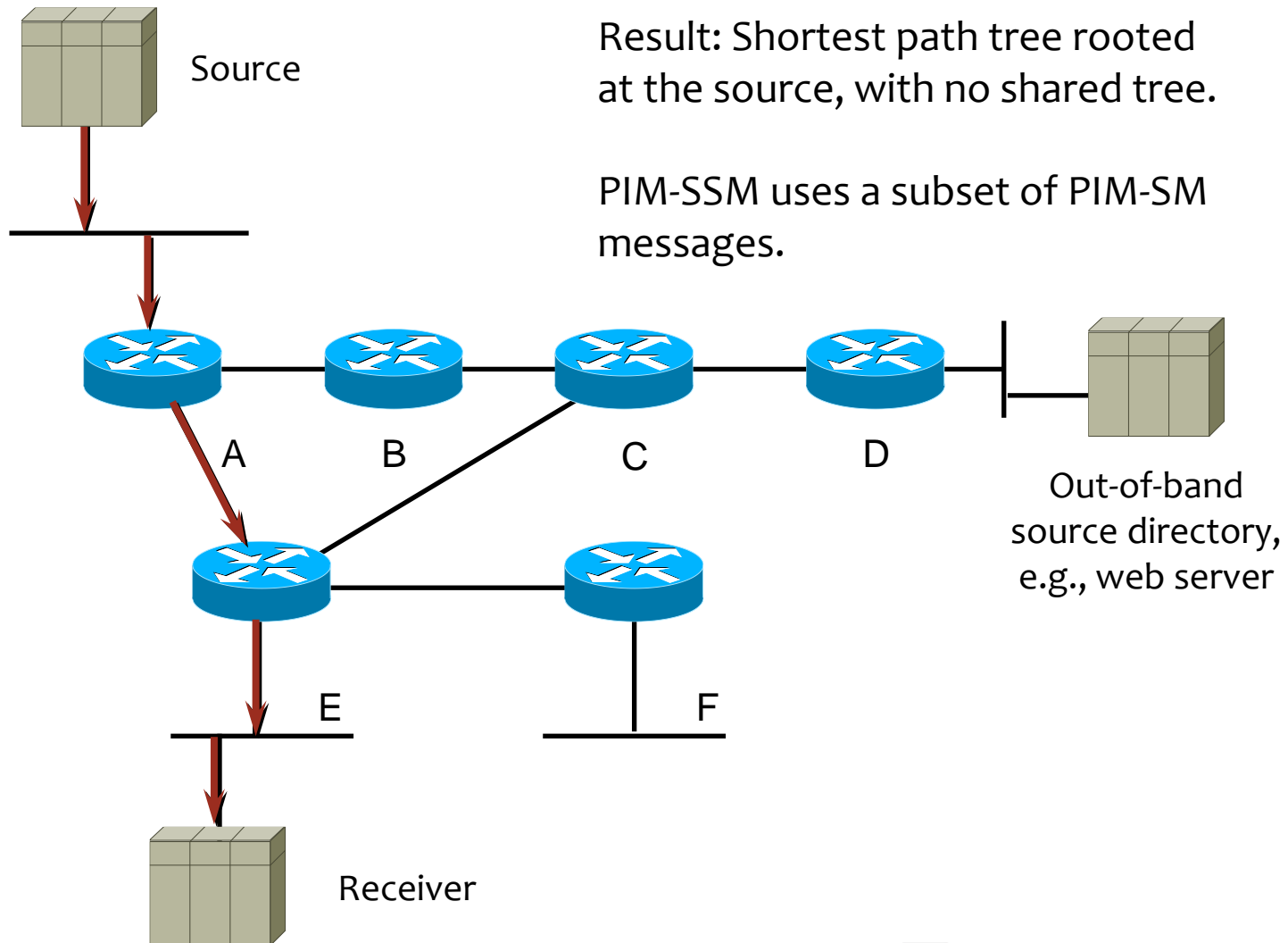
# SSM Overview

- Hosts join a specific source within a group
  - Content identified by specific (S,G) instead of (\*,G)
  - Hosts responsible for learning (S,G) information
  - Requires IGMPv3 or MLDv2
    - Alternatively, the last hop router may use static or DNS-based mappings to find sources in a group to join them all when a receiver uses IGMPv2 or v1
- Last-hop router sends (S,G) join toward source
  - Shared Tree is never Joined or used
  - Eliminates possibility of content jammers
  - Only specified (S,G) flow is delivered to host
- Makes address allocation trivial
  - Dissimilar content sources can use the same group without fear of interfering with each other

# PIM-SSM



# PIM-SSM



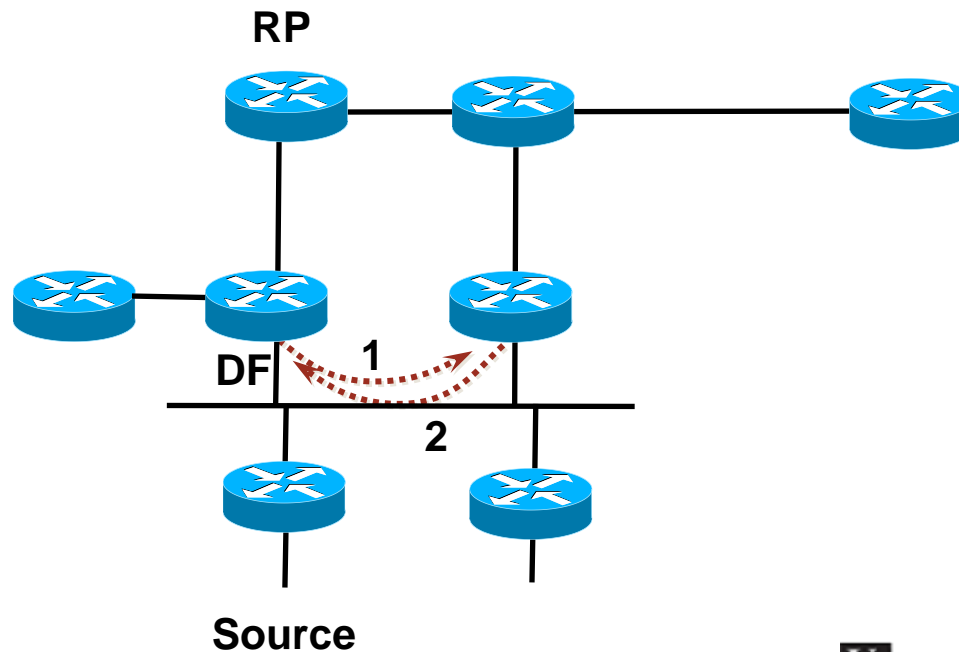


# Bidirectional PIM (BIDIR-PIM)

- Source trees work great in most cases, but not all
  - Groups with many senders create huge amounts of (S,G) state
  - E.g., large videoconference with many participants
- Solution: use a single shared tree for all sources
  - This idea appeared first in CBT
- Single bidirectional tree from RP to receivers
  - Single (\*,G) in each router for all sources in group
  - Traffic from sources / to receivers follows the same path if on the same branch of the RP
- RP Address needs not belong to an existing router
  - The RP performs no specific function in Bidir-PIM
  - May be any address in the subnet of an existing link (RP Link)

# BIDIR-PIM: Designated Forwarder

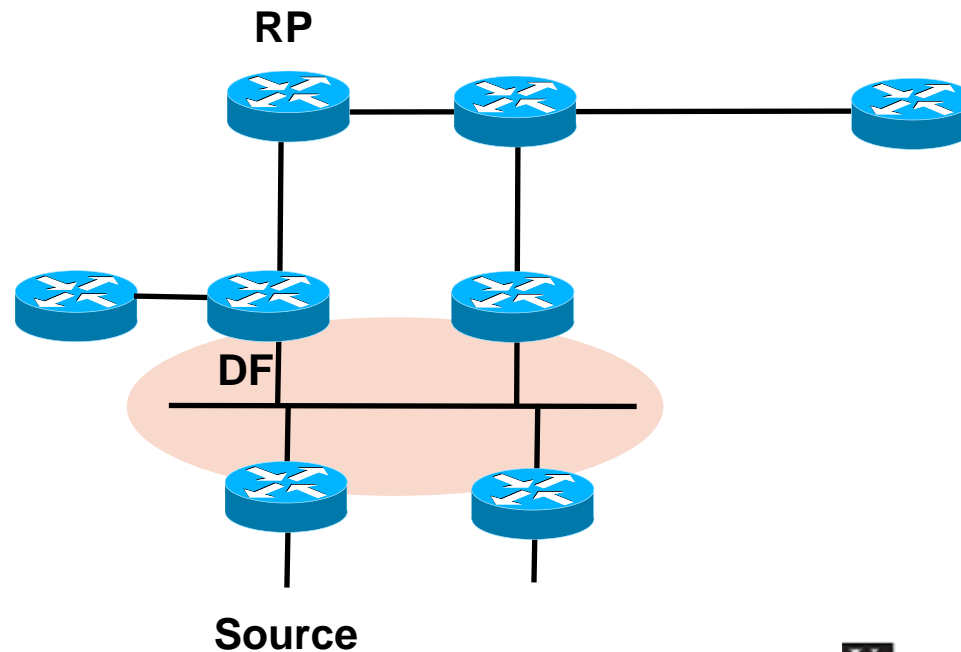
- Designated Forwarder elected per subnet based on
  - Higher preference
  - Lower metric to the RP
  - Higher IP address (for tie-breaking)
- Election is performed at RP discovery time
  - May be re-run when network conditions change



Multicast

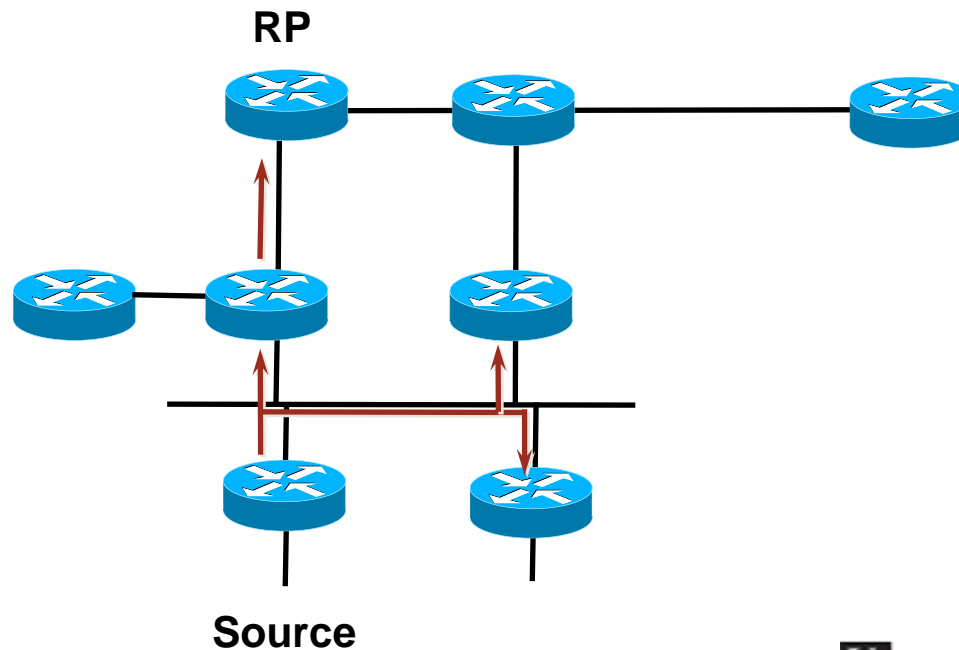
# BIDIR-PIM: Designated Forwarder

- The DF is the only router that
  - Forwards packets traveling downstream onto the link
  - Picks up packets traveling upstream off the link and forwards them to the RP
- The use of DFs allows multicast traffic to flow natively to the RPL without requiring source-specific state



# BIDIR-PIM: Sources

- Traffic is forwarded natively toward the RP, rather than registered / tunneled
- A DF without  $(*,G)$  state (i.e., not in the tree) for the group forwards the packet toward the RP

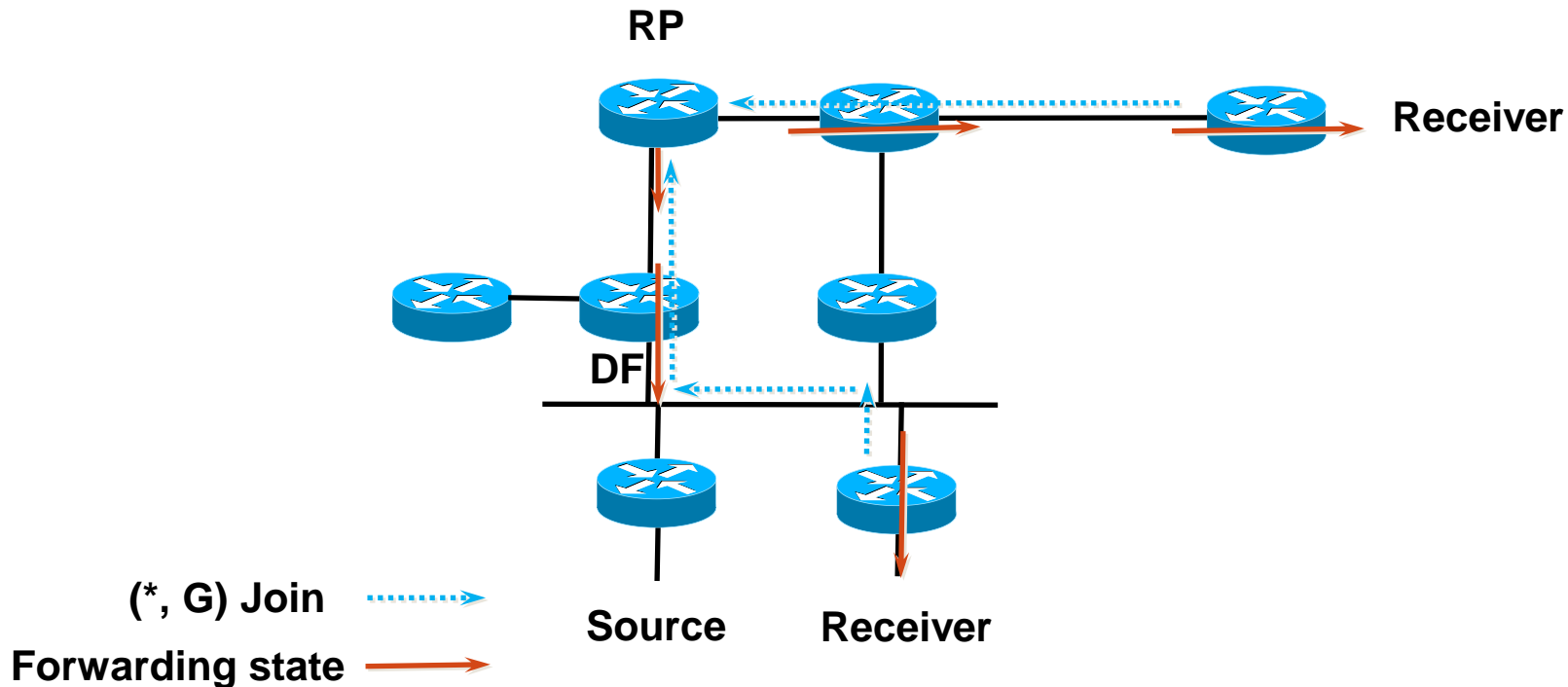


Source

Multicast

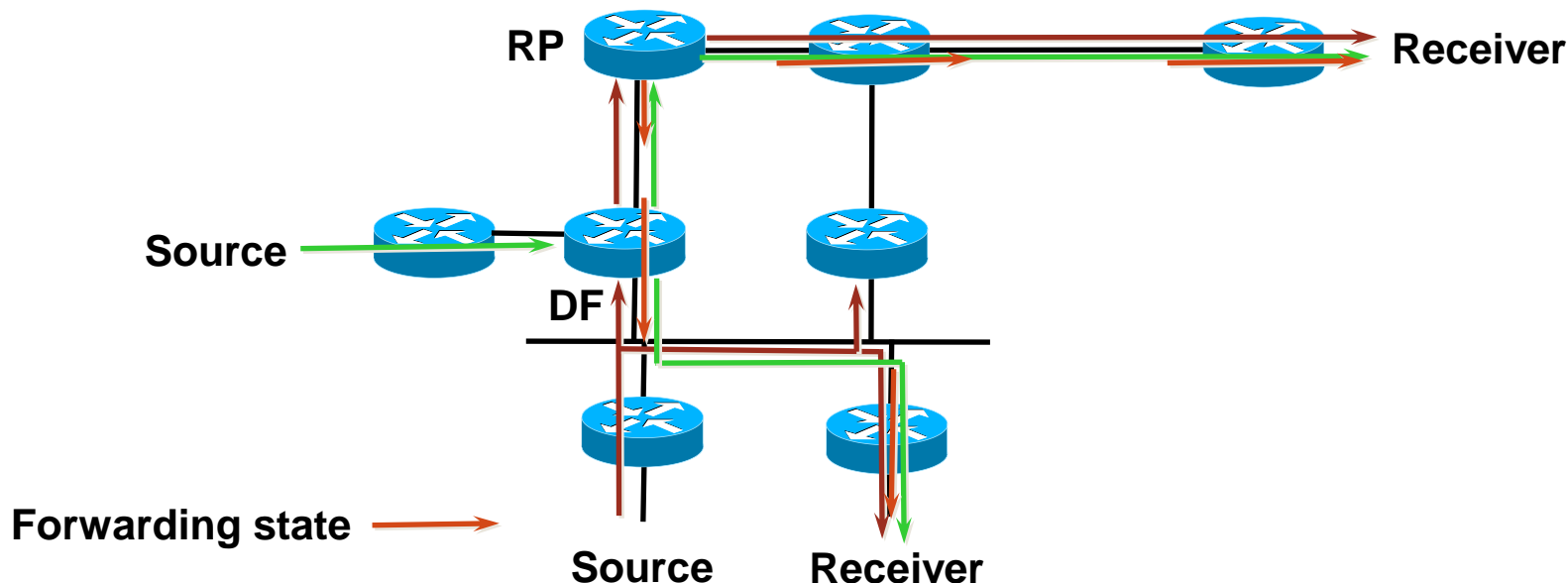
# BIDIR-PIM: Receivers

- Routers with receivers join toward the RP
- $(*,G)$  state created for the RPF interface toward the RP and DF interfaces



# BIDIR-PIM: Traffic Flow

- Packets sent from source to RP by the DFs
  - Can be forwarded directly toward interested receivers on a branch
- Traffic for all sources in group G forwarded based on the same  $(*,G)$  entry
  - But NEVER re-sent to the receiving interface



# PIM: Use Cases

- BIDIR-PIM
  - Many sources (many-to-many or many-to-few applications)
  - Drastically reduces total (S,G) state in network
- PIM-SSM
  - One-to-many applications
  - Simple
  - Eliminates the need for a RP
- PIM-SM
  - General purpose
  - Complex
- PIM-DM
  - Receivers in many (most) links
  - Eliminates the need for a RP
  - Scales poorly

# Multicast Routing Protocols: Summary of Characteristics

- Building delivery trees
  - Broadcast then prune
  - Direct joins toward sources
  - Rendezvous points
- Types of trees
  - Unidirectional, per-source, per-group trees
  - Unidirectional, per-group trees, shared by all sources
  - Bidirectional, per-group trees, shared by all sources

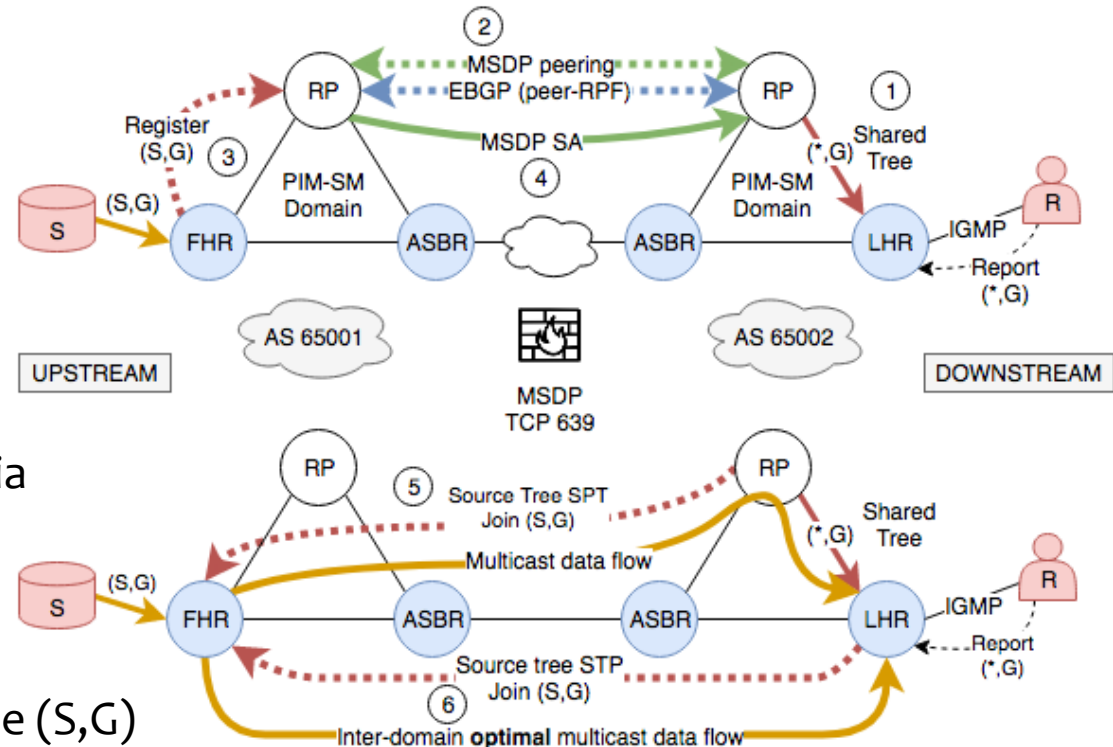


# Inter-Domain Multicast Protocols

- See [RFC5110](#), “Overview of the Internet Multicast Routing Architecture”
- Multiprotocol extensions to BGP-4 (MBGP; [RFC4760](#))
  - Reachability information for efficient multicast distribution and avoiding loops
- IPv4
  - Multicast Source Discovery Protocol (MSDP; [RFC3618](#))
  - Connect multiple PIM-SM domains together
- IPv6
  - Use unicast-prefix-based multicast groups
  - Embed RP address in multicast address ([RFC3956](#))
- Alternative (IPv4 & IPv6): use SSM and be happy ☺

# Example inter-domain advertisement

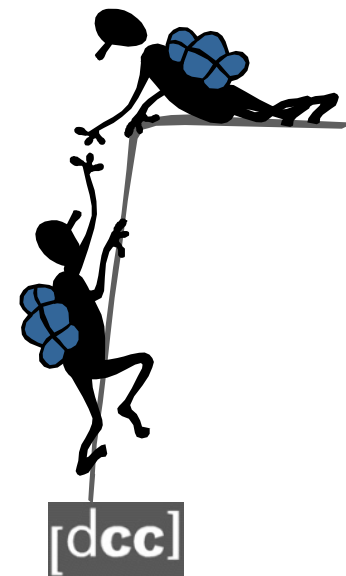
1. Receiver joins  $(*,G)$ ; LHR already in shared tree
2. MSDP peering; Source appears, FHR registers to local RP
3. MSDP shares information about source with peers via Source Active message; MBGP to ensure peer-RPF check PASS (no loops)
4. Remote RP joins the source  $(S,G)$  and starts delivering multicast traffic via shared tree to interested receivers
5. LHR may decide to join the Source (SPT switchover is enabled), resulting in inter-domain optimal multicast data flow.



Source: Cisco

# Reliability

- Objective: no losses, no duplicates or data corruption
- Using TCP-like acknowledgements would not scale
  - Arbitrary number of members → arbitrary number of ACKs (ACK implosion)



# Reliability

## Solution 1

- Designated Routers along the path cache packets
- Packets with (per source) sequence number
- Receiver sends NACK when it detects losses
- When it receives a NACK, DR (or source) retransmits
- Ex.: MTP ([RFC1301](#)), PGM ([RFC3208](#)), SRM (Scalable Reliable Multicast)

## Solution 2

- Redundancy
- Two options
  - Sending N copies of packet
  - Using Forward Error Correction Codes
- Ex.: [RFC3453](#), [RFC3048](#) (also uses NACK)

See [Reliable multicast transport IETF group](#)

# The end

# Acronyms – Multicast

- CBT – Core-Based Trees
- IGMP – Internet Group Management Protocol
- IIF – Incoming Interface
- MALLOC – Multicast-Address Allocation
- MASC – Multicast Address Set Claim
- OIF – Outgoing Interface
- PIM – Protocol Independent Multicast
- PIM-DM – PIM Dense Mode
- PIM-SM – PIM Sparse Mode
- RP – Rendezvous point
- SA – Source Activation