Advanced Internet Technologies

Chapter 6
IP Multicast

Prof. Dr.-Ing. Georg Carle

Chair for Computer Networks & Internet
Wilhelm-Schickard-Institute for Computer Science
University of Tübingen

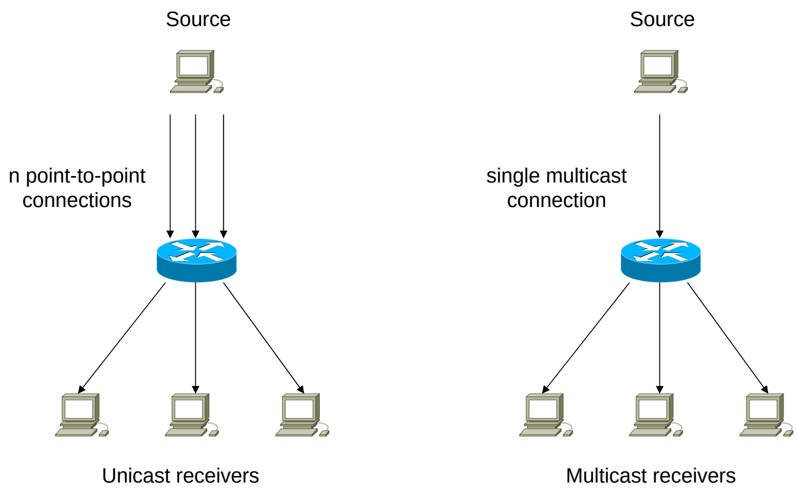
http://net.informatik.uni-tuebingen.de/
carle@informatik.uni-tuebingen.de

Chapter 6
IP Multicast

Group Communication

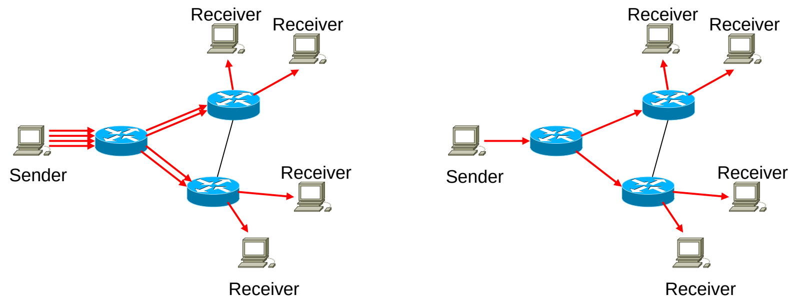
- Multiple partners communicate in a closed group
- Types of group communication
  - Unicast: 1:1
  - Concast: m:1
  - Multicast: 1:m
  - Multipeer: m:n (typically emulated using multicast)
- Other types of communication
  - Broadcast
  - Anycast
- Scalability
  - Group size
  - Topology
  - Dynamics

Principles of Multicast



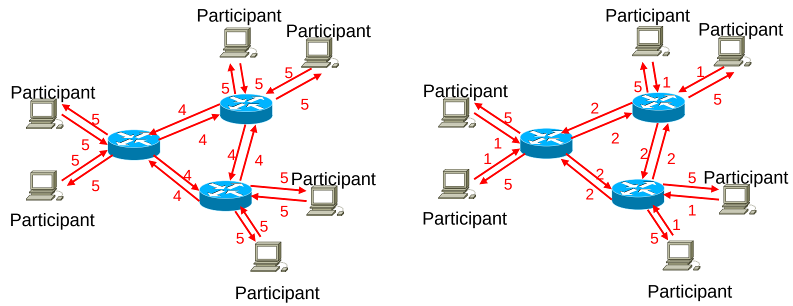
Typical Scenarios: one-to-many

- TV broadcast
- time synchronization (NTP)
- distribution of data, e.g., stock exchange rates



Typical Scenarios: many-to-many

- video conferences
- multiplayer games



Aspects of Group Communication

- Addressing
  - IP Multicast addresses
- Group maintenance
  - Internet Group Management Protocol
- Routing
  - Distribution trees
  - Routing protocols

IP Multicast Addressing

- Multicast addresses = Class D addresses
  - address range: 224.0.0.0/4
  - only for destination address
  - source address is still the unicast source address
- Link-Local multicast addresses
  - only available in the subnet (will not be forwarded)
  - address range: 224.0.0.0/24
  - reserved addresses (examples):
    - 224.0.0.1 - all systems
    - 224.0.0.2 - all routers
    - 224.0.0.5 - OSPF routers
    - 224.0.0.6 - OSPF designated routers

Address Scoping

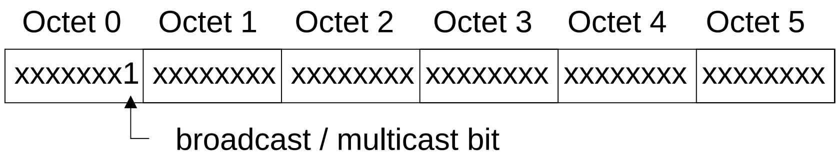
- Globally scoped addresses
  - 224.0.1.0 - 238.255.255.255
  - are to be used globally in the internet
- Source-specific multicast
  - 232.0.0.0/8
- GLOP addresses (RFC2770)
  - 233.0.0.0/8
  - reserved for statically defined addresses by organizations that already have an AS number reserved
  - address: 233.<AS>.0/24
- Administratively scoped addresses
  - 239.0.0.0/8
  - like RFC1918 addresses for local use only
  - not routed in the internet

TTL Thresholds

- same principle as in IP unicast
- thresholds are also used to limit multicast traffic to a particular region

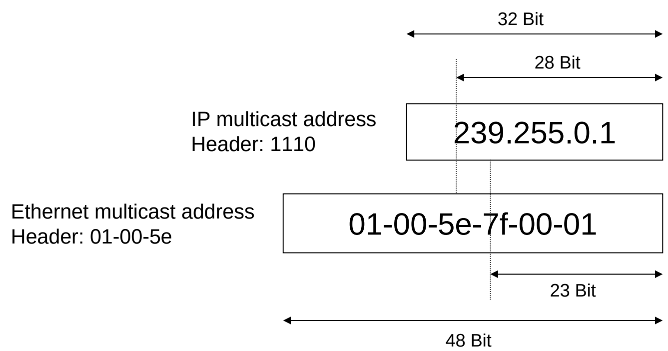
TTL scope	**Initial TTL value**	**TTL threshold**
Local net	1	-
Site	15	16
Region	63	64
World	127	128

Multicast Ethernet Addresses



IP Multicast to Ethernet Address Mapping

- Why only 23 bit?
  - In the early 90s Steve Deering tried to get 16 OUIs from the IEEE but could not pay for it.
- Any problems?
  - 32 IP multicast addresses can be mapped to a single ethernet address. This may lead to performance problems!



Internet Group Management Protocol (IGMP)

- "The membership of a host group is dynamic; that is, hosts may join and leave groups at any time. There is no restriction on the location or number of members in a host group. A host may be a member of more than one group at a time. A host need not be a member of a group to send datagrams to it." [RFC1112]

- IGMPv1 (RFC 1112)
  - Message Format
  - Query-Response Process
  - Join Process
  - Leave Process
- IGMPv2 (RFC 2236)
  - Message Format
  - Enhanced Leave Process
- IGMPv3 (RFC 3376)
  - Ideas
  - Message Format

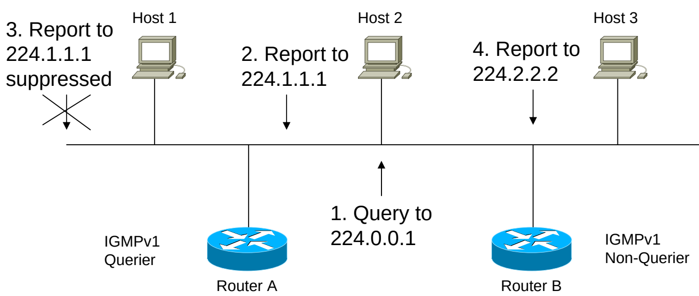
IGMPv1 – Message Format

- Version: 1
- Type: message type
  - Membership query
  - Membership report
- Checksum: for the whole IGMP packet
- Group address:
  - Multicast address for membership report
  - Null for membership query

Version	Type	Unused	Checksum
Group address			

IGMPv1 – Query-Response Process

1. Router A (IGMP querier) sends periodically (every 60 sec.) membership query messages to all multicast hosts (224.0.0.1)
2. Host 2 responses first by sending a membership report for group 224.1.1.1
3. Host 1 (also member in 224.1.1.1), receives this report and suppresses any additional report
4. Host 3 reports to 224.2.2.2.



IGMPv1 – Querier

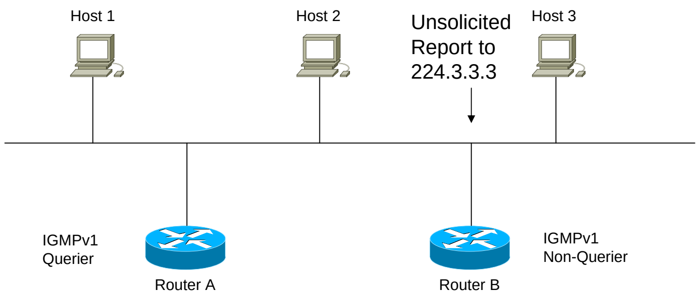
- IGMPv1 does not define an election mechanism

- Solution by the multicast routing protocol:
  Designated Router (DR) is also querier

- IGMPv2 defines its own election mechanism

IGMPv1 – Join Process

- technically, a JOIN is a membership report
- join is required only to **receive** multicast traffic
- **to send** multicast packets, no join is required first
  - ↳ problems in connecting sparse and dense mode networks



IGMPv1 – Leave Process

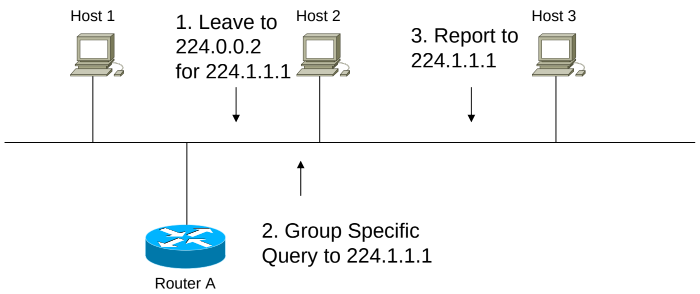
- There is no Leave-Group-Message in IGMPv1!
- Solution: time-out
  - every 60 sec. a query is sent to the group
  - if there is no report after 3 queries, the group state is removed
- Problem: leave latency up to 3 min.

IGMPv2 – Message Format

- Type: message type
  - Membership query (0x11)
    - General query
    - Group-Specific query
  - V1 membership report (0x12)
  - V2 membership report (0x16)
  - Leave group (0x17)
- Maximum response time field
  - for tuning of membership reports and leave latency
- Checksum: for the whole IGMP packet
- Group address:
  - Multicast address for membership report
  - Membership query: Null if general query, multicast address else

Type	Max.Resp.Time	Checksum
Group Address		

IGMPv2 – Leave Process



IGMPv3

- Problem: every member of a group gets all the traffic to this group
- Extension of the join/leave messages by (S,G)-pairs
- Support for source filtering
- Basis for Source-Specific Multicast
- Example:
  - Join (1.1.1.1, 224.1.1.1)
  - Leave (2.2.2.2, 224.1.1.1)

IGMPv3 – Query Message Format

- Type = 0x11: query
- Maximum response time field
- Checksum: for the whole IGMP packet
- Group address: Null if general query, multicast address else
- S: S flag, indicates that processing by routers is suppressed
- QRV: Querier Robustness Value, affects timers and number of retries
- QQIC: Querier's Query Interval Code, query interval
- Number of sources: # of sources in this query
- Source address [1..N]: address of source

Type = 0x11		Max.Resp.Time	Checksum	
Group Address				
	S	QRV	QQIC	Number of sources (N)
Source Address [1]				
Source Address [2]				
. . .				
Source Address [N]				

IGMPv3 – Report Message Format

- Type = 0x22: report
- Checksum: for the whole IGMP packet
- Number of group records: # block fields containing information regarding the sender's membership with a single group
- Record type: group record type
  - MODE_IS_INCLUDED - to receive on from these sources
  - MODE_IS_EXCLUDED - to receive from any sender but from these sources
- Number of sources: # of sources
- Group address: multicast address in this record
- Source address [1..N]: address of source
- Aux. data len / Auxiliary data: for future enhancements

Type = 0x22	Reserved	Checksum
Reserved		Number of group records (N)
Group record [1]		
Group record [2]		
. . .		
Group record [N]		

Record type	Aux. data len	Number of sources (N)
Group address		
Source address [1] Source address [2] . . . Source address [N]		
Auxiliary data		

Layer-2 Multicast Mechanisms

- Normal case: multicast = broadcast, i.e. flooding trough the LAN
- IGMP snooping
  - Intelligent switches process all multicast packets, look for IGMP messages and analyze them
  - Prerequisite for a broad use: layer-3-aware switches
- Cisco Group Management Protocol (CGMP)
  - Intelligence only at the router, which informs 'its' local switches

IGMP snooping

- Join
  - A host sends an IGMP join for group 224.1.2.3 to 0x0100.5E01.0203. Because there is no entry in the CAM table of the switch for this address, the packet is flooded to all ports (including the internal CPU port).
  - The CPU receives the packet and decodes the IGMP information. Then it generates an CAM entry and adds the ports of the CPU, the host and the router.
- Leave
  - A host sends a leave-group message to 224.0.0.2 (All-Routers)
  - The CPU of the switch gets the message and sends a general query back to this port (there may be more than one host behind the same port!)
  - If there is no answer to the query, the port is removed from the CAM entry.
  - If there are no more ports in the CAM entry (except CPU and the router), the CAM entry is discarded and a leave-group message is sent to the router.

IGMP snooping II

- Performance
  - IGMP packets do use the same group address as data packets so the CPU has to scan EVERY packet which travels over this group for IGMP messages. This may result in performance problems at simple layer-2 switches.
  - Solution: Layer-3-aware switching. Special ASICs scan for IGMP messages and only these IGMP messages are forwarded to the CPU port.
- Question: How does the switch know to which port(s) the router(s) is(are) connected?
  - ‚It's magic!'
  - At least the switch can watch for general query messages
  - Typically, it watches also for OSPF hellos, PIMv1/v2 hellos, DVMRP probes, IGMP queries, CGMP self-joins, HRSP messages

CGMP

- Cisco Group Management Protocol
- Messages via the well-known CGMP MAC multicast address 0x0100.0cdd.dddd
- The router processes the information for the switch(es)
- No processing power at the switch is required

CGMP Messages

GDA (Group Destination Address)	USA (Unicast Source Address)	Join/Leave	Meaning
Mcst MAC	Client MAC	Join	Add port to group
Mcst MAC	Client MAC	Leave	Delete port from group
0000...0000	Router MAC	Join	Assign router port
0000...0000	Router MAC	Leave	Deassign router port
Mact MAC	0000...0000	Leave	Delete group
0000...0000	0000...0000	Leave	Delete all grouos

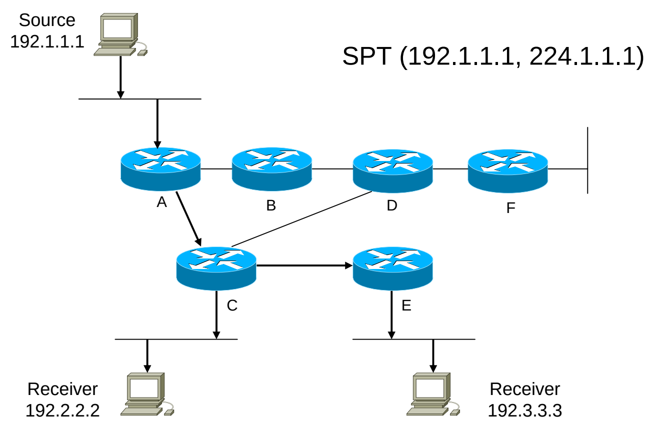
Multicast Forwarding

- Source address is used for forwarding decision (unlike destination address in unicast)
- A distribution tree is created to let the packets flowing from the root to the leaves
- Reverse Path Forwarding (RPF)
  - Check on the basis the source address whether the package arrived at the expected interface (depending upon the multicast routing protocol, there are different sources for the RPF check)
  - The packet is forwarded if RPF check is OK, otherwise the packet is dropped

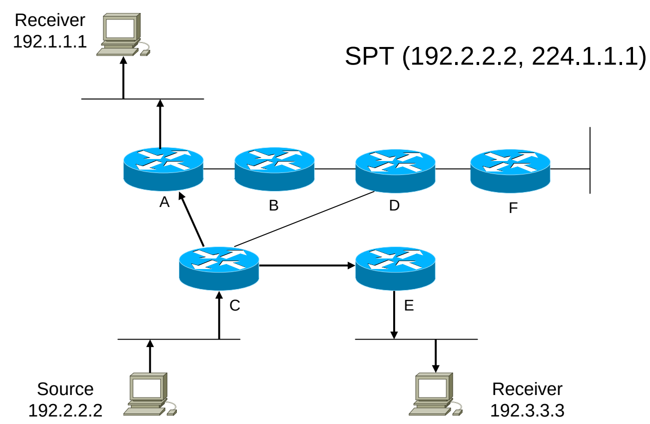
Multicast Distribution Trees

- Why a tree?
  - IP unicast: single path from source to destination
  - IP multicast: ‚branched‘ path = tree
- Source tree
  - Also known as Shortest Path Tree (SPT)
  - Different tree for each source
  - Calculation e.g. via ‚Steiner tree‘
  - Source is the root of the tree
  - Notation: (S,G)
    - S ... source IP address
    - G ... multicast group address

Shortest Path Tree



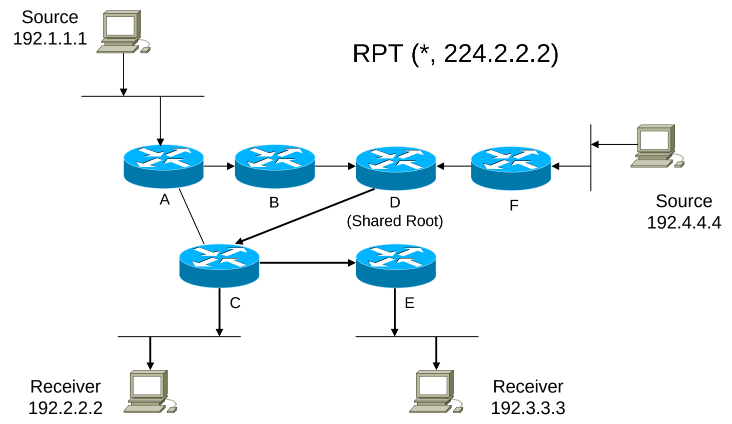
Shortest Path Tree II



Multicast Distribution Trees II

- Shared tree
  - single root for each source (Rendezvous Point (RP) or Core)
  - also known as RP Tree (RPT) or Core-Based Tree (CBT)
  - Notation: (*,G)
  - Bidirectional shared trees
    - can be used for data transfer up toward and down from the RP
  - Unidirectional shared trees
    - different path toward the RP
    - via SPT (PIM Sparse Mode) / via IP unicast (CBT)

Shared Distribution Tree



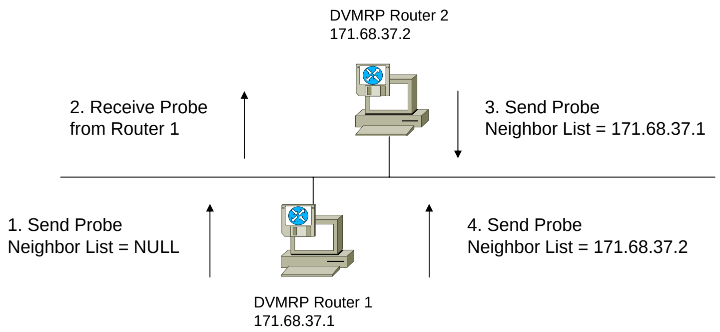
Multicast Routing Protocols

- Dense mode protocols
  - **DVMRP**, **PIM-DM**
  - ‘push‘ principle
- Sparse mode protocols
  - **PIM-SM**, **SSM**, CBT
  - ‘pull’ principle
- Link-state protocols
  - MOSPF
  - Mixture of dense and sparse mode

DVMRP

- Distance Vector Multicast Routing Protocol
- The ‘old‘ MBone was based on DVMRP
- Characteristics:
  - Distance vector protocol (like RIP)
  - Periodical updates of routing information (every 60 sec.)
  - Infinity = 32 hops (RIP: 16)
  - Classless
  - Flood-and-prune mechanism (every 2 min.)
- Scalability?
  - Limits of distance vector protocols
  - Update 50.000 routes every 60 sec.?

DVMRP – Neighbor Discovery



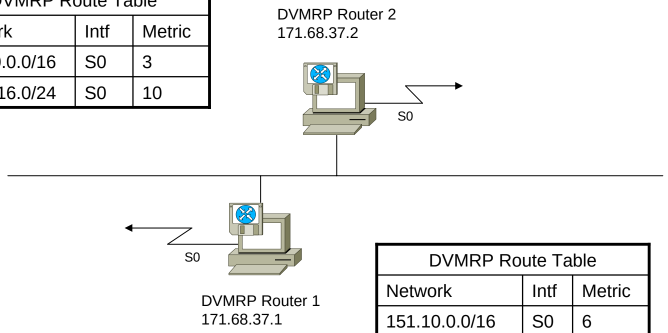
DVMRP – Routing Table

- DVMRP maintains its own routing table besides unicast routing table
- Responsible for
  - building the source distribution trees
  - multicast forwarding (RPF check)
- Example:

```
DVMRP Routing Table - 8 entries
130.1.0.0/16 [0/3] uptime 00:19:03, expires 00:02:13
    via 135.1.22.98, Tunnel0, [version mrouted 3.8]
  [flags: GPM]
```


DVMRP – Route Exchange

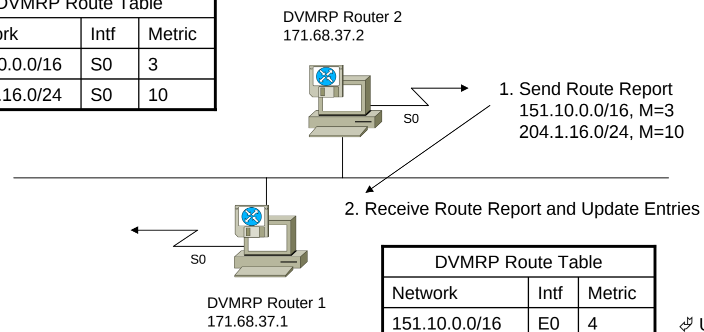
DVMRP Route Table		
Network	Intf	Metric
151.10.0.0/16	S0	3
204.1.16.0/24	S0	10



DVMRP Route Table		
Network	Intf	Metric
151.10.0.0/16	S0	6
198.14.32.0/24	S0	3

DVMRP – Route Exchange II

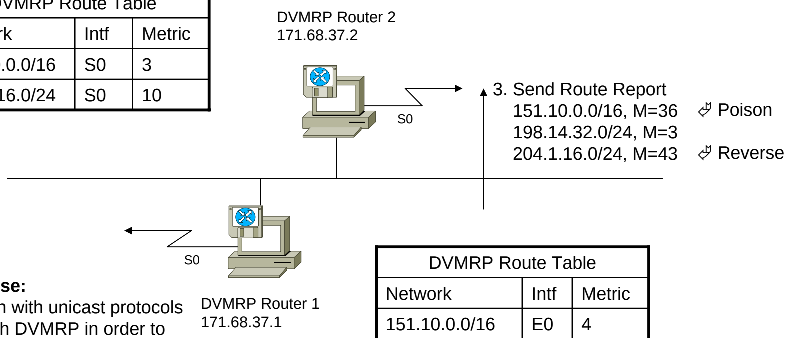
DVMRP Route Table		
Network	Intf	Metric
151.10.0.0/16	S0	3
204.1.16.0/24	S0	10



DVMRP Route Table			
Network	Intf	Metric	
151.10.0.0/16	E0	4	Updated
198.14.32.0/24	S0	3	
204.1.16.0/24	E0	11	Added

DVMRP – Route Exchange III

DVMRP Route Table		
Network	Intf	Metric
151.10.0.0/16	S0	3
204.1.16.0/24	S0	10



DVMRP Route Table		
Network	Intf	Metric
151.10.0.0/16	E0	4
198.14.32.0/24	S0	3
204.1.16.0/24	E0	11

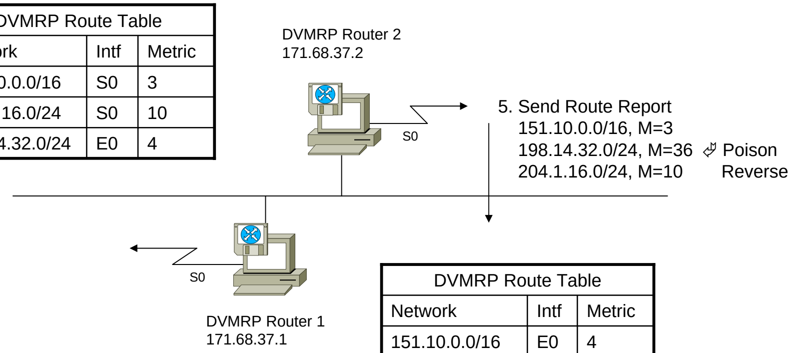
Poison Reverse:
Differently than with unicast protocols PR is used with DVMRP in order to tell the upstream neighbor that my router is ‚downstream' in the multicast distribution tree.

DVMRP – Route Exchange IV

4. Receive Route Report and Update Entries

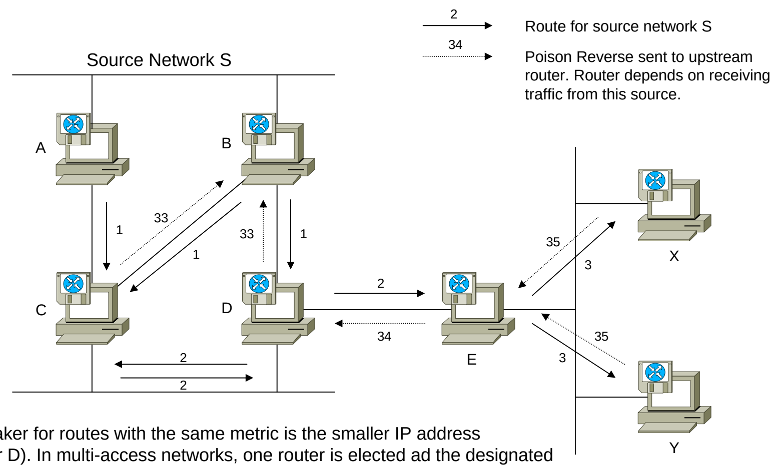
DVMRP Route Table		
Network	Intf	Metric
151.10.0.0/16	S0	3
204.1.16.0/24	S0	10
198.14.32.0/24	E0	4

Added



DVMRP Route Table		
Network	Intf	Metric
151.10.0.0/16	E0	4
198.14.32.0/24	S0	3
204.1.16.0/24	E0	11

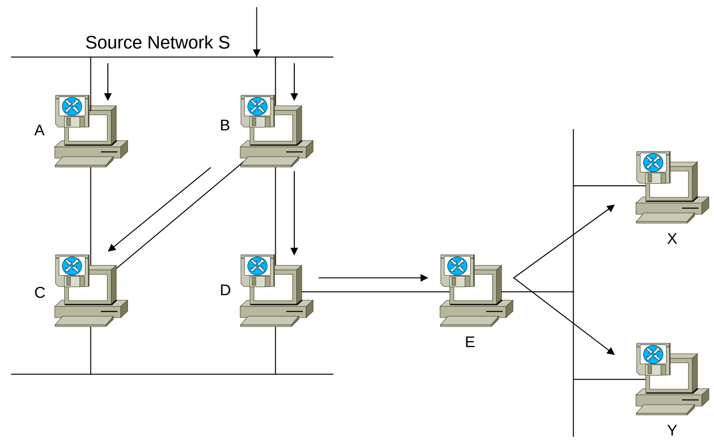
DVMRP – Truncated Broadcast Tree



Tiebreaker for routes with the same metric is the smaller IP address (Router D). In multi-access networks, one router is elected ad the designated forwarder (Metric + smaller IP address) (Router D, E).

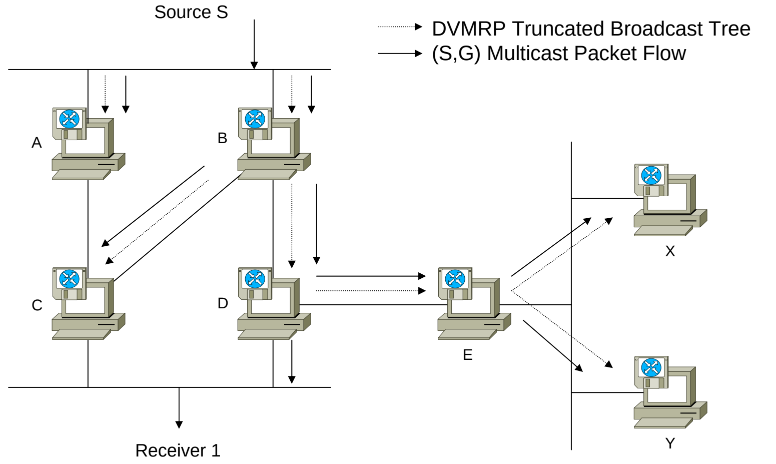
DVMRP – Distribution Tree

- For source network S



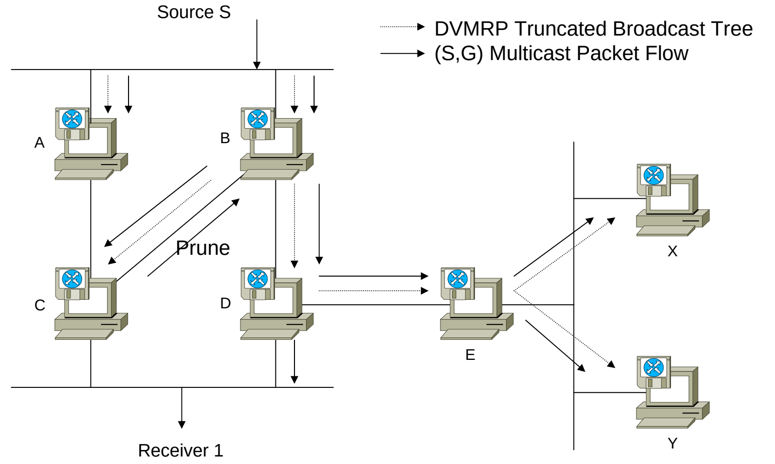
DVMRP – Pruning

- Initial flooding



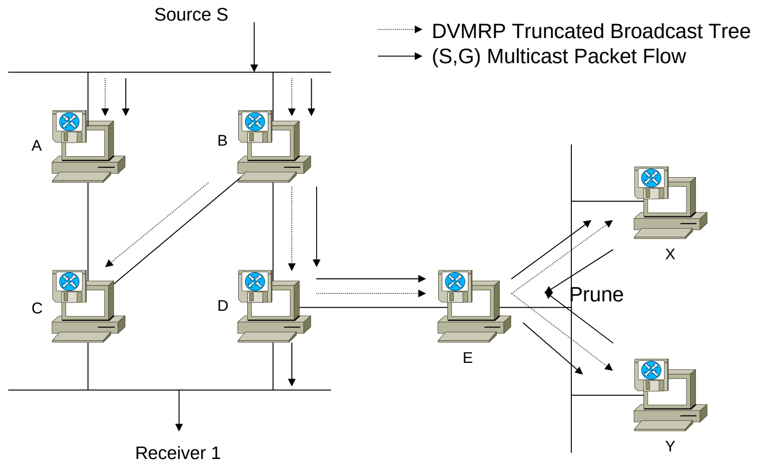
DVMRP – Pruning II

- Step 1 (C is not the DR)



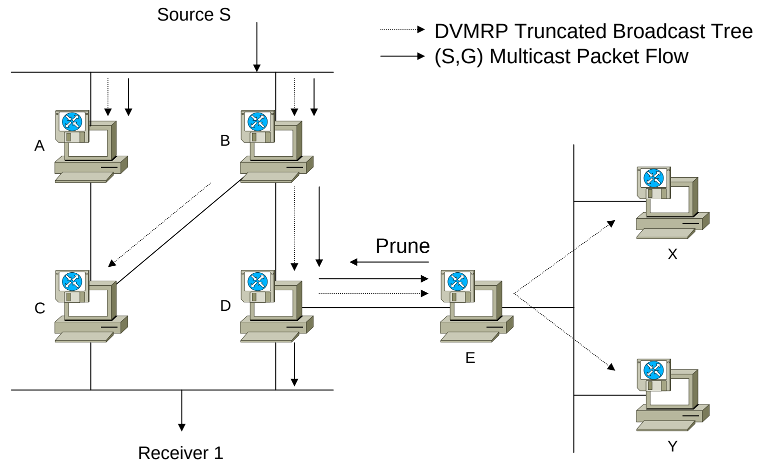
DVMRP – Pruning III

- Step 2 (X, Y without connected receivers)



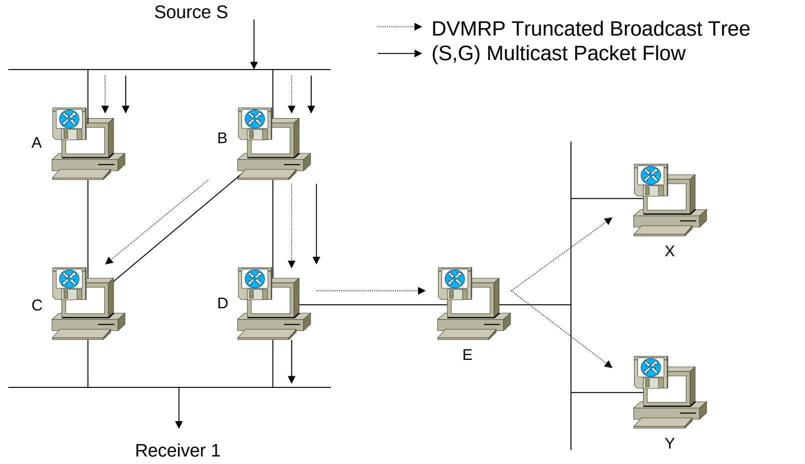
DVMRP – Pruning IV

- Step 3 (E has pruned all (S,G) traffic)

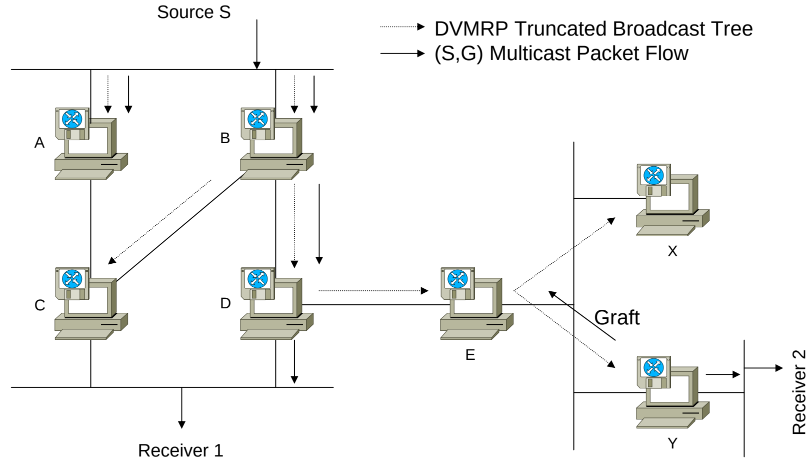


DVMRP – Pruning V

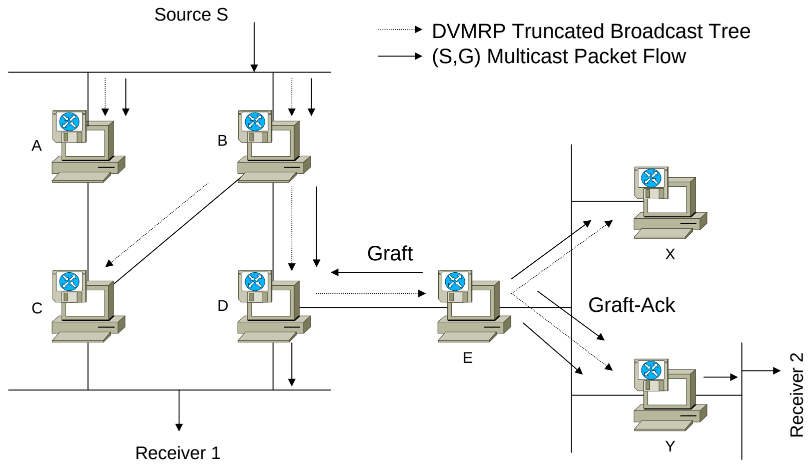
- Final pruned state



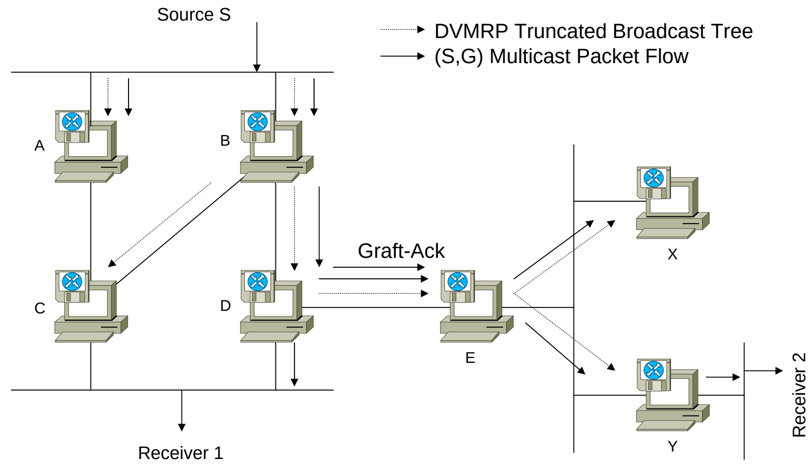
DVMRP – Grafting



DVMRP – Grafting II



DVMRP – Grafting III



Protocol Independent Multicast (PIM)

- PIM neighbor discovery
  - By sending hello messages to 224.0.0.13 (All-PIM-Routers)
  - PIMv1: to 224.0.0.2 (All-Routers)
  - Hello interval: 30 sec.
  - Goal: Creation of a table with neighborhood relations
  - And: election of a designated router (DR) (tiebreaker is the highest IP address)

- Example

```
reliant> sh ip pim neighbor
PIM Neighbor Table
Neighbor Address  Interface          Uptime    Expires   Ver  Mode
131.188.7.8       Vlan7              2w4d      00:01:23  v2
131.188.7.7       Vlan7              2w4d      00:01:17  v2
131.188.7.89      Vlan7              2w4d      00:01:28  v2
131.188.7.211     Vlan7              2w4d      00:01:19  v2        (DR)
131.188.7.131     Vlan7              2w4d      00:01:21  v2
131.188.7.88      Vlan7              2w4d      00:01:18  v2
131.188.7.3       Vlan7              2w4d      00:01:25  v2
131.188.7.66      Vlan7              2w4d      00:01:21  v2
131.188.7.5       Vlan7              2w4d      00:01:17  v2
131.188.7.58      Vlan7              2w4d      00:01:35  v2
```

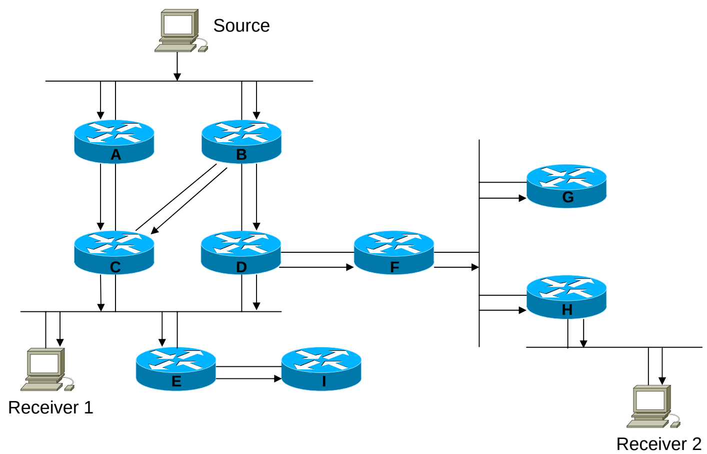

PIM Dense Mode

- Characteristics
  - Protocol independent (uses the unicast routing table for RPF checks)
  - ‘push’ principle
  - Flood-and-prune mechanism (every 3 min.)
  - Classless (so far the unicast routing protocol is classless)

- Source distribution tree
  - Differently than DVMRP (minimum spanning tree is built by its own multicast routing table and the poison reverse mechanism) PIM-DM uses its neighborhood information
  - An initial SPT is built with the input interface toward the source and all other neighbors as destinations
  - This initial SPT is also known as broadcast tree
  - Problem: duplicated packets if there is more than one upstream router
  - Tree is cut back gradually

PIM-DM – Distribution Tree

- Initial flooding

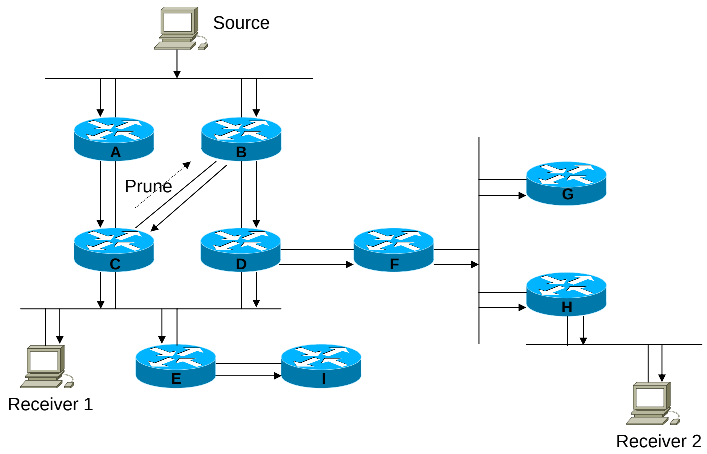


PIM-DM – Pruning

- Conditions
  - Traffic arrived at a non-RPF interface
  - Leaf router without directly connected receivers
  - Non-leaf router which received a prune over a point-to-point link
  - Non-leaf router which received a prune over a LAN segment and no other neighbor has overwritten the prune

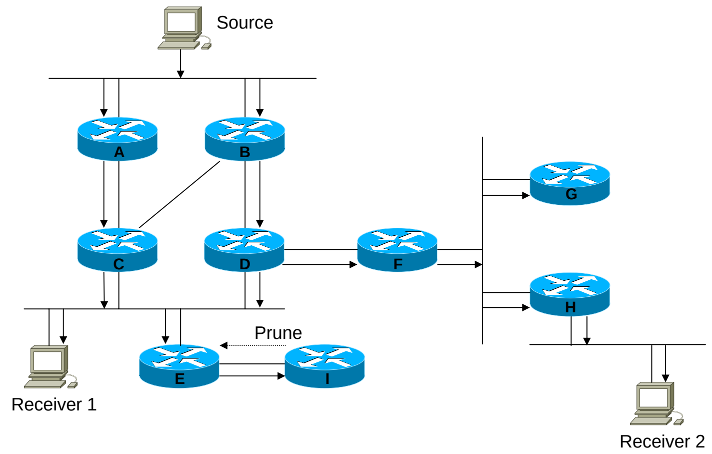
PIM-DM – Pruning II

- Pruning of non-RPF interfaces



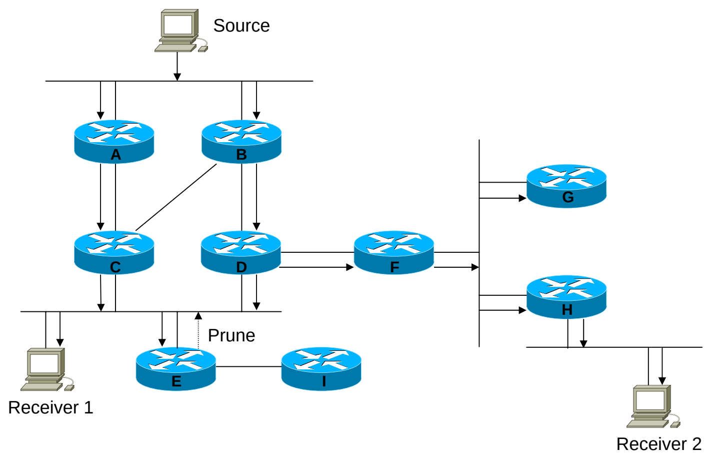
PIM-DM – Pruning III

- Leaf router without receivers, step 1



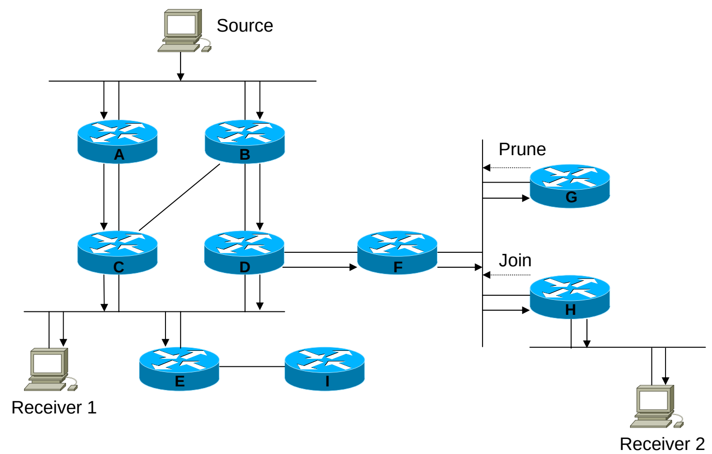
PIM-DM – Pruning IV

- Step 2



PIM-DM – Pruning V

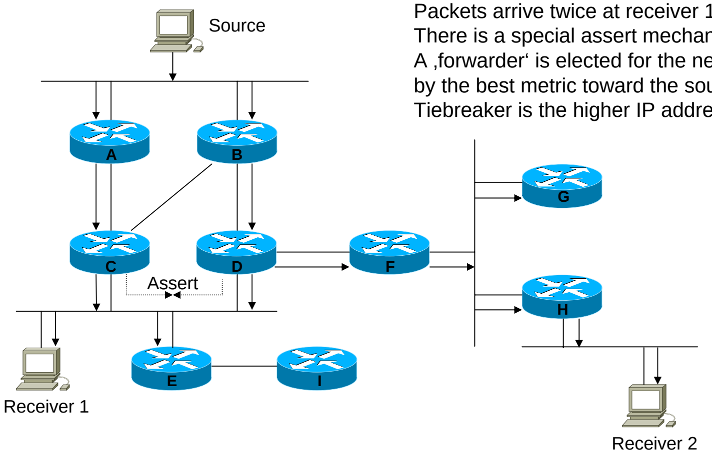
- Prune override



PIM-DM – Pruning VI

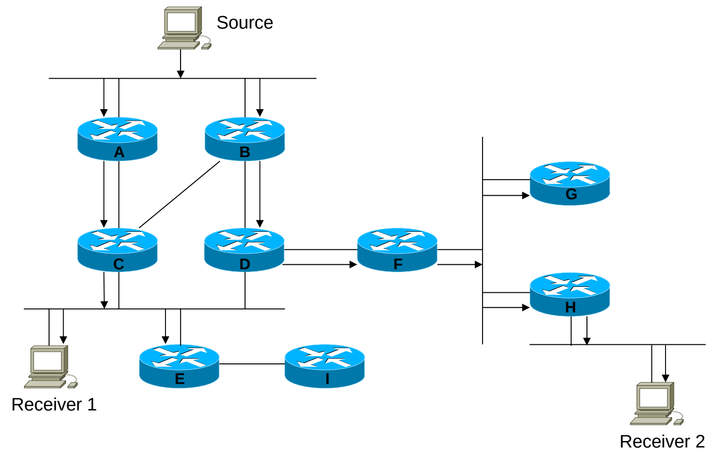
- Assert

Packets arrive twice at receiver 1.
There is a special assert mechanism.
A ‚forwarder' is elected for the network by the best metric toward the source.
Tiebreaker is the higher IP address.

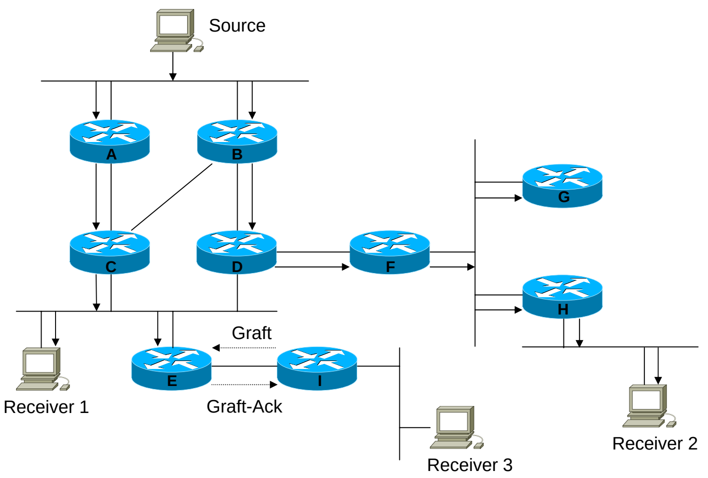


PIM-DM – Pruning VII

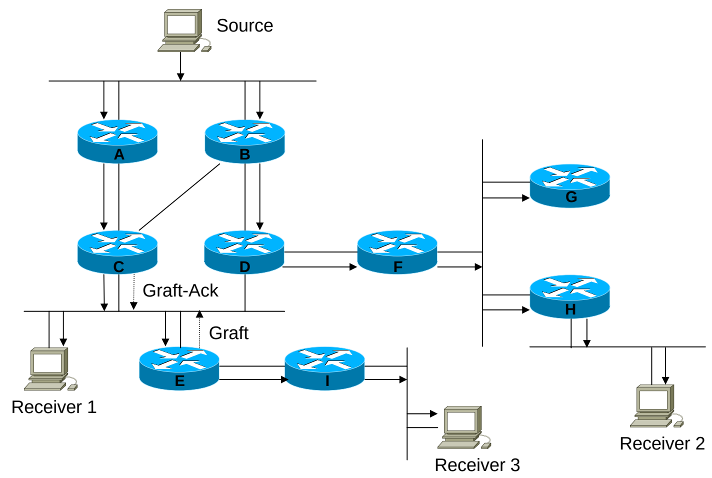
- After assert



PIM-DM – Grafting



PIM-DM – Grafting II



PIM Sparse Mode

- Characteristics
  - Protocol independent (uses unicast routing table for RPF checks)
  - Multicast forwarding via (1) RPT (also known as shared tree) and (2) SPT
  - 'pull' principle (an explicit join is required)
  - Classless (so far as the unicast routing protocol is classless)
- Shared Tree (RP-Tree, RPT)
  - Single tree rooted at the RP leading to all receivers (regardless of the sender)
  - Created using join/prune messages
- Shortest Path Tree (SPT)
  - Shortest path tree rooted at a source leading to all receivers (different trees for different sources)
  - Same mechanisms of join/prune messages for RPT and SPT

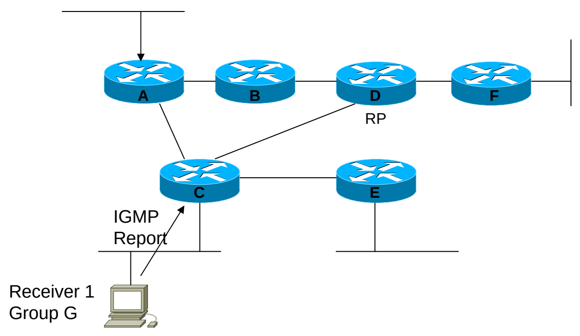
PIM Sparse Mode II

- Advantages of SPTs
  - Direct path between source and destination
  - Minimization of the latency
  - Minimization of the load of the RP
- Disadvantages
  - Number of required (S,G) entries may be very large
  - Requires much more resources within the network
- Question: What is the need of the RPT?
  - The problem is to find active multicast sender!

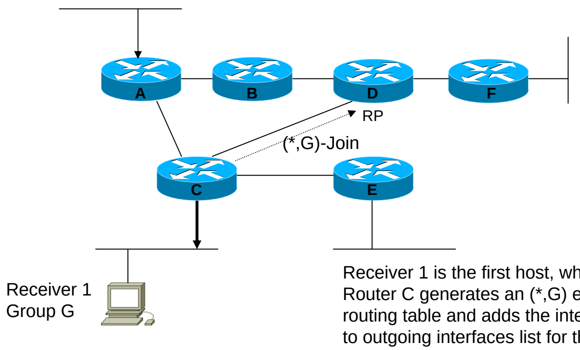
PIM Sparse Mode III

- Join/Prune messages
  - Each message contains a list of joins and a list of prunes
  - Each entry contains:
    - Multicast source address - source address or RP, if WC-bit
    - Multicast group address
    - WC-bit (wildcard flag) - indicate (*,G) join/prune
    - RP-bit (RP tree flag) – this information if for the RP and has to be forwarded toward the RP

PIM-SM – Shared Tree Join

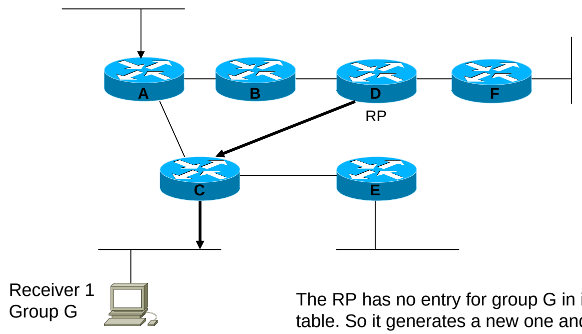


PIM-SM – Shared Tree Join II



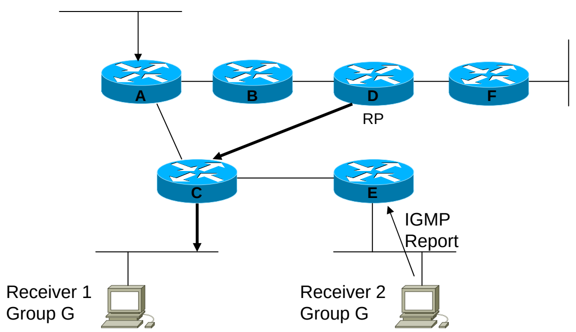
Receiver 1 is the first host, which joins group G. Router C generates an (*,G) entry in its multicast routing table and adds the interface toward receiver 1 to outgoing interfaces list for this (*,G) entry. Also, it sends an PIM (*,G) join toward the RP.

PIM-SM – Shared Tree Join III

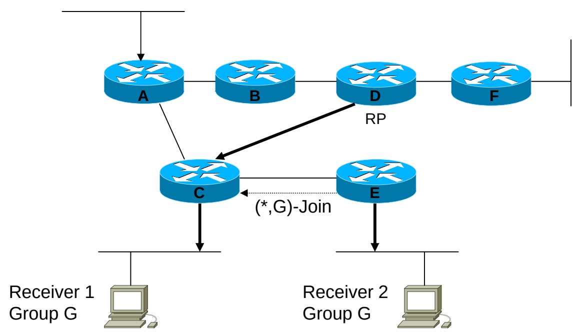


The RP has no entry for group G in its multicast routing table. So it generates a new one and adds the interface toward router C to the outgoing interfaces list of the new (*,G) entry.

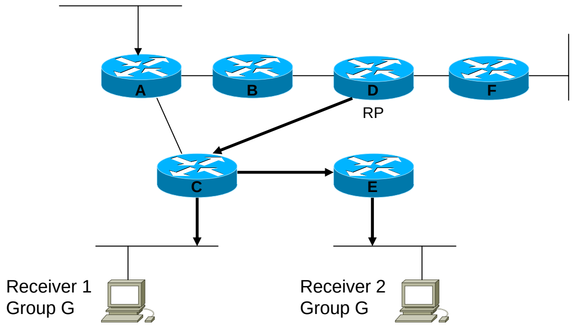
PIM-SM – Shared Tree Join IV



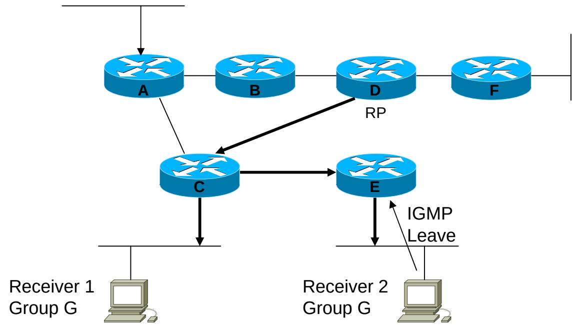
PIM-SM – Shared Tree Join V



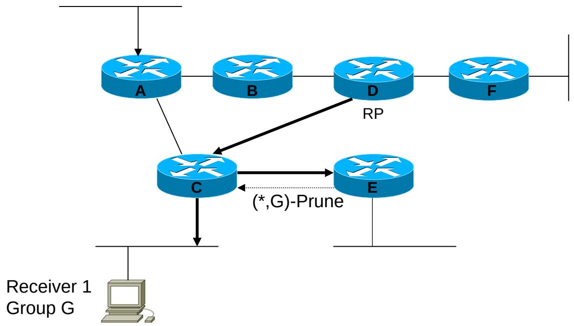
PIM-SM – Shared Tree Join VI



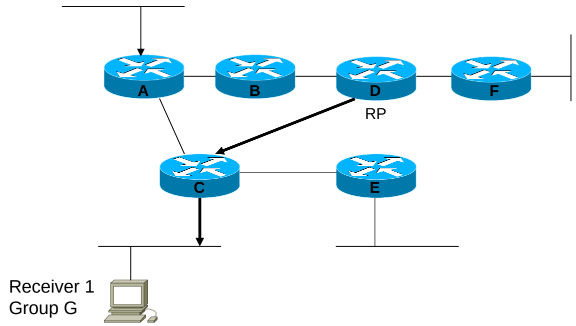
PIM-SM – Shared Tree Prune



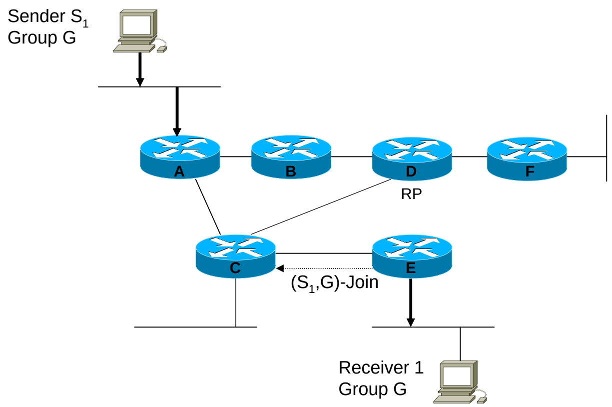
PIM-SM – Shared Tree Prune II



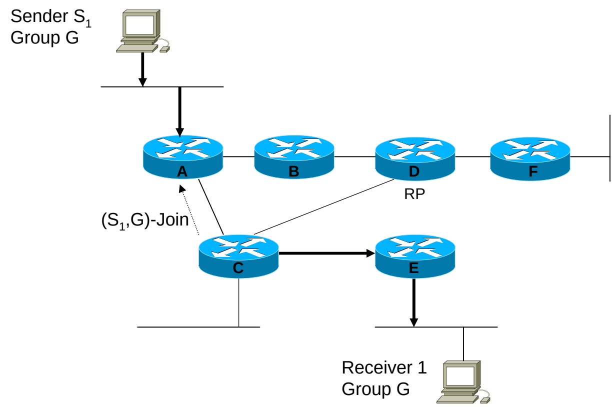
PIM-SM – Shared Tree Prune III



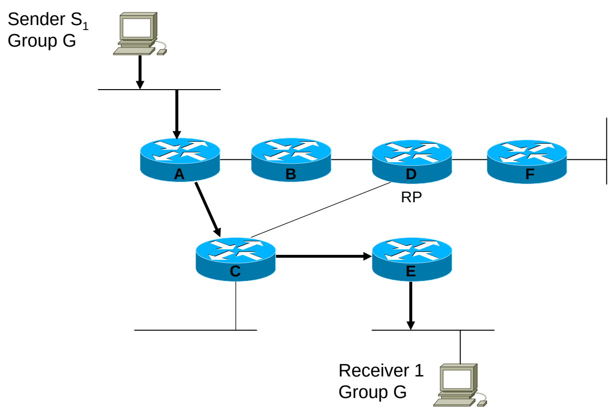
PIM-SM – Shortest Path Tree Join



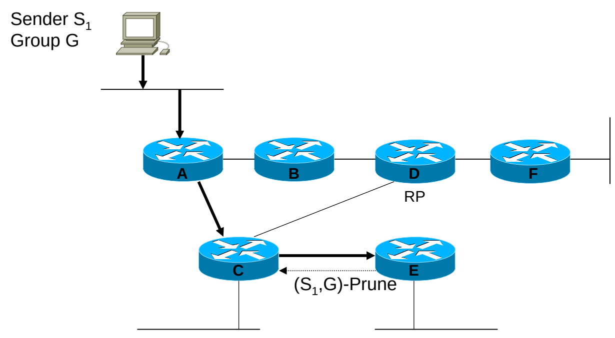
PIM-SM – Shortest Path Tree Join II



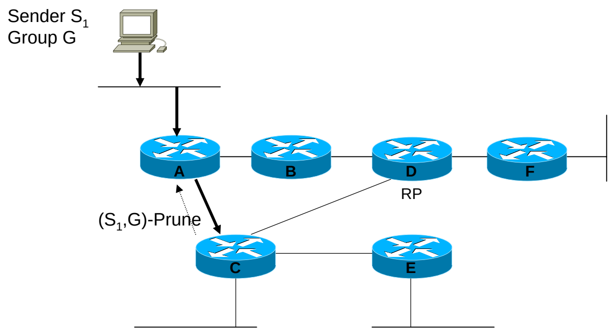
PIM-SM – Shortest Path Tree Join III



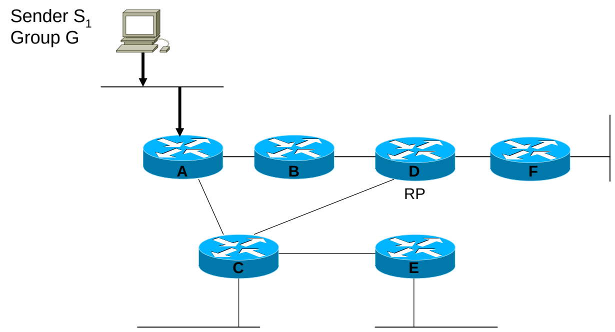
PIM-SM – Shortest Path Tree Prune



PIM-SM – Shortest Path Tree Prune II



PIM-SM – Shortest Path Tree Prune III



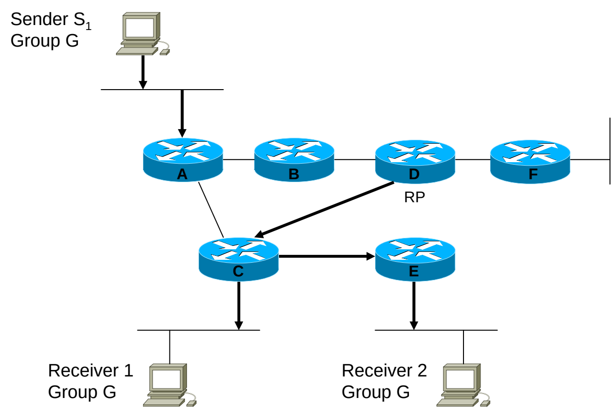
PIM-SM – Source Registration

- PIM-SM uses the RPT for get multicast packets to all receivers
- But, how do the packets get to the RP?
  By registering the active source at the RP!

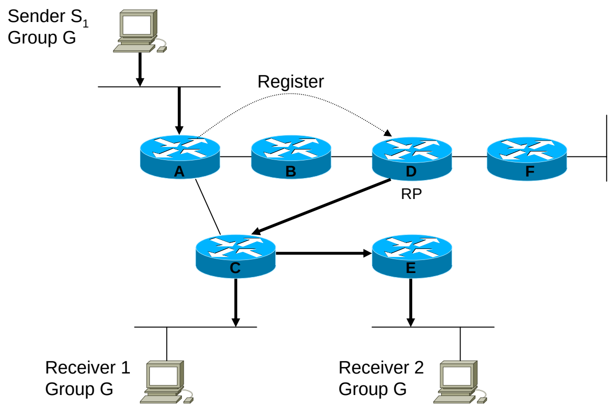
- PIM register messages
  - Tell the RP that source S_i sends packets to group G
  - Send the first multicast packets from source S_i (encapsulated into PIM register messages) to the RP
- PIM register-stop messages are sent, if
  - The RP already receives traffic from S_i via (S_i,G) SPT
  - The RP has no use for this traffic because there is no active RPT

- Please note: register and register-stop messages are unicast between the first hop router and the RP

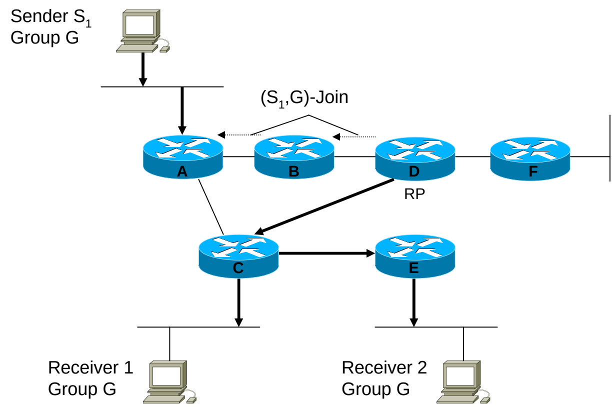
PIM-SM – Source Registration II



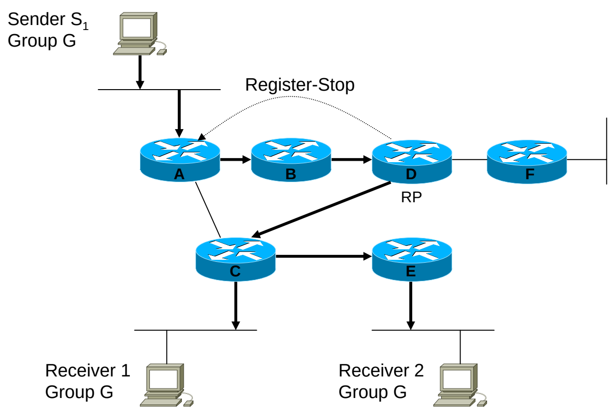
PIM-SM – Source Registration III



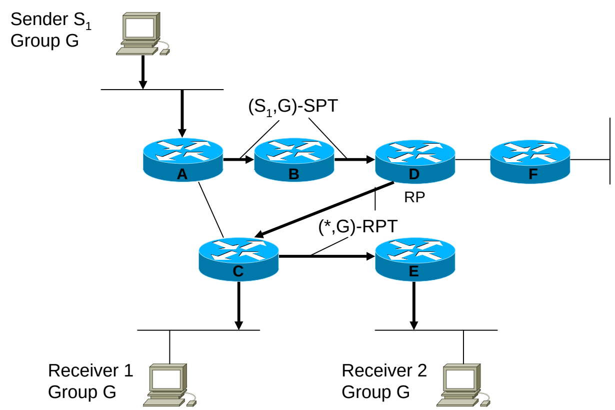
PIM-SM – Source Registration IV



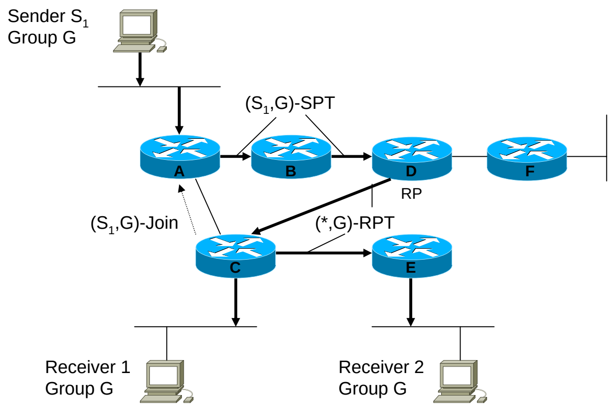
PIM-SM – Source Registration V



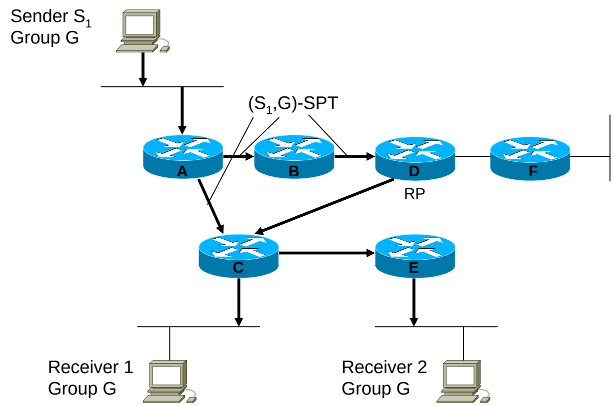
PIM-SM – Source Registration VI



PIM-SM – SPT switch-over

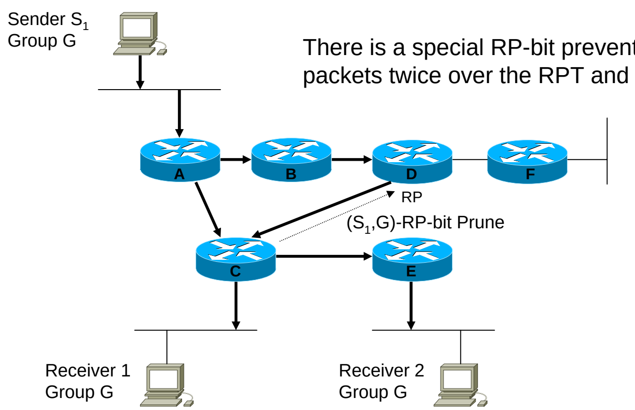


PIM-SM – SPT switch-over II

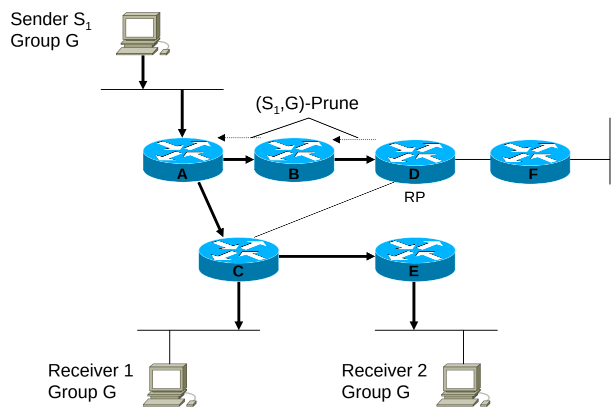


PIM-SM – Pruning Sources from RPT

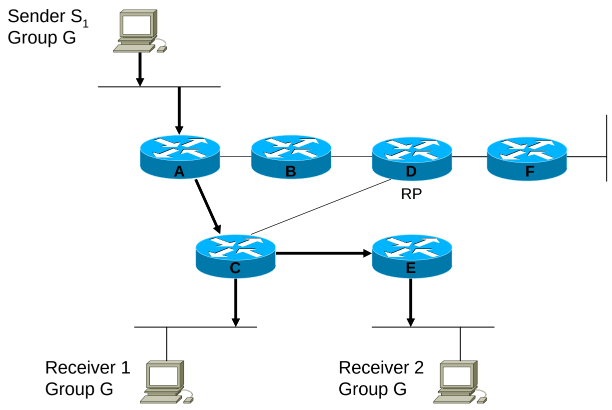
There is a special RP-bit prevent receiving packets twice over the RPT and the SPT.



PIM-SM – Pruning Sources from RPT II



PIM-SM – Pruning Sources from RPT III



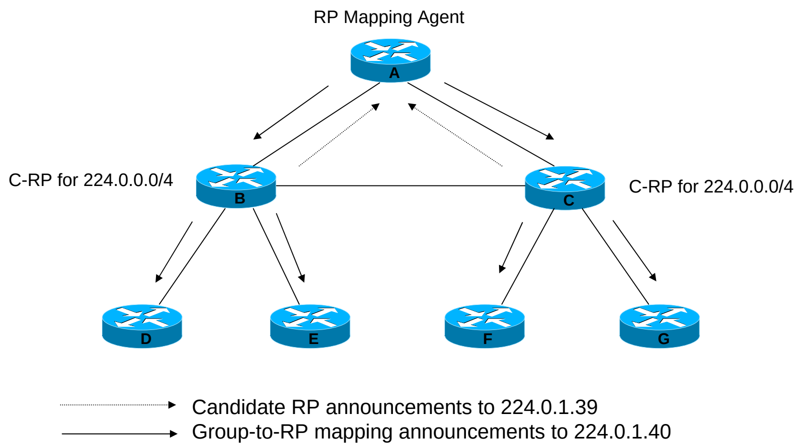
PIM Rendezvous Point

- Where is the RP?
  - Static configuration
    - Easy deployment and debugging, no redundancy
  - Cisco Auto-RP
    - Dynamic, redundancy is possible
  - PIMv2 Bootstrap Router (BSR)
    - Standardized in PIMv2
    - The mechanism most vendors understand

Cisco Auto-RP

- Two mechanisms
  - Mapping agent
    - Information via 224.0.1.40 (Cisco-RP-Discovery)
    - Provide a group-to-RP mapping
  - Candidate RPs
    - information via 224.0.1.39 (Cisco-RP-Announce)
    - potential RPs (maybe for a reduced address space)
- What, if there is no RP (for a particular group)?
  - Fallback to a statically configured RP
  - If there is still no RP, switch to dense mode (for this group)

Cisco Auto-RP II



Source Specific Multicast (SSM)

- PIM-SM
  - Applications ‘join’ to an multicast address
  - If two applications use the same address, both applications get the unwanted traffic
  - Everyone can send data to this group (ideal for a denial-of-service attack)

- SSM
  - The closest router sees the request of a receiver to get data for a multicast group from a specific source (via IGMPv3)
  - Thus, the SPT can be established without the need of an RPT

  - Extension to PIM-SM
  - Allows an efficient data delivery for one-to-many communications such as TV broadcasts
  - Prevents from finding / using a single RP
  - Simplifies the intra-domain routing by removing the requirement for MSDP to announce active sources
  - Solves the IP multicast address collision problem

Multicast Inter-Domain Routing

- DVMRP?
  - Scalability (flooding)
- PIM-SM?
  - Requires knowledge about RPs
  - Scalability / Reliability (RP)

- BGMP (Border Gateway Multicast Protocol)
  - Inter-domain multicast protocol
  - Supports RPTs, SPTs
  - Distribution via BGP-4

- MSDP (Multicast Source Discovery Protocol)
  - To interconnect sparse mode networks
  - Distributes information about active sources
  - Still scalability issues!