

## IP Multicast

---

**Muhammad Jaseemuddin**

**Dept. of Electrical & Computer Engineering  
Ryerson University  
Toronto, Canada**

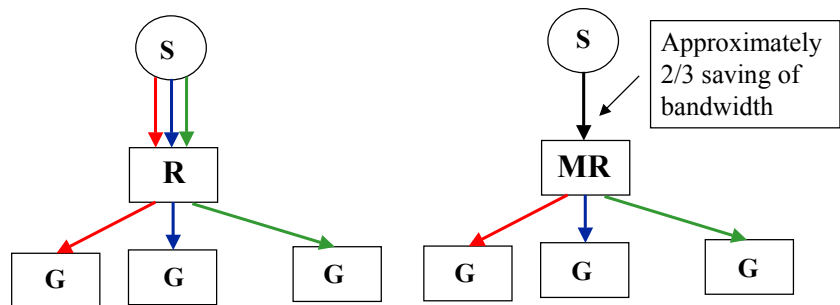
## References

---

- Greg Shepherd, Juniper Networks, *IP Multicast Tutorial*, APRICOT 2002.  
<http://www.shepfarm.com/juniper/multicast/McastApricot2002.ppt>
- *Multicast Routing Algorithms and Protocols: A Tutorial*, IEEE Network, January/February 2000.
- C. Diot et al, *Deployment Issues for the IP Multicast Service and Architecture*, IEEE Network, January/February 2000.
- K. Almeroth, *The Evolution of Multicast: From the Mbone to Interdomain Multicast to Internet2 Deployment*, IEEE Network, January/February 2000.
- B. Edwards, L. Giuliano and B. Wright, *Inter-domain Multicast Routing – Practical Juniper Networks and Cisco Systems Solutions*, Addison-Wesley, 2002.
- B. Williamson, *Developing IP Multicast Networks*, Vol. 1, Cisco Press, 2000.
- S. Deering et al, *The PIM Architecture for Wide-Area Multicast Routing*, IEEE/ACM Transactions on Networking, Vol. 4, No. 2, April 1996.

## Why Multicast?

---



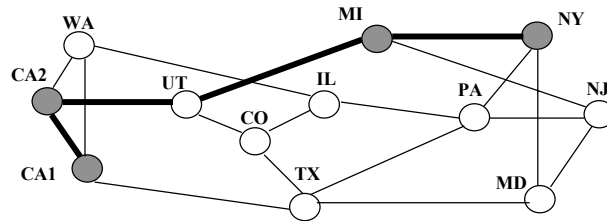
## Multicast – Problem Statement

---

- Let
  - $V$  be the set of nodes
  - $E$  be the set of edges
  - $G = (V, E)$  be the undirected connected graph
  - $M \subseteq V$  be the subset of nodes forming multicast group

Then the problem of multicast is to find a tree **T** in graph **G** such that **T** spans all vertices in the multicast group **M**

## Steiner Tree Problem Formalism



- Let
  - $G = (V, E)$
  - $C_{u,v} = C(u,v)$ , a cost function assigns positive real cost to the link  $(u,v)$
  - $M \subseteq V$Find a tree  $T = (V_T, E_T)$ , which spans  $M$  such that its cost  $C_T = \sum C_{u,v}$  for all  $(u,v) \in V_T$  is minimized.  
 $M = \{CA1, CA2, MI, NY\}$   
UT is a Steiner node

## Steiner Tree Problem in Network

- **SPN is a NP complete problem, but polynomial time algorithm exists for the following trivial cases**
  - $|M| = 2$  unicast case  $\rightarrow$  reduced to shortest path problem
  - $|M| = |V|$  broadcast case  $\rightarrow$  reduced to minimum spanning tree problem
  - $G$  is a tree  $\rightarrow$  There is only one subtree that is the SPN

## Optimization

---

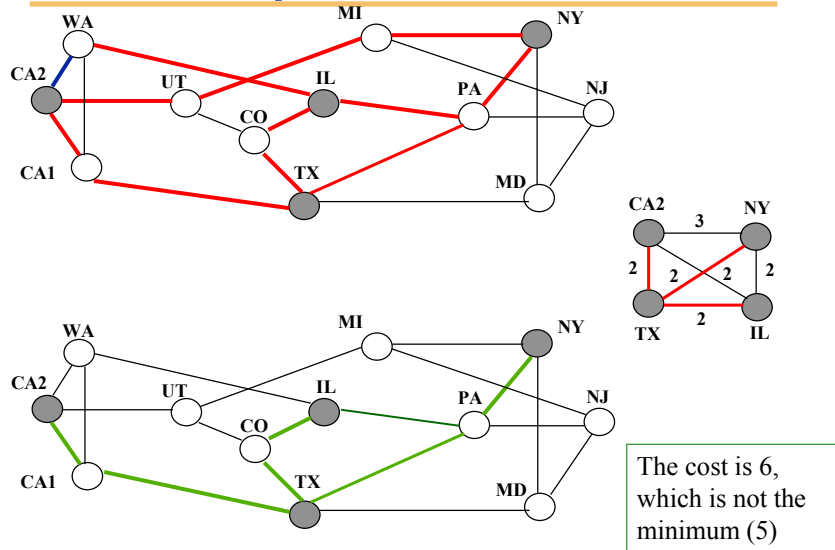
- **Cost optimization**
- **Delay optimization**
- **Scalability**
- **Dynamic Multicast Groups**
- **Survivability**
- **Fairness**
- **Cost-Delay trade-off**

## KMB Algorithm

---

- **Find group shared minimum cost Steiner tree**  
→ **Cost optimization**
- **An approximation algorithm that assumes symmetric links**
- **Algorithm**
  - Construct an undirected closure graph  $G_1$ , such that for every  $(u,v)$  pair in  $M$  there is a link in  $G_1$  and  $C'_{uv}$  is  $d_{uv}$  in  $G$
  - Find the minimum spanning tree  $T_1$  of  $G_1$
  - Construct  $G_2$  by replacing each link in the spanning tree  $T_1$  of  $G_1$  with the corresponding shortest path in  $G$ 
    - May introduce steiner nodes
  - Find the minimum spanning tree  $T_2$  of  $G_2$
  - Construct multicast tree  $TM$  by deleting links in  $T_2$ , if necessary, such that all the leaves in  $TM$  belong to the multicast group

## KMB Example



## Difficulties with ST Formulation

- Computational Complexity
  - General solution is NP complete
  - Approximate algorithms exist that give sub-optimal solutions
  - How much worse?
    - Cost-performance trade-off?
- The algorithms require knowledge of the topology
  - Multicast group membership is dynamic
    - Group members join and leave the tree anytime
    - Controlled group membership management to be able to propagate the topological change throughout the network will be difficult

## IP Multicast

---

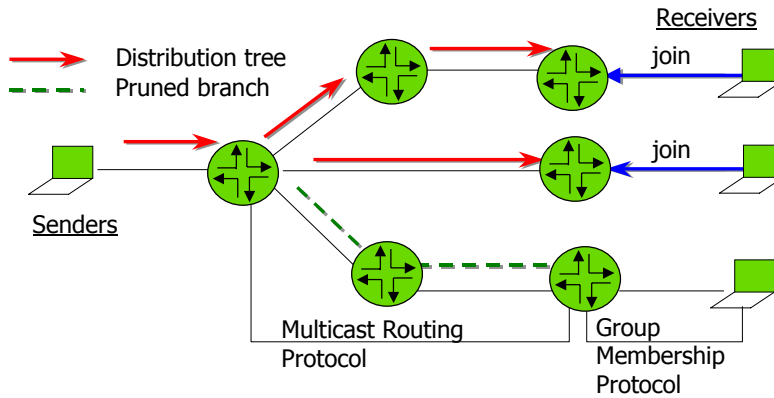
- IP Multicast Model
  - Receivers join the distribution tree → receiver-built tree
  - Senders have no knowledge of receivers
    - Makes the model scalable
  - Senders need not be the group member
  - Any Source Multicast (ASM) model
- The SENDERS send
  - Multicast Addressing - rfc1700
  - class D (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 subnets
  - Multicast Routing Protocols (PIM-SM/SSM)
  - RPF (reverse path forwarding) – against source address

## Reverse Path Forwarding

---

- What is RPF?
  - A router forwards a multicast datagram only if received on the interface that leads a unicast datagram 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 uses unicast route back to source for RPF.
  - RPF ensures no loop in multicast forwarding.
  - Loop is especially bad in case of multicasting
    - Paths fan-out in each cycle
- RPF Table
  - Two ways to populate the RPF table
    - A companion protocol of the multicast protocol updates the RPF table → DVMRP approach
    - Any unicast protocol updates the RPF table → PIM approach

## IP Multicast Components



- Group Membership Protocol - enables hosts to dynamically join/leave multicast groups. Membership info is communicated to nearest router
- Multicast Routing Protocol - enables routers to build a distribution tree between the sender(s) and receivers of a multicast group

## IP 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 of 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

## IP Multicast Addressing

---

- draft-albanna-iana-ipv4-mcast-guidelines-01
- <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
  - 224.0.0.1 - All local hosts
  - 224.0.0.2 - All local routers
  - 224.0.0.4 - DVMRP
  - 224.0.0.5 - All OSPF routers
  - 224.0.0.6 - All OSPF Designated Routers
  - 224.0.0.9 - RIP2
  - 224.0.0.13 - PIM
  - 224.0.0.15 - CBT
  - 224.0.0.18 - VRRP

## IP Multicast Addressing

---

- Administratively Scoped Addresses – rfc2365
  - 239.0.0.0–239.255.255.255
  - Private address space
    - Similar to RFC1918 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

## Private Multicast Addresses

- GLOP addresses
  - Provides globally available private Class D space
  - 233.x.x/24 per AS number
  - RFC2770
  - AS number = 16 bits
    - Insert the 16 bits of ASN into the middle two octets of 233/8

Online Glop Calculator:

[www.shepfarm.com/juniper/multicast/glop.html](http://www.shepfarm.com/juniper/multicast/glop.html)

## IP – MAC Multicast Address Mapping

- **RFC 1700 - Ethernet**

	224.	10.	8.	5	← Class D IP address
0000 0001	0000 0000	0101 1110	0xxx xxxx	xxxx xxxx	xxxx xxxx ← MAC Address
----- block 01-00-5E -----					
	IANA reserved			0 = Internet Multicast	
0000 0001	0000 0000	0101 1110	0000 1010	0000 1000	0000 0101
----- 01-00-5E-0A-08-05-----					

- 224.10.8.5 multicast stream maps to 01-00-5E-0A-08-05 Ethernet MAC address.
- 32 multicast addresses map to a single Ethernet multicast address
  - RFC 1469 TR
  - RFC 1390 FDDI
  - RFC 2226 & 2022 - ATM
  - RFC 1209 SMDS (broadcast)

## IGMP Details - Router

---

- **Router:**
  - sends Membership Query messages to All Hosts (224.0.0.1)
    - query-interval = 125 secs default
  - router with lowest IP address is Querier (rest non-queriers)
  - If lower-IP address query heard, backoff to non-querier state
    - Other Querier Present Interval default:  $(\text{robust-count} \times \text{query-interval}) + (0.5 \times \text{query-response-interval}) = 255$  secs
  - listens for reports (whether querier or not) and adds group to membership list for that interface
    - query-response-interval = 10 secs default
  - Soft state
    - State timeout (Group member interval) default:
      - $(\text{robust-count} \times \text{query-interval}) + (1 \times \text{query-response-interval}) = 260$  sec
  - robust-count - provides fine-tuning to allow for expected packet loss on a subnet. Default = 2 (tunable from 2-10)

## IGMP Details - Host

---

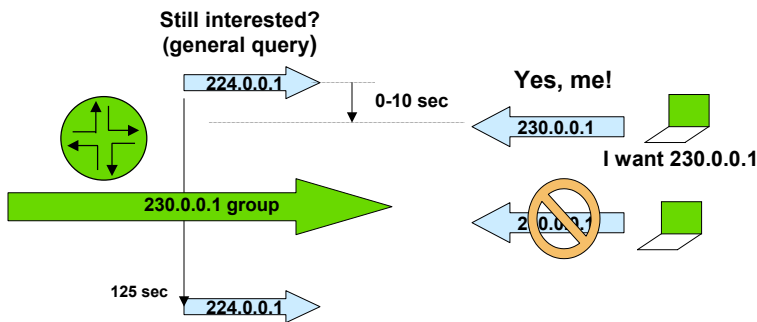
- **Host:**
  - sends Membership Report messages, if joined
    - Waits for a random time between 0-10 sec (def).
    - Hosts listen to other host reports
      - **Feedback suppression**
      - Only 1 host responds
  - Join messages (unsolicited Membership Report) to group address (e.g. 224.10.8.5)
  - Leave messages to All Routers (224.0.0.2)

## IGMP Report (Join)



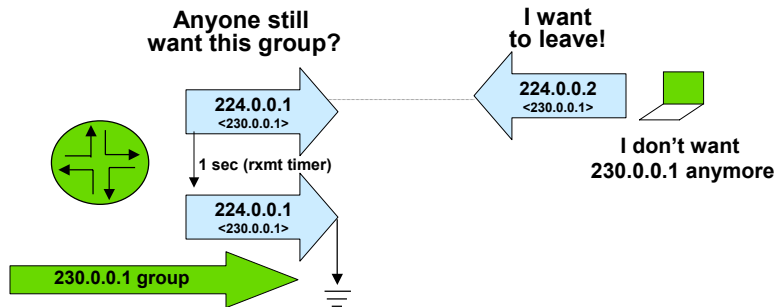
- Hosts can send unsolicited *join* membership messages – IGMP Report
  - Can send more than one
- Or hosts can join by responding to periodic query from router

## IGMP Query



- 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 soft state timer=260 sec expires with no response from any host, router times out group state

## IGMP Leave



- Hosts that support IGMP v2 send *leave* messages to all routers group indicating group they're leaving.
  - Router follows up with 2 *group-specific queries* messages
- IGMP v1 hosts leave by not responding to *queries* (260 sec timeout)

## IGMP Enhancements

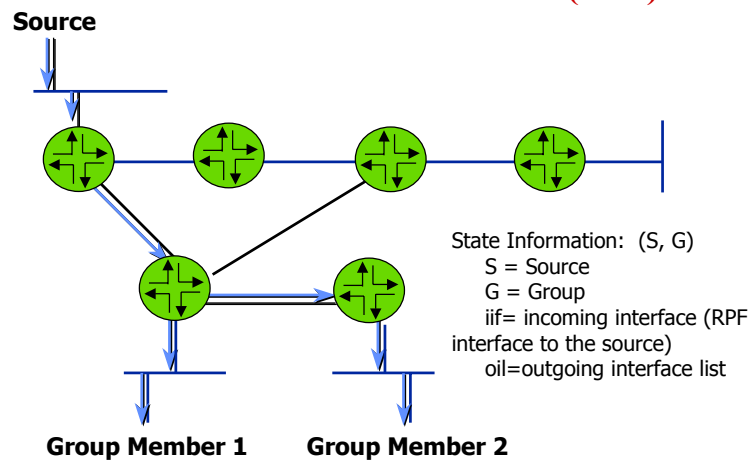
- **IGMP Version 2**
  - multicast router with lowest IP address is elected querier
    - IGMPv1 was multicast protocol specific and potentially conflicted.
  - Group-Specific Query message is defined. Enables router to transmit query to specific multicast address rather than to the "all-hosts" address 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.

## IP Multicast Routing Protocol

- Multicast Routing Protocol needs to build and maintain multicast distribution tree
- Two fundamental approaches of building a multicast tree
  - Source-based tree
    - Shortest path from source to every destination
  - Shared tree
    - Shortest path from core to every destination
- Broadcast-and-prune or dense mode
  - Routers keep huge state information
  - E.g. DVMRP, PIM-DM
- Core-based or sparse mode
  - A shared tree is formed
  - Router keeps less state
  - E.g. CBT, PIM-SM

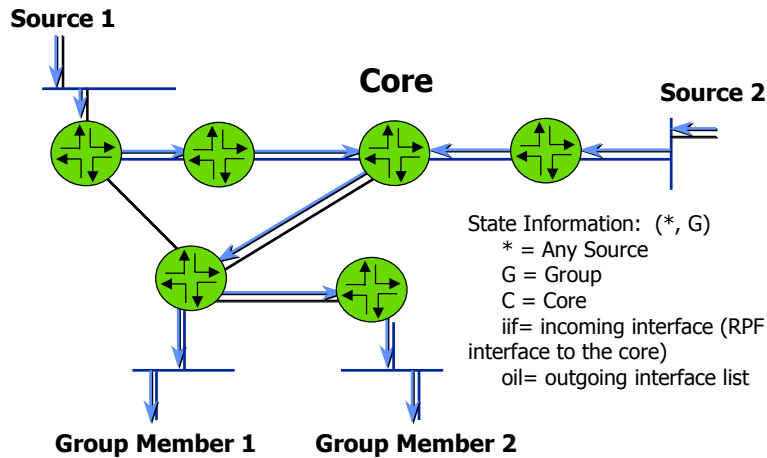
## Multicast Distribution Tree – SPT

- **Shortest Path or Source-based Tree (SPT)**



## Multicast Distribution Tree – RPT

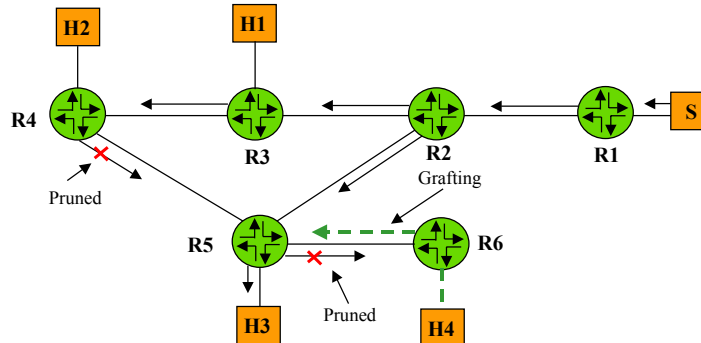
### • Shared or Core-based or Rendezvous Point Distribution Tree (RPT)



## Multicast Distribution Tree

- **Source-Specific or Shortest Path Trees (SPT)**
  - More resource intensive; requires more state  $\rightarrow o(S \times G)$
  - It gives optimal paths from source to all receivers  $\rightarrow$  minimizes delay
  - Traffic is distributed
  - Best for one-to-many distribution, data intensive application (e.g. video conferencing)
- **Shared or Core-Based or Rendezvous Point Trees (RPT)**
  - Uses less resources; less state one per group  $\rightarrow o(G)$
  - It may give sub optimal paths from source to some or all receivers, depending on topology
  - The choice of Rendezvous Point RP (core) *may* affect performance
  - Traffic is concentrated along fewer paths
  - Best for many-to-many distribution, less data intensive application (e.g. resource discovery)
  - May be necessary for source discovery (PIM-SM)

## IP Multicast – Dense Mode



- **Broadcast phase**
  - router broadcasts multicast packet on all interfaces
  - router sends prune to the upstream router if the interface is not RPF interface to the source
    - router receives multicast packet only through RPF interface
- **Prune Process**
  - router sends prune to the upstream router no receiver has joined to any of its interface
- **Grafting Process**
  - router sends graft message to the upstream router to graft an already pruned link when a receiver joins any of its interface

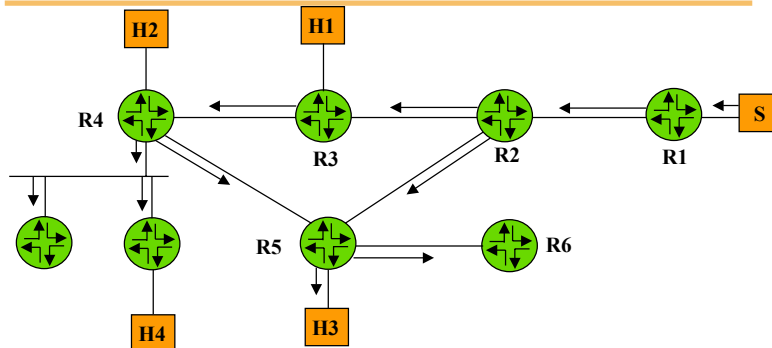
## PIM

- Protocol Independent Multicast (PIM)
  - It uses unicast routing table built by any unicast routing protocol for RPF
    - Does not rely on any specific unicast routing protocol
  - Dense Mode → **PIM-DM**
    - Unlike DVMRP it uses underlying unicast routing protocol
  - Sparse Mode → **PIM-SM**
    - Like CBT it builds a core-based shared tree rooted at Rendezvous Point, called RP-rooted Shared Tree
    - Switch to Shortest Path Tree (SPT)
  - **PIM Neighbor Discovery**
    - PIM routers send Hello message periodically (every 30 seconds) to 224.0.0.13 (all PIM routers)
    - Hold time defined in hello message (90 seconds) specifies the time if elapsed without receiving hello message from a neighbor indicates that the neighbor is down
    - DR election
      - PIM routers on a broadcast link elects DR through Hello message
      - The router with the highest IP address is elected as the DR

## PIM-DM

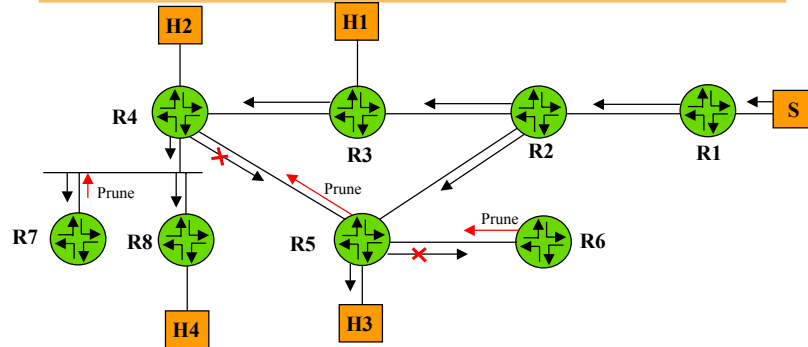
- Unlike DVMRP it uses underlying unicast routing protocol
- A PIM-DM router broadcasts packets to all outgoing interfaces **disregarding the fact that some routers connected to those interfaces never use this router to send packets to the source**, the packets fail RPF check at those routers causing prune message to be sent by them
  - It simplifies broadcast
  - But causes unnecessary generation and handling of prune messages

## PIM-DM: Flood-and-Prune



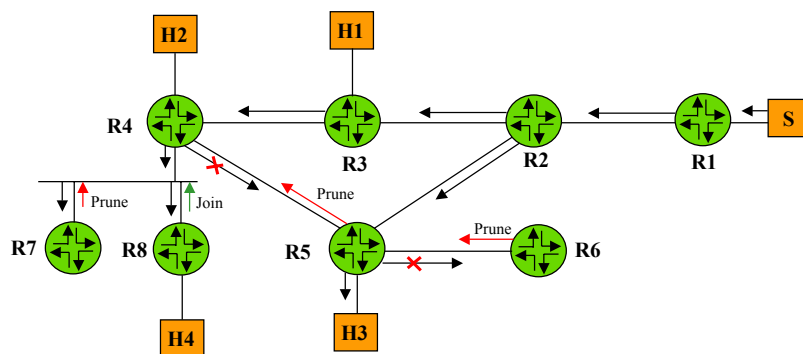
- A router when receives a multicast packet through the RPF interface for the source S, then it installs (S,G) state and forwards the packet downstream through all other interfaces and include those interfaces in the (S,G) oil
  - The (S,G) state times out after 3 minutes
  - Every time the router receives a packet, it refreshes the (S,G) state and resets the expiration timer to zero
  - the router keeps every interface in the outgoing interface list either in forwarding or prune state; prune state changes into forwarding state after 3 minutes
- When the source stops sending packets, (S,G) states in all the routers time out

## PIM-DM: Flood-and-Prune



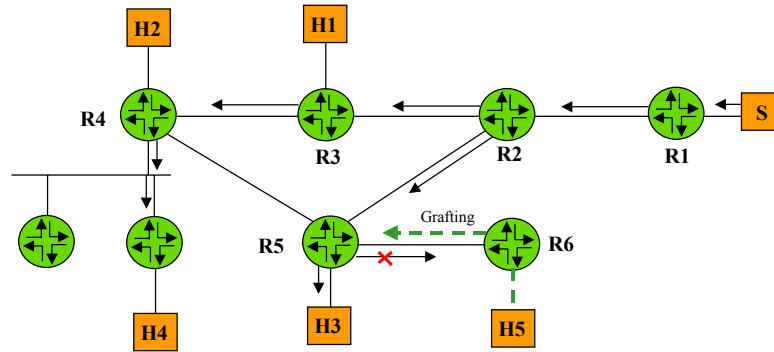
- A router can prune a branch by sending prune message to upstream neighbor if:
  - it has no group member downstream, e.g. R6 and R7
  - it receives the packet on non-RPF interface, e.g. R5
- The router that prunes the tree maintains multicast state including the link identifies the upstream router
- Router disables forwarding on the outgoing interface through which it receives prune
  - prune state is a soft state and typically expires after 3 min, then the router resumes forwarding through that interface
- The router can send prune again every time it receives the packet, I.e. after every 3 min
- When source S stops transmission, the multicast states at routers time out

## PIM-DM: Prune Override



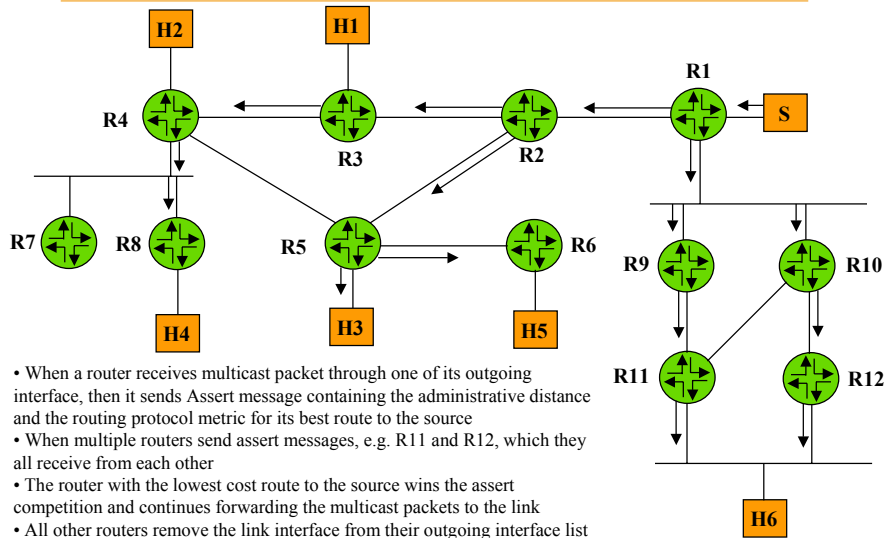
- A router does not accept prune from another router on a multi-access link for about 3 seconds
  - e.g. R4 does not accept prune generated by R7
- Another router within this time can override the prune by sending Join immediately after seeing a prune on the line
  - e.g. R8 sends Join immediately after it sees R7's prune, which is accepted by R4

## PIM-DM: Grafting



- A router (e.g. R6) that has previously pruned the multicast tree for group G, when receives a join its downstream for the same group, it sends Graft message to the upstream router to resume multicast forwarding on the pruned link immediately
  - it can alternatively wait for the prune state at the upstream router (R5) to timeout, but then would lose the traffic for up to 3 minutes
- Graft can trigger grafting process upstream up to the point of active multicast forwarding router

## PIM-DM: Assert



- When a router receives multicast packet through one of its outgoing interface, then it sends Assert message containing the administrative distance and the routing protocol metric for its best route to the source
- When multiple routers send assert messages, e.g. R11 and R12, which they all receive from each other
- The router with the lowest cost route to the source wins the assert competition and continues forwarding the multicast packets to the link
- All other routers remove the link interface from their outgoing interface list (oil) and prune the tree if needed, e.g. R11 and R9

## PIM-SM

---

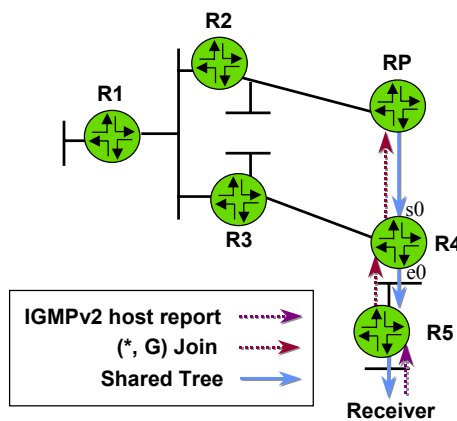
- Each group has its own unique *Rendezvous Point* (RP)
- Some routers are configured as RP
- RP bootstrapping
  - Routers are manually configured to learn the addresses of RPs
  - A bootstrap protocol is designed that automate RP discovery – called PIM Bootstrap Router (**BSR**)
    - PIM-SMv1 does not have the bootstrap protocol, hence every vendor has its proprietary RP discovery mechanism, e.g. Cisco supports proprietary **Auto-RP** protocol
    - PIM-SMv2 contains BSR
- PIM-SM supports both RPT and SPT
  - A single shared tree rooted at RP is formed
  - The tree is a Reverse Shortest Path Tree (RSPT)
- Receiver explicitly joins the tree. It sends:
  - RP-Join towards the RP to join RPT
  - SP-Join towards the Source to join SPT
- Each router along the path creates soft state that expires in 3 minutes
  - Periodically join message is sent upstream to refresh the state typically once every minute

## PIM-SM

---

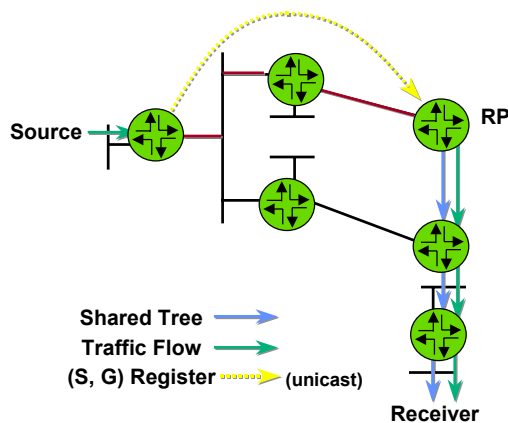
- Source sends multicast data packets encapsulated in the *registration packet* unicast to the RP
  - The RP when receives the registration packet from the source:
    - It strips off the unicast encapsulation and send multicast packet down the shared tree if there are some receivers who have joined the tree
    - If it doesn't have forwarding state, that is no receiver has joined the tree yet, it sends register-stop message to the source causing the source to stop wasting bandwidth
    - It may send a join message to the source, which causes all routers along the path to the source establish multicast state. Thereafter the RP will receive packets from the source as multicast packets avoiding encapsulation overhead

## PIM-SM Shared Tree Join



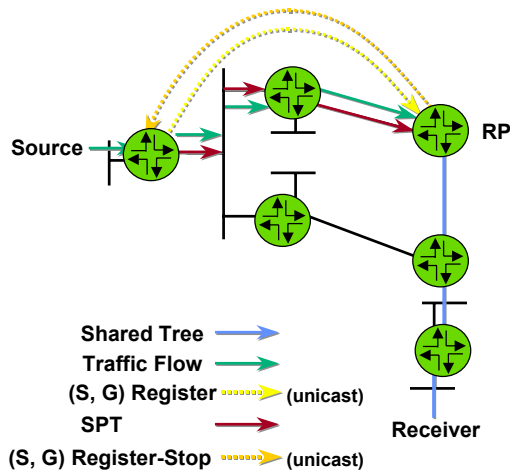
- Host joins group G by sending IGMP Report message (\*, G) to the last-hop router
- Last hop router sends (\*, G) join towards RP
- Routers along the way to the RP create (\*, G) State with RPF interface pointing to RP
- (\*, G) state:
  - Group
  - RP address
  - input interface (iif)
  - output interface (oil)
  - WC-bit: any source
  - RPT-bit: shared tree
- For example, R4 maintains state
  - (\*, G):
    - Group = G
    - RP address = RP
    - iif = s0
    - oil = {e0}
- An expiration timer is attached to the outgoing interface e0, if it expires (typically in 3 min) the interface is removed from oil
- Downstream routers need to refresh Join by periodically sending typically once every min join towards RP
  - e.g. R5 periodically sends join to R4, thus R4 never removes e0 from its state

## PIM-SM Sender Registration



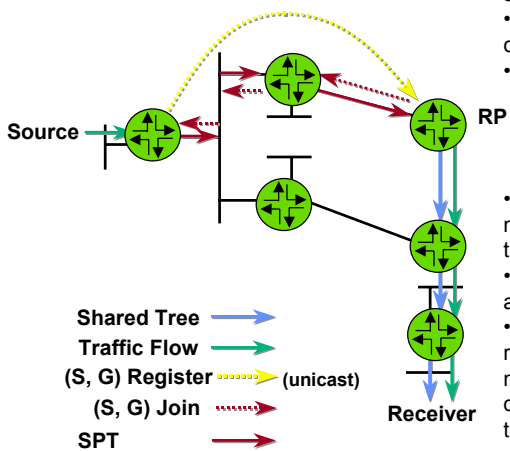
- The first hop router to the source encapsulates the data packet into the register packet and unicasts it to the RP
- The RP decapsulates the Register packet and sends the multicast packet down the RP-shared tree

## PIM-SIM Sender Deregistration (1)



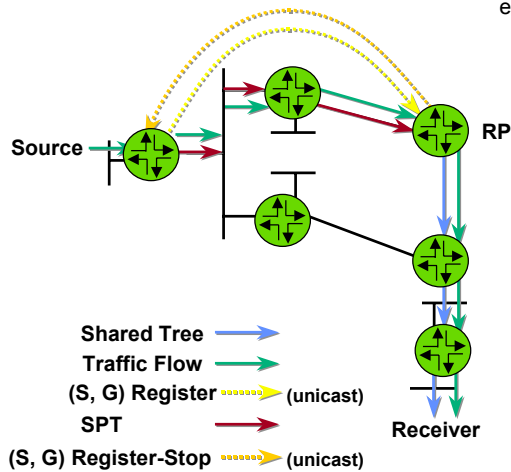
- The RP has no group member downstream
  - e.g. soft state expired
  - It sends Register-Stop message to the first hop router to the source.
- The first hop router starts Registration suppression timer, which if expires causes the router to go back to send the encapsulated packet down the RP-tree

## RP-initiated SP-Tree



- RP sends (S,G) join towards the source
- Routers on the path to the source create (S, G) state along the SPT
- (S,G) state:
  - S: source
  - G: group
  - SPT-bit: SPT is active
  - RPT-bit:
- Only RP and routers with local members can initiate switching to the SP-tree
- all the internal router between RP and source keeps (S,G) state
- Group members connected to any router along the SPT receive multicast packets through the SPT during the packets forwarded through the SPT to the RP

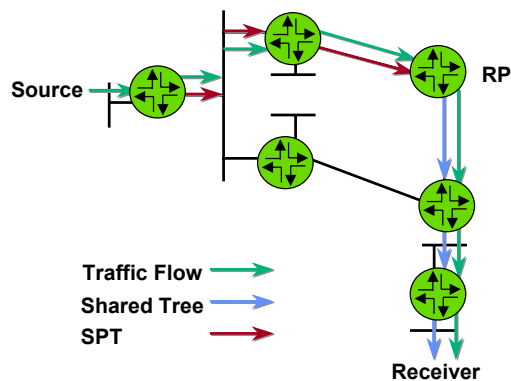
## PIM-SIM Sender Deregistration (2)



The source transmission rate exceeds certain threshold

- The RP sends source specific Join (S,G) to the source to build the SP-tree
- Once RP starts receiving native multicast packets through SPT, if it receives any subsequent Register message then it sends Register-Stop message to the first hop router to stop receiving multicast packets encapsulated in the register message
- The first hop router stops encapsulating multicast packets
- (S,G) traffic continues arriving at the RP through the SP-Tree

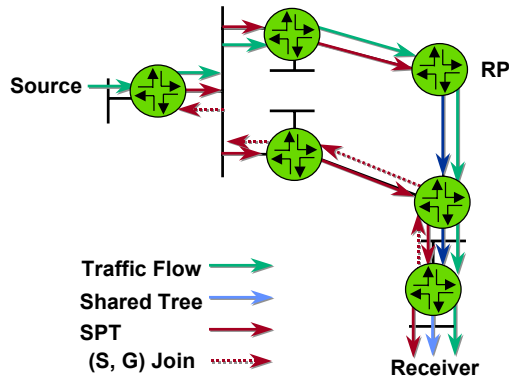
## Source to RP Traffic through SPT



• The multicast traffic originated at the source flows along SP-Tree to RP.

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

## Receiver-initiated SP-Tree



- Last-hop router to the receiver determines that the source rate exceeds SPT threshold

- sends source specific Join (S,G) to the source to join SPT

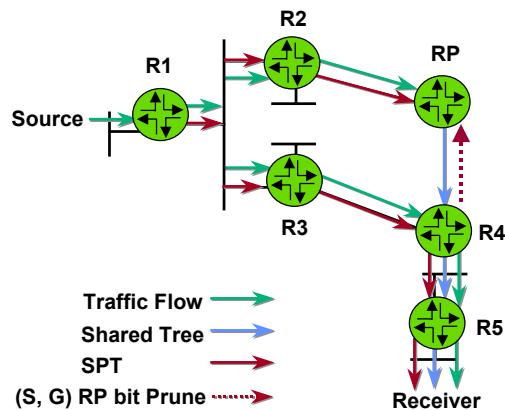
- in Cisco routers default value of SPT-threshold is zero kb/s, i.e. once the last hop router receiver receives first packet along RPT, it sends (S,G) join to the source

- once every second the router calculate the rate of traffic received along RPT

- if the rate exceeds the threshold, then it sends (S,G) join to the source of the first packet received

- SP-Tree from the source to the receiver is created by inducting (S, G) State along the path

## Pruning the Shared Tree for (S,G)



- The router at the cross-road of the SP and RP Trees (e.g. R4) sends (S,G) RP-bit Prune message to the RP

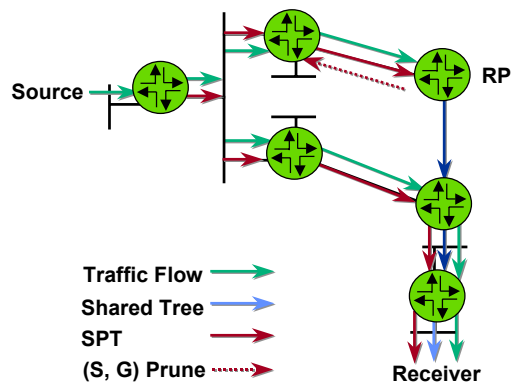
- The iif for (S,G) is different from the iif for (\*,G) state

- Once the SP-Tree is active, that is receive the packet causing SPT-bit to set indicating active SP-Tree

- The prune message contains S in the prune list and RPT-bit set indicating that it no longer needs packets originated at S over the RP-tree

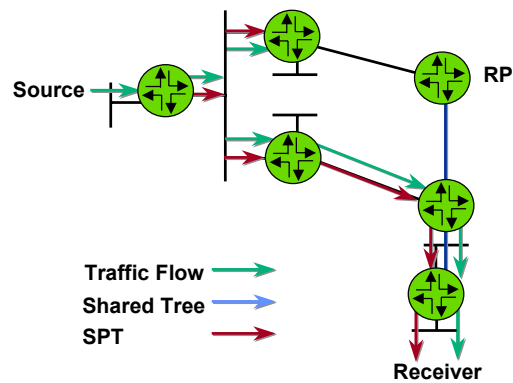
- RP creates (S,G) entry, a child entry of the parent (\*,G) entry, if it doesn't exist, and copies all the interfaces of (\*,G) into (S,G) except the interface through that it receives the join

## PIM-SM SPT Cutover



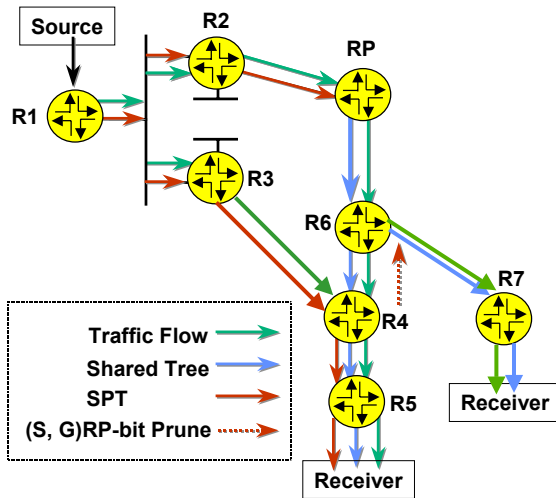
- RP no longer has any member connected to the RPT for (S,G)
- It Prunes the flow of (S, G) traffic by sending (S,G) prune to the source

## PIM-SM SPT Cutover



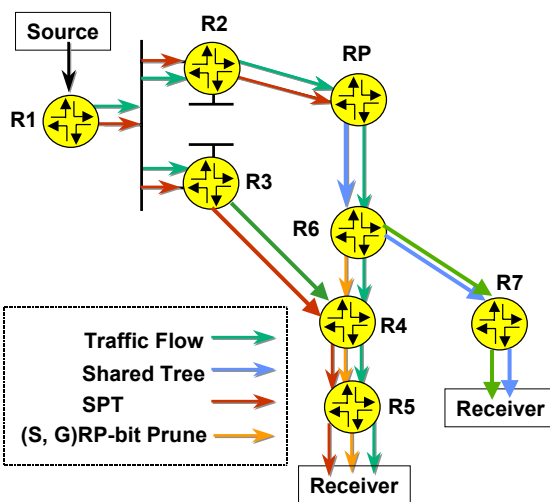
(S, G) Traffic flow is now only flowing to the Receiver via a single branch of the Source Tree.

## PIM-SM SPT Cutover



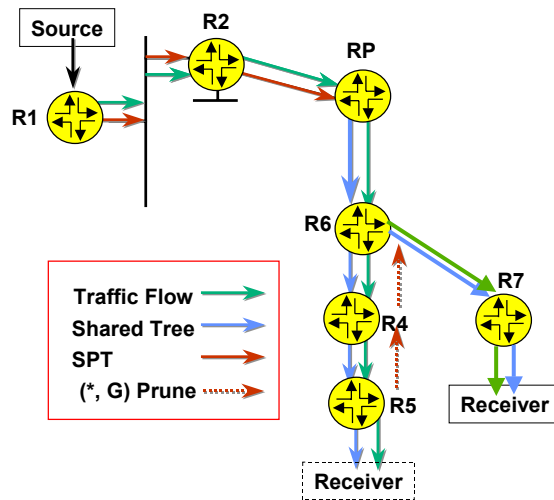
- Traffic begins flowing down the SPT
- The crossover router (e.g. R4) sends (S,G) RP-bit prune towards the RP
- Additional (S, G) State is created along the Shared Tree to prune (S,G) traffic.
- A router along the RPT
  - If it already has (S,G) state it deletes the interface from (S,G) oil through which it received the prune message
  - If it only has (\*,G) state (e.g. R6), it creates an entry for (S,G) state with RPT-bit set, copies all the interfaces from (\*,G) oil to (S,G) oil except the one through which it receives the prune message; the RPF interface for this (S,G) state directs towards the RP because it receives source traffic along the shared tree

## PIM-SM SPT Cutover



- R6 receives (S, G) traffic through the RPT, it forwards the packets to R7 but not to R4
- R4 receives (S,G) traffic only through the SPT and forwards the packets to R5

## Pruning Shared Tree



- When a leaf router, e.g. R5, finds no receiver for a group G attached to its directly connected network

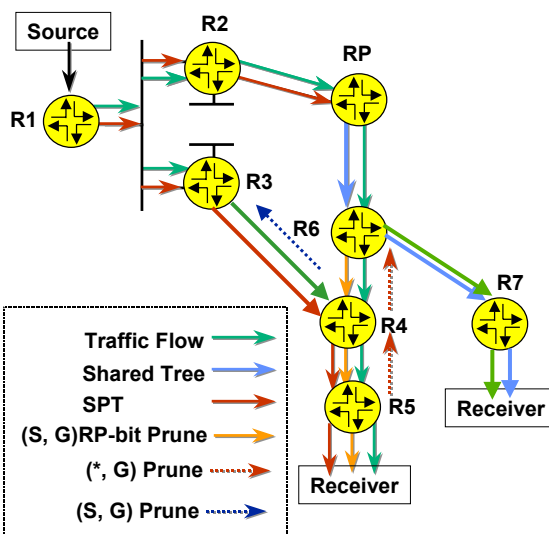
- It sends (\*,G) prune message upstream to the RP to prune the RPT

- A router along the RPT

- If its (\*,G) oil contains only one interface through that it received the (\*,G) prune, then it deletes the interface from the list and forwards the prune message upstream to the RP (e.g. R4)

- If its (\*,G) state contains multiple interfaces, then it simply removes the interface from the oil through that it received the prune message; it does not forward the prune further (e.g. R6)

## Pruning Shortest Path Tree



- When a leaf router, e.g. R5, finds no receiver for a group G attached to its directly connected network

- It sends (\*,G) prune message upstream along the SPT to the RP to prune the SPT

- it stops refreshing (S,G) state

- A router along the SPT removes the interface through that it received the prune message from both (\*,G) and the child (S,G) states, e.g. R5

- it lets (S,G) state expires, after that it counts down the expiration of (\*,G) state

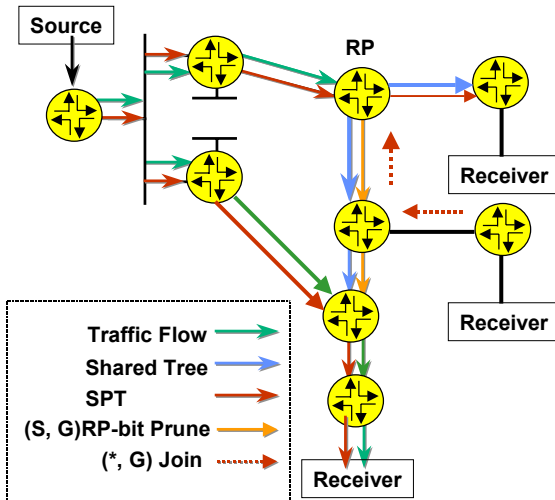
- it sends (\*,G) prune upstream to the RP

- it marks (S,G) state pruned but doesn't send (S,G) prune to the upstream router

- if it receives packets from the source, then it sends (S,G) prune to the upstream router where it received the multicast packets, e.g. to R4

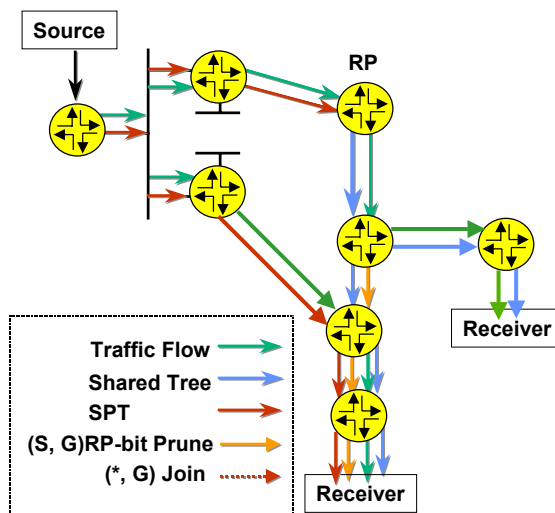
- if the source stops sending the multicast packet, then the (S,G) states along the SPT timeout

## Repairing (S,G) Pruned Segment



- A new receiver can join the segment of the RP-tree that was pruned by a (S,G) RPT-bit prune message causing (S,G) RPT-bit entries inducted there
  - It sends (\*,G) Join message
- A router when receives (\*,G) join message for the group G such that it already has (S,G) RPT-bit entry, then:
  - It updates the (S,G) RPT-bit entry (reset RPT-bit)
  - It sends (\*,G) Join upstream to cause similar updates to happen at upstream routers
- If the (\*,G) Join message contains in its prune list source S, then (S,G) RPT-bit entry remains unchanged

## Repairing (S,G) Pruned Segment



- Some receivers receive through shared tree, while others through SPT
- (S,G) RPT-bit entries along the pruned segments are modified to carry (\*,G) traffic
- Some pruned segments remain untouched by the (\*,G) join

## Sparse vs Dense Mode

---

- Advantages
  - Offers better scalability in terms of routing state
    - Only routers on the path between a source and a group member keeps state, whereas dense mode requires state in all routers in the network
  - More efficient because of explicit join message
    - Multicast traffic only flows on the links that are explicitly added, whereas dense mode assumes all links multicast enabled unless they are explicitly pruned
- Disadvantages
  - RP is a single point of failure
    - Bootstrap protocol mitigates this problem
  - RP becomes hotspot for multicast traffic
  - Traffic through source and shared trees means non-optimal paths may exist in the multicast tree
    - PIM-SM allows the receiver to join source specific shortest path tree, whereby traffic stops taking detour through RP

## Multicast Deployment - MBone

---

- Multicast Backbone has been in operation since 1992 when it first multicast worldwide a live event of IETF meeting in San Diego to 20 sites
- MBone is an overlay network connecting multicast enabled routers with each through the Internet
  - The multicast routers on MBone exchange multicast packets over IP tunnels
    - Each tunnel connects two endpoints via unicast over the Internet
  - Each router runs *mrouterd*
    - *mrouterd* receives multicast packet encapsulated in a unicast packet, multicasts it within the router domain and sends via tunnel to other multicast router on the Mbone
    - *mrouterd* runs DVMRP for building multicast tree

## MBone Experience

---

- As the Mbone has grown in size it experiences **state explosion** and has become more **susceptible to misconfiguration**
- Mbone problems
  - Scalability
    - At its peak Mbone had 10,000 routes and most of these routes had long prefixes (/28 to /32)
    - Low aggregation
  - Manageability
    - Lack of procedures cause Mbone to grow randomly generating inefficiencies
    - Virtual topology (tunnel) management
      - Single physical link carrying multiple tunnels
      - MCI restricted the tunnel to end on designated endpoints
    - Inter-domain policy management
      - Flat topology connected through tunnels causes routing problems spread throughout the overlay network
- Mbone community realized the need for the deployment of **hierarchical inter-domain routing**

## Multicast Deployment Motivations

---

- **Market Motivations**
  - Sources want to scale their services to large audience
  - Low-capacity domains require multicast when many redundant high-bandwidth streams threaten the capacity of incoming links
- **Possible Applications**
  - Audio and Video Distribution (Webcast)
  - Push application (Information Delivery)
  - Audio and Video Conferencing
  - File transfer (webcaching, distributed database, remote logging)
- **Customer Requirements**
  - Ubiquitous global access to multicast service
  - It is easy and transparent to install
  - Senders expect group membership to be controlled
  - Content providers expect that their assigned group addresses are unique (at least for the duration of a session)
  - Reliable transmission may be required

## Multicast Deployment Issues

---

- **Router Migration**
  - Carrier router migration model is from core to edge
  - The current multicast model requires multicast support at the edge
- **Domain Independence**
  - For application with many low-rate sources, it might be more efficient if all sources share the tree
  - Inter-domain deployment of PIM-SM or CBT requires MSDP to join the RPs located in distinct domains
  - Problems when RPs and sources are in distinct domains:
    - Traffic sources in other domains require traffic controls, such as rate and congestion control
    - ISP has very little control over the service its customers receive from the remote RP located in other domain
    - ISPs don't want to be the core of a session for which they have no receivers or sources since it is a waste of their resources
    - Scalable and low-latency RP advertisement
- **Management**
- **Cost Recovery and Profitability**

## Functionality Not Addressed

---

- **Group Management**
- **Multicast Security**
- **Address Allocation**
- **Network Management**
- **Additional Services**
  - Service Level Agreement (SLA) and Virtual Private Networking (VPN)
  - Network Performance Measurement
    - Measurement data, e.g. highest transmission delay to a group member, if sent to the sender may help it adjust its transmission
  - Subcasting
  - **Congestion Control**
    - Without congestion control multicast sessions may threaten to steal bandwidth from unicast TCP sessions
  - Low-latency Inter-domain Routing
    - Inter-domain routing should be as fast as intra-domain routing
  - Unidirectional Links
    - Satellite links are unidirectional and form asymmetric routing paths

## **Near-term Inter-domain Deployment**

---

- The near-term solution has three components
  - An intra-domain multicast protocol → PIM-SM
    - Most vendors are now supporting PIM-SM
  - The RP in domains that have receivers must know the source IP address
    - A new protocol to build trees across domain boundaries → MSDP
  - All routers along the path from the source to a receiver must have an entry corresponding to the shortest path route to the source in its RPF table
    - Extension to BGP to carry multicast source information → MBGP