Advanced Technologies TCP/IP

Master (RITC) course y.2

Chapter 1. Multicast Protocols and Applications

Eugen Borcoci

University POLITEHNICA of Bucharest
Electronics Telecommunication and Information Technology Faculty
Eugen.Borcoci@elcom.pub.ro

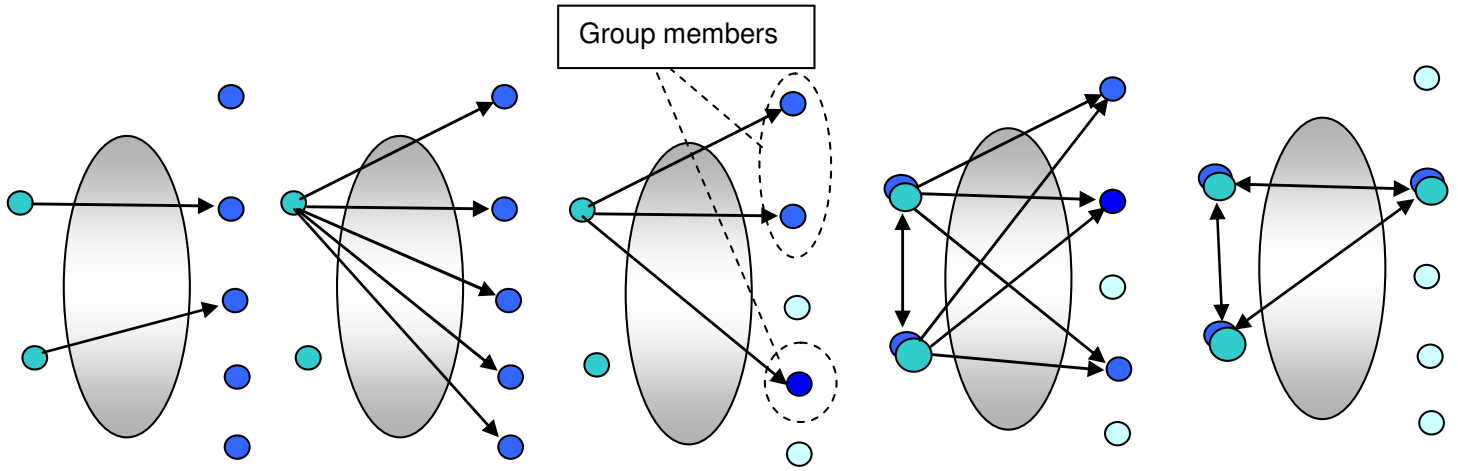
2014-2015

Contents

- 1. Multicast General Overview
- 2. Multicast Routing Algorithms
- 3. IP Level Intradomain Multicast
- 4. IP Multicast Networks (1)
- 5. Inter-domain IP level multicast protocols
- 6. Overlay Multicast
- 7. Multicast based applications and services
- 8. Open issues in IP networks multicast

1.Multicast General Overview

- **Unicast versus multicast**



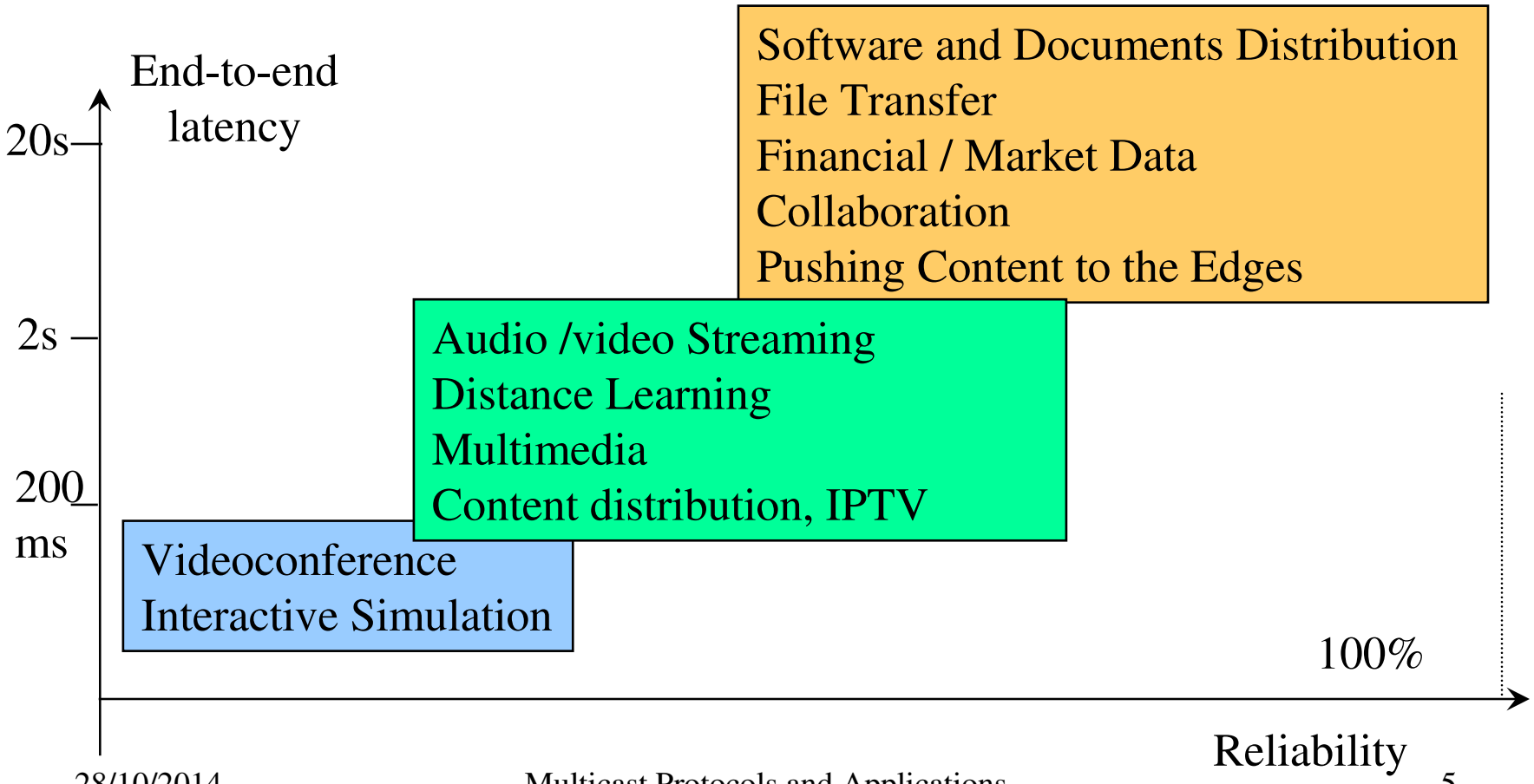
1-to-1 unicast	1-to-M broadcast	1-to-n n ≤M Multicast	n-to-m Panel Multicast	n-to-n Symmetric Multicast

1.Multicast General Overview

- **Unicast versus multicast** (previous slide)
- The multicast is a group communication in which one or more senders or receivers are involved.
- This slide presents the main scenarios for a multicast communication.
- The basic form is unicast in which we have one sender/source and one receiver. Of course,
- the communication can be also bidirectional.
- The trivial extension is the *broadcast (one-to-all)* : here the source information is broadcasted to all possible receivers.
- The multicast itself suppose that only some receivers which subscribe to a group will get the information. In the most simple case we have one sender and several receivers.
- The most general case is (m-to-n) where m denotes the number of senders and n the number of receivers. Despite not shown in the figure, in practice, depending on applications, a host can play the both roles: sender and receiver.

1. Multicast General Overview

- **Applications using multicast**

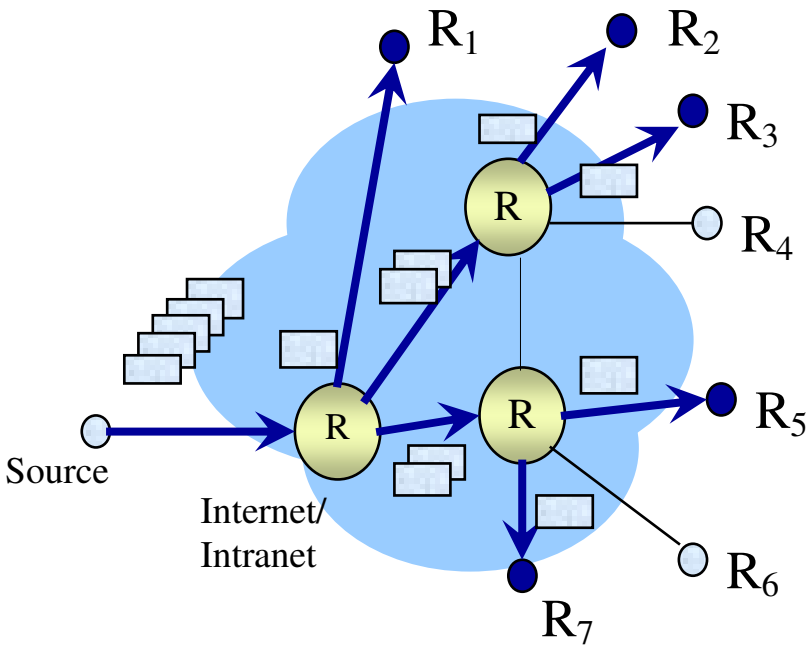


1. Multicast General Overview

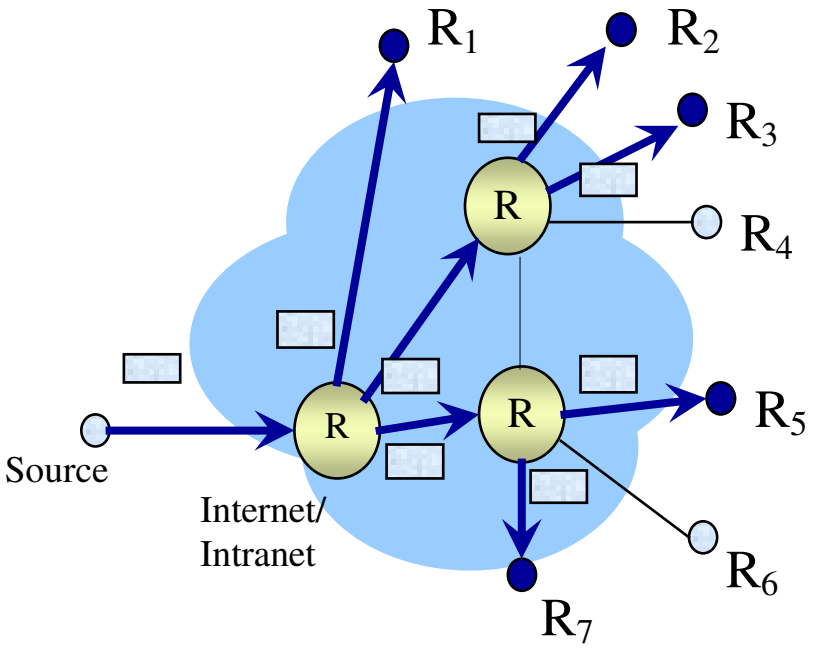
- **Applications using multicast**
- **(previous slide)**
- This slide presents some applications needing multicast communication.
- They are shown depending on their reliability and latency requirements.
- The most sensitive applications w.r.t. latency are the Videoconference and Interactive Simulation. The reliability requirements for such applications are low in comparison to others. This is the consequence of the nature of these applications.
- The upper left group of applications require high reliability. Here we have some examples: Software and Documents Distribution, File Transfer, Financial / Market Data Collaboration and Pushing Content to the Edges.
- In the middle of the range (latency- reliability) are the Audio / video Streaming, Distance Learning and Multimedia entertainment application.
- The latency and reliability requirements put some constraints on the algorithms and multicast protocols designed to support different kind of application. It is one reason that we can find a lot of protocols standardised or proposed, trying to solve these problems.

1. Multicast General Overview

- **Multicast versus Unicast Efficiency**
  - **Example at IP level**



Unicast: N ≥ 1, TCP/IP connections
multiple packet copies



Multicast: one multicast association
one packet generated by source

1. Multicast General Overview

- **Multicast versus Unicast Efficiency (previous slide)**
- One of the multicasting challenges is to minimise the amount of resources employed when delivering the same message to multiple destinations.
- To illustrate this point the figure presents two approaches to distribute one packet to several destinations. In unicast approach we should have as many connections as the number of receivers. Multiple copies of the packets circulate on the links, the result being the overloading of the network with redundant traffic.

- The multicasting (when deployed at the network layer) is much more efficient in bandwidth utilisation because it avoids the packet multiplication on the links. Instead of many 1-to-1 connections, we have now a tree structure (in our example rooted in the source) overlaying the physical network. So, one router sends only one copy of a multicast data packet on one output interface, no matter how many the recipients reachable via this link are.

- Some problems appear with reliable multicast protocols (usually deployed at the transport layer), because a lot of control packets are necessary to solve the reliability. This problem appears especially when the number of receiver is large. The so called "ACK and NACK implosion problem" refers to the significant number of control packets that can overload the network. One challenge of the transport reliable multicast protocols is to minimise the number of these control packets while preserving the reliability.

1. Multicast General Overview

- **Multicast groups taxonomy**

- ***Geographical Density***
  - ***Dense groups*** have members on most of the links or subnets in the network
  - ***Sparse groups*** have members only on a small number of widely separated links.
- ***Open/Close groups***
  - ***Open***– the sender/source need not be a member of the group
  - ***Closed*** - allow only members to send to the group.
- ***Permanent/transient groups***
  - ***Permanent*** - exist forever or for a longer duration
  - ***Transient –*** exist for a short time
- ***Static/Dynamic groups***
  - *Static* - whose membership remains constant in time
  - *Dynamic groups* - allow members to join/leave the group during a mc session.

1. Multicast General Overview

- **Multicast advantages**
  - Efficient utilization of bandwidth
    - Avoid excessive multiplication
  - Reach N user (~simultaneously)
    - If the distribution tree is enough balanced and satisfies delay constraints
  - Suitable for group targeted applications
  - Lower load for servers
    - Avoid n >> 1 connections starting at servers
  - Congestion avoiding (! some problems with reliable multicast)
    - Due to previous point
    - Problems in reliable multicast when Acks are generated by many sources- ACK/NACK implosion
  - Suitable for “push” applications

1. Multicast General Overview

- **Multicast algorithms and protocols requirements and problems (1/3)**
- **They should be able to offer/solve different problems/challenges**
  - **Scalability**
    - signalling overhead at tree building
    - Number of states to be stored in multicast tree nodes
  - **Complexity** of processing in mc routing
  - **Network heterogeneity** (routing protocols, link technology, different hosts)
  - Internet cannot support multicast everywhere **(legacy routers)**
    - Crossing not-multicast aware routers (tunnelling techniques)
  - **Group management** (quorum, join actions, leave actions, not-enough capable receivers, etc.)
  - **QoS related problems** – for multimedia and real-time communications
    - Finding QoS constrained trees
    - Maintaining the QoS properties after initial tree construction
    - Resource reservation?/Diffserv philosophy?

1. Multicast General Overview

- **Multicast algorithms and protocols requirements and problems(2/3)**
- **They should be able to offer/solve different problems**
  - **Mobility and multicast**
    - variable topology – additional difficulties, especially when seamless mobility is wanted
  - **TCP friendlliness-** not all mc protocols satisfy that
  - **Security**
    - Members authentication and access rights
    - Confidentiality of distributed information
    - Protection of distributed information against altering, etc.
  - **Fairness** (for routers, receivers, etc.)
  - **Survivability of multicast trees**
    - In case of central points failure
    - In case of node/link failures

1. Multicast General Overview

- **IP Multicast protocols requirements and problems (3/3)**
- **Multicast works over UDP, therefore we have (in traditional TCP/IP stack)**
- **Best Effort Delivery**
  - IP multicast offers non-reliable data delivery
  - Drops are to be expected
  - Reliable Multicast is still a separate issues (solutions exist at transport layers)
- No Congestion Avoidance
  - No window based flow control, or "slow-start"-like TCP
  - Network congestion is possible
  - Multicast applications should attempt to detect and avoid congestion conditions.
- Duplicate packets
  - Some multicast protocol mechanisms may result in the occasional generation of duplicate packets
  - Multicast applications should be designed to expect occasional duplicate packets.
- Delivery order alteration - possible
  - IP possible delivery order alteration
  - Some protocol mechanisms may also result in out of order delivery of packets.

1. Multicast General Overview

- **Multicast algorithms and protocols requirements and problems (1/3)- previous slides details**

- **Scalability** means mainly two aspects:
- -multicast tree construction requires a reasonable computation effort
- - the routers in a WAN have to suuport many simultaneous multicast trees.
- **Network heterogeneity** reffers to different routing protocols, different link technology and different hosts w.r.t resources and operating systems.
- **Internet cannot support multicast everywhere**. There exist many legacy routers in the network that cannot support IP multicast. They have to be crossed by using special encapsulation techniques (tunnelling).

- **The multimedia and real-time communications** have special requirements and guarantees for Quality of Services (QoS). Examples of QoS important parameters are : delay, jitter, packet loss, bandwidth. The multicast protocols have to take into account these requirements.

- **Mobility and multicast**
- Additional complexity is encountered in multicast transfer when mobile network are involved. The multicast tree construction algorithm has to adapt dynamically when the hosts change their location in the network.

- **Security** problems are more complex in multicast communications especially when the multicast deals with large groups and sparse locations. The security requirements are varying also with the type of applications that use multicast.

- **Fairness** means that a good multicast tree has to provide a minimum quality of service to each member of the multicast group. Also the multicasting effort has to be evenly divided among the participant nodes.
- **Survivability** requires that a multicast tree must be able to survive multiple node and link failures.

1. Multicast General Overview

- **Multicast related functions**
- **group management**
  - join, leave, source filtering
- **routing, source discovery**
  - tree searching algorithms
    - source routed tree, shared tree (unidirectional/bidirectional)
    - pruning, grafting the tree
- **scalability related** (large number of members / networks)
- **reliability**
  - FEC, retransmission (ACK, NACK)
- **fault tolerance** (for essential tree elements)
- **QoS guarantees** (QoS routing related to IntServ, Diffserv, MPLS)
- **congestion control**
- **mobility related functions**
- **security**

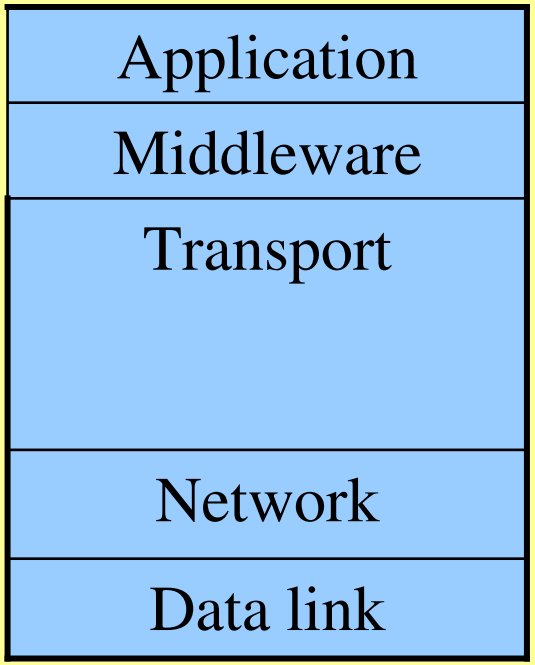
1. Multicast General Overview

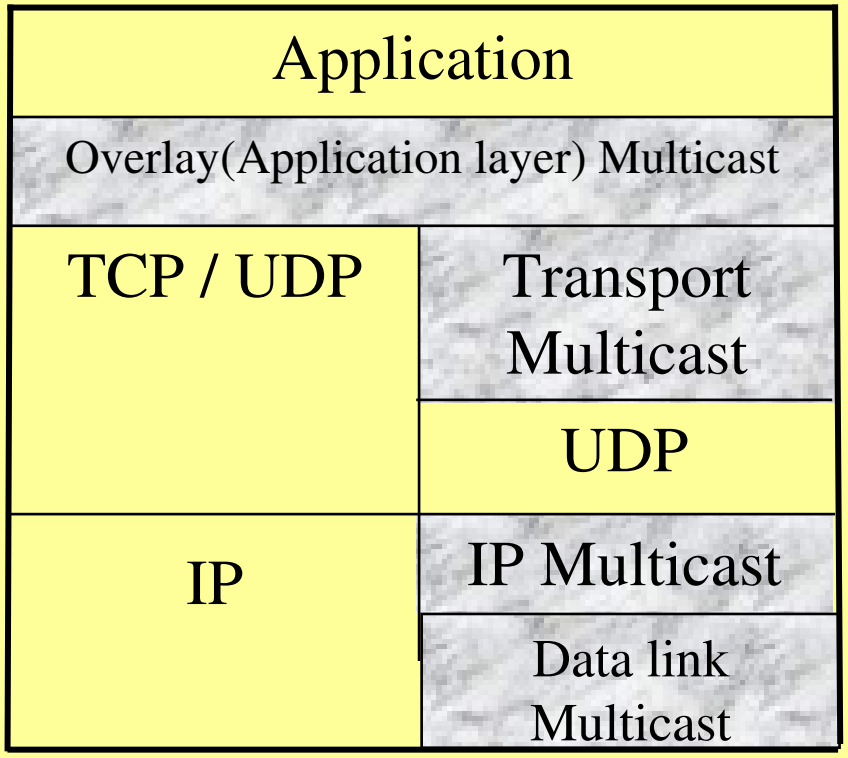
- **Multicast related functions (previous slide)**

- This slide lists the principal functions related to multicast. Depending on the objective to fulfill, a multicast protocol has to deal with a lower or a larger part of this list.
- The group management: members of the group should be permitted to dynamically join or leave the group. Some source filtering is necessary to be implemented in some cases.
- A basic function of multicasting is the routing. Many algorithms proposed led to many routing protocols. The source discovery is a function needed especially in large networks having several administrative domains. Basically the routing function is supported by some tree searching algorithms. One can find two basic form of a tree used in multicast protocols: source routed tree (having the root in the source and leaves at destination recipients) and shared tree (unidirectional/bidirectional) used for distribution by several sources.
- The tree can be pruned (cutting some leaf branches at request) or grafted (adding new branches).
- The scalability is a requirement. Some protocol functions are influenced by this requirement (especially in case of large number of members / networks)
- The reliability is usually accomplished by error detection and correction functions based on Forward Error Control, ACKnowledges, No-ACKnowledges mechanisms usually associated with timer based mechanisms.
- The fault tolerance is performed by functions dedicated to re-establish the tree in case of network elements failures.
- Assuring a certain level of QoS in multicast communication is a a complex and open issue. We just mention some directions of research: Finding QoS constrained routes, maintaining QoS on the tree, the heterogeneous QoS problem, Inter-domain QoS Routing, integration with the Resource Reservation Protocol. The integration of Multicast with new QoS related technologies like DiffServ and MPLS are also subjects of interest.
- The congestion control functions are targeted to control the amount of data and control packets on the multicast tree in order to avoid the overloading the network.
- The main problem in a mobile environment is to establish and maintain the multicast trees in conditions of terminal mobility. Adding QoS requirements puts more complexity on the solutions.
- Last but not least the security is a crucial issue in some multicast applications. Here one can include: Routing Protection, Content Protection and Tree Access Control

1. Multicast General Overview

- Which layer – to deploy multicast?





1. Multicast General Overview

- **Which layer – to deploy multicast? (previous slide)**

- This slide shows the different layers in an architecture stack where multicast protocols can exist.

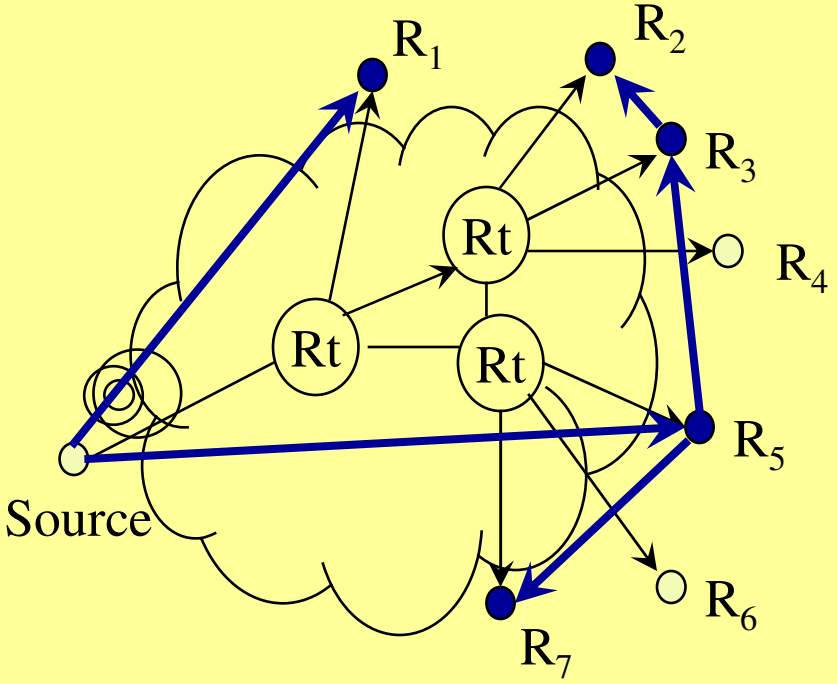
- **The Network layer** can include multicast protocols – permitting distribution of one copy of a packet to several destination points. Special multicast tree-finding algorithms and IP layer multicast routing protocols should be developed. The multicast aware routers must coexist with the legacy ones.

- **The Transport layer** multicast usually deals with problems like reliability or real-time transfer of data. Usually the multicast transport protocols are optimised for certain applications.

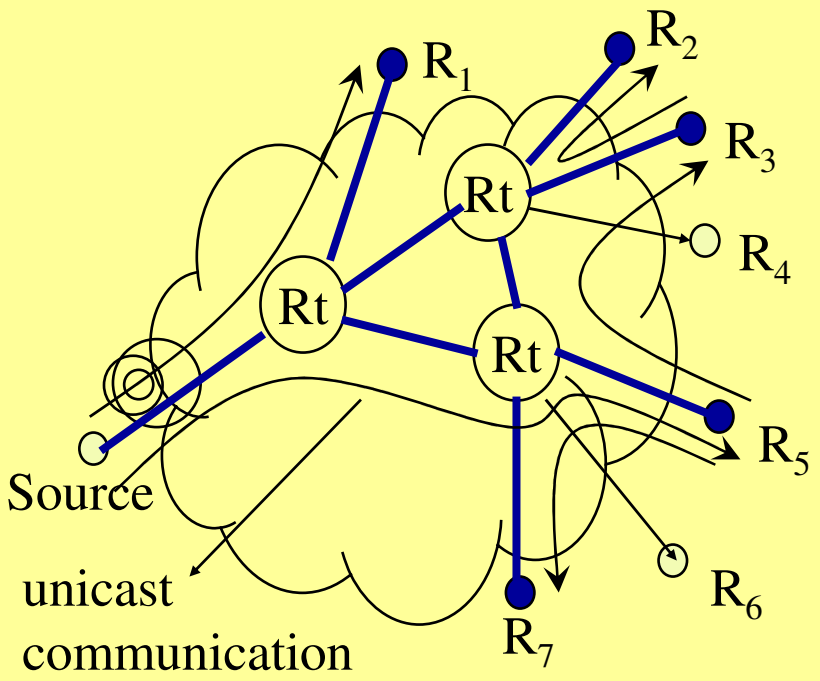
- **The Overlay Multicast** (End System Multicast = Application Layer Multicast) is a new approach which does not assume that IP layer is multicast aware. The task of realising the multicast distribution is put on some End Systems (servers).

1. Multicast General Overview

- **Overlay (application) layer multicast (1/2)**



Logical view
(overlay distribution tree)



Real traffic flow
(unicast communications)

1. Multicast General Overview

- **Overlay (application) layer multicast (previous slide)**
- We recall that one problem of the current internet (despite many IP routing protocols developed) is that it cannot support multicast everywhere. We have seen that some solutions have been found and applied (MBone, Inter-domain protocols). But in practice, the IP Multicast, IP Security, IP v6, QoS- aware protocols are not currently widely deployed.
- An alternative solution seems to be a new Infrastructure Software called Overlay Multicast (Application Layer Multicast)
- The basic characteristics of **IP Multicast** is that uses the routers to replicate packets and therefore requires multicast capable routers. For reliable delivery one can rely on multicast reliable transport protocols. The IP multicast protocols are (a lot of them already standardised).
- **The Overlay (Application) Multicast** - uses overlay servers to replicate packets. But it can take advantage of IP multicast if available. Currently these protocols are not yet standardised. The general approach is that applications
- self-organize into a logical overlay network, and transfer data along the edges of the overlay network using unicast transport services. Each application communicates only with its neighbors in the overlay network. By forwarding packets from neighbor to neighbor, multicast forwarding is performed at the application layer.
- The figure shows an example of a logical view for application layer multicast and the actual connections setup in the network. We can see that an unicast capable network is sufficient.

1. Multicast General Overview

- **Overlay (application) layer multicast (2/2)**

Characteristic	a. IP multicast	b. Overlay multicast
Packet Replication	routers	servers
Multicast capable routers	required	not required (Can take advantage if routers knows multicast)
Deployment	slow	quick
Reliability	transport layer or higher level	higher level
Latency	lower	higher
Bandwidth utilization efficiency	higher	lower (multiple copies on the same link)
Security	lower	higher
Standardised	yes	proprietary

1. Multicast General Overview

- **Overlay (application) layer multicast (previous slide)**
- This table shows a comparison between IP layer based multicast) multicast aware routers) and Overlay Multicast.
- The *packet replication* is performed by the routers in case a. while this task is done by the servers in case b.
- The immediate consequence of this is that in the case a. we need multicast capable routers while in b. there is no need to have such routers. In first case the routers have to maintain per/group state which violates in a sense the stateless principle of IP layer. In case b. all multicast state are maintained in end systems. Computation at forwarding points simplifies support for higher level functionality.
- Therefore the deployment of IP multicast is slower than overlay multicast.
- The *reliability* of the transfer is assured by some transport protocols in case a. (e.g. PGM) or is not guaranteed at all. In case b. the reliability is solved by the application layer.
- *Latency* is usually higher in case b. because of the transfer principle (see later discussion).
- The *bandwidth utilization* efficiency is better in case a. because at the network layer the packet replication is avoided.
- *Security* is lower in IP multicast because the IP multicast model permits to any source to send to any group. This makes the network more vulnerable to flooding attacks.
- The *Overlay multicast* protocols are still proprietary while for *IP multicast* there exist many IETF RFCs standards.

1. Multicast General Overview

- **Multicast protocols in fixed Internet – examples (1/3)**
- **Data link layer – no special protocols**
- **IP Layer:** ***Any Source Multicast - model***
  - IGMP – Internet Group Management Protocol (V.1, V.2, V.3)

- **IP Intra-domain Routing Protocols - Dense Mode :**
  - DVMRP – Distance Vector Multicast Routing Protocol
  - MOSPF – Multicast Open Shortest Path First
  - PIM-DM - Protocol Independent Multicast – Dense Mode

- **IP Intra-domain Routing Protocols - Sparse Mode :**
  - CBT – Core Based Tree, OCBT – Ordered Core Based Tree
  - PIM-SM – Protocol Independent Multicast – Sparse Mode

1. Multicast General Overview

- **Multicast protocols in fixed Internet – examples (1/3) previous slide**
- At data link layer no special multicast protocol is necessary. In the simplest case the broadcast feature of usual LANS at data link layer is used to distribute mc-data units to all potential receivers on a LAN. Some mapping of the IP group addresses onto the MADC addresses can be used. This can avoid the processing IP mc packets, in all stations on a LAN at IP level, to discover whether or not the station is destination for the current packet.

- At IP layer several protocols have been proposed and standardised. They are basically dealing with group management and mc-tree construction, tree maintenance and data delivery. All these protocols will be discussed later in this tutorial.

- **Group Management protocol**
- *IGMP - Internet Group Management Protocol* - is the protocol by which hosts report their multicast group memberships to neighboring routers.
- The routing protocols can be divided in two classes – depending on the policy adopted to build the mc distribution tree: *dense-mode* - for groups with many members in a region and *sparse –mode* – for widely distributed groups.
- Examples of dense-mode *protocols are:*
- *DVMRP – Distance Vector Multicast Routing Protocol* - is the multicast equivalent of unicast routing protocol based on distance vector algorithm - RIP
- *MOSPF – Multicast Open Shortest Path First* – is the multicast equivalent of unicast routing protocol based on link state – OSPF
- *PIM-DM - Protocol Independent Multicast – Dense Mode* – is a protocol independent of routing protocol used by the network layer. It is used if the multicast receivers are located close to one another.
- ***DVMRP, MOSPF, PIM-DM* – build spanning trees that are the the shortest path from each source**.

- **Examples of sparse-mode *protocols are:***
- *CBT – Core Based Tree, OCBT – Ordered Core Based Tree and PIM-SM – Protocol Independent Multicast – Sparse Mode* – build multicast spanning trees that are the shortest path from a known central node called *Rendez-vous Point (RP).* All sources in the session share the same spanning tree.
- *BGMP – Border Gateway Multicast Protocol* – is the multicast extension of the unicast Border Gateway Protocol BGP.

1. Multicast General Overview

- **Multicast protocols in fixed Internet – examples (2/3)**
- **IP Layer:**
- **IP Inter-domain Routing Protocols**
  - *First solution (>1992)*
    - Multicast Backbone (MBone)- flat network linking multicast capable islands
  - *Currently developed (>1999) Near-Term solutions*
    - MBGP- Multi-protocol Extension of BGP4
    - PIM-SM used as inter-domain multicast routing protocol
    - MSDP – Multicast Source Discovery Protocol
  - *Long-Term Proposal – for Internet-wide inter-domain multicast*
    - BGMP – Border Gateway Multicast Protocol
    - MASC – Multicast Address Set Claim

- Source Specific Multicast (SSM) – simplified model of ASM

1. Multicast General Overview

- **Multicast protocols in fixed Internet – examples (2/3) (previous slide)**
- The deployment of multicast network is related to that of inter-domain routing.

- The first solution of multicast network deployment has been the MBone. This structure is a flat multicast backbone overlaying the internet and permitting interconnection of multicast capable routers/islands. The technique used to cross the non-multicast aware part of the Internet has been to encapsulate (tunneling) the multicast packets in unicast packets tunneled between multicast capable routers. The main protocol used in MBone implementations has been DVMRP.

- MBone has some drawabacks (we will discuss them later in greater details) – a main one being the flat structure of addresses with all unpleasant consequences for large networks.
- Therefore other solutions have been investigated.

- A near term solution permitting a hierarchical approach is the development of inter-domain routing protocols, like PIM-SM, complemented with an Multicast extension of the Border Gateway Protocol BGP4 (MBGP). Discovering the sources in other domains is accomplished by MSDP – Multicast Source Discovery Protocol.

- A long-term proposal for solving the inter-domain routing problem is *BGMP – Border Gateway Multicast Protocol* – is the multicast extension of the unicast Border Gateway Protocol BGP. The key idea of BGMP is to build bidirectional shared (by several sources) trees between domains using a single root.
- The MASC – Multicast Address Set Claim protocol is used to support allocation of addresses between domains.

1. Multicast General Overview

- **Multicast protocols in fixed Internet – examples (3/3)**
- **Transport layer:**
  - reliable :
    - SRM – Scalable Reliable Multicast Protocol
    - RMP - Reliable Multicast Protocol
    - RMTP, RMTP–II - Reliable Transport Multicast Protocol
    - PGM – Pretty Good Multicast
    - ALC – Asynchronous Layer Coding
    - MFTP – Multicast File Transfer Protocol
  - real time :
    - RTP + RTCP – Real Time Protocol + RT Control Protocol
    - RTSP – Real Time Streaming Protocol
- **Application layer (related to P2P communications) :** Narada, Gnutella, BitTorrent, Skype, Joost, etc.

1. Multicast General Overview

- **Multicast protocols in fixed Internet – examples (3/3) (previous slide)**
- The multicast transport protocols are usually stacked over UDP transport layer. Their main task is depending on the application supported (e.g Reliable Transport Multicast Protocol – for reliable transfer or Real Time Protocol for real-time or multimedia stream transport over IP networks).
- The presentation of multicast transport protocols is out of the scope of this lecture. The slide simply lists some significant examples of transport protocols.
- We divided them in two groups:
- - having the reliability as the main objective
- - having the real-time transfer as main objective.
- There are some protocols aiming to solve (partially) the both requirements. This is not a trivial thing because the two requirements can be (at least partially) contradictory in some cases.

1. Multicast General Overview

- **IETF Standards examples for multicast (1/3)- partial list**
  - **General:**
    - RFC1112 Host Extensions for IP Multicasting (STD)
    - RFC1458 Requirements for Multicast Protocols (INFORMATIONAL)
    - RFC2627 Key Management for Multicast: Issues and Architectures (STD)
  - **IGMP:**
    - RFC2236 IGMPv2 (updates 1112)(PROPOSED STD)
  - **ATM:**
    - RFC2226 IP Broadcast over ATM Networks (PROPOSED STD)
    - RFC2149 Multicast Server Architectures for MARS-based ATM multicasting.

1. Multicast General Overview

- **IETF Standards examples for multicast (2/3)**
- **Network Layer:**
  - RFC1075 Distance Vector Multicast Routing Protocol (EXP)
  - RFC1584 Multicast Extensions to OSPF (PROPOSED STD)
  - RFC2189 Core Based Trees (CBT v.2) Multicast Routing (EXP)
  - RFC2201Core Based Trees (CBT) Multicast Routing Architecture (EXP)
  - RFC2858 Multiprotocol Extensions for BGP-4 (PROPOSED STD)
  - RFC4601 Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised) 2006
  - RFC 3569 An Overview of Source-Specific Multicast (SSM) (2003)
  - RFC 4607 –Source Specific Multicast – 2006
  - RFC 4608 Source-Specific Protocol Independent Multicast in 232/8

1.Multicast General Overview

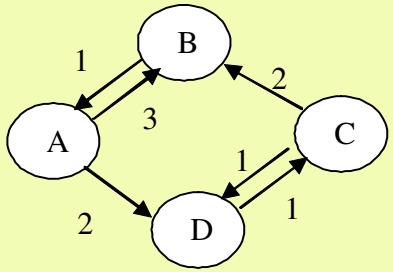
- **IETF Standards examples for multicast (3/3)**
- **Transport Layer:**
  - RFC1301 Multicast Transport Protocol (STD)
  - RFC1889 RTP: A Transport Protocol for Real-Time Applications(STD)
  - RFC2490 A Simulation Model for IP Multicast with RSVP (INFO)
  - RFC2887 The Reliable Multicast Design Space for Bulk Data Transfer (INFO)
  - RFC3048 Reliable Multicast Transport Building Blocks for One-to-Many Bulk-Data Transfer(INFO)

Contents

- *1. Multicast General Overview*
- *2. Multicast Routing Algorithms*
- **3. IP Level Intradomain Multicast**
- **4. Inter-domain IP level multicast protocols**
- **5. IP Multicast Networks**
- **6. Overlay Multicast**
- **7. Multicast based applications and services**
- **8. Open issues in IP networks multicast**

2. Multicast Routing Algorithms

- **Multicast Routing Problem (1/3)**
- **WAN network – graph (directed, undirected)**
- G(V,E); V- node set; E – edge set
  - Weights associated to links
  - Asymmetric links (A,B)
  - Symmetric links (D,C)



- Undirected graph – all links are symmetric
- **Graph problems**
  - Shortest path finding problems (between two points)
  - Minimum weight trees – sum of all weights – minimum
- Multicast communications:
  - ***Source specific*** (1-to-n, n<m, m = |V|, one source, n receivers)
  - ***Group shared*** (n-to-n; any node or some of them can be sender or receiver)

2. Multicast Routing Algorithms

- **Multicast Routing Problem (1/3) (previous slide)**
- A WAN network can be modelled as an directed or undirected graph, where the nodes represent the routers/switches and the edges represent the links.
- The figure shows a directed graph with asymmetric links (having different weights on the two directions).
- The links are associated with weights (costs). The links can be Asymmetric like (A,B) or Symmetric links (D,C).
- In an undirected graph all links are symmetric.
- The shortest path problem is to find the route between two points having the minimum cost.
- A multicast communications needs a tree containing the nodes involved in the multicast group. The minimum cost tree is of interest. Generally we can have two sort of trees:
  - Source specific (1-to-n, where n<m, m = |V| - the number of vertices. There is one source and n receivers)
  - Group shared (n-to-n; any node or some of them can be sender or receiver)

2. Multicast Routing Algorithms

- **Multicast Routing Problem (2/3)**
- Graph $G = (V, E)$ – undirected; $M \subseteq V$, M = multicast group
- find a tree $T \subset G$, s.t. T spans all vertices in M
- Tree types:

  A - *source routed tree-* employs unidirectional links

  – **What we are interested in?**
    - Interest to find the **shortest path** between two points (source, $dest_k$)
      – then a minimum cost tree $\Rightarrow$ **Shortest Path Tree** (SPT) **algorithms**
      – Dijkstra, Bellman-Ford – centralised or distributed – also used in unicast routing
    - Interest to find the **minimum of total cost** of the tree rooted in S
    - **Optimisation and/or constraints**

  – **B – *group shared tree-* employs uni or bidirectional links**
    - Interest to find **a minimum cost tree (sum of all costs = minimum)**
    - **Optimisation and/or constraints**

2. Multicast Routing Algorithms

- **Multicast Routing Problem (2/3) (previous slide)**
- Given an undirected graph, we are interested to find a tree that spans all vertices belonging to set of multicast nodes M (vertices of G).
- We consider the two Tree types:
  - A - source routed tree- which employs unidirectional links
  - B – group shared tree- which employs bidirectional links
- We are interested to find a minimum cost tree (no matter of what type is source routed or shared)
- For case A (source routed tree) there exist some shortest path algorithms (SPT) (e.g. Dijkstra, Bellman-Ford) – run in centralised or distributed manner. On such algorithms are based the unicast routing protocols. Here, minimum cost tree means that tree that has the sum of all costs = minimum.
  - Some popular implementations of such SPT algorithms does exist:
    - Distance Vector Algorithm - used in RIP protocol
    - Link state + Dijkstra algorithm –used in OSPF protocol

2. Multicast Routing Algorithms

- **Multicast Routing Problem (3/3)**

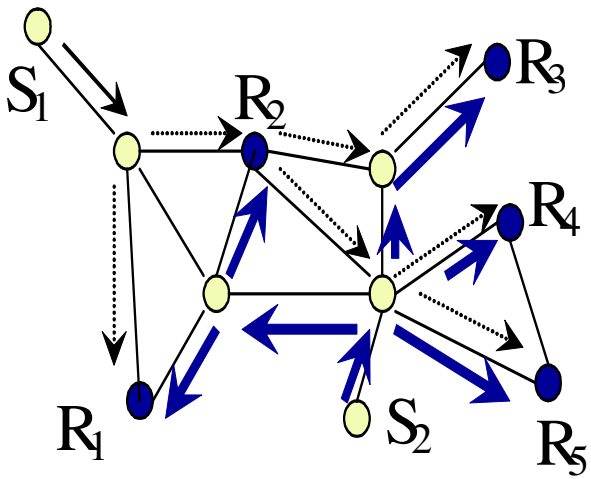
  Implementation examples of SPT:

  - Distance Vector Algorithm - used in RIP protocol
  - Link state + Dijkstra algorithm –used in OSPF protocol

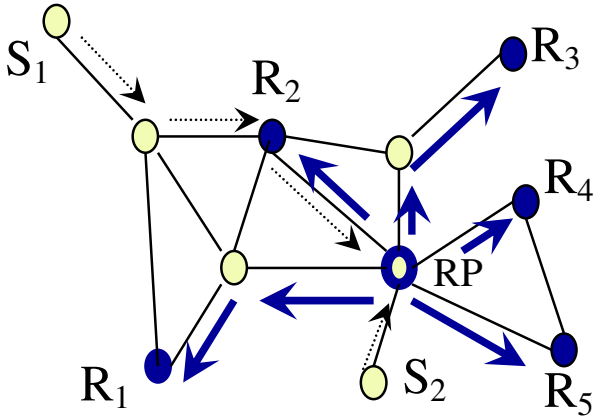
- **How much memory is required in routers?**
  - **Source or Shortest Path trees**
    - Uses more memory ***O(S * G)*** but you get optimal paths from source to all receivers; can minimize delay
  - **Shared tree**
    - Uses less memory ***O(G)*** but you may get sub-optimal paths from source to all receivers; may introduce extra delay

2. Multicast Routing Algorithms

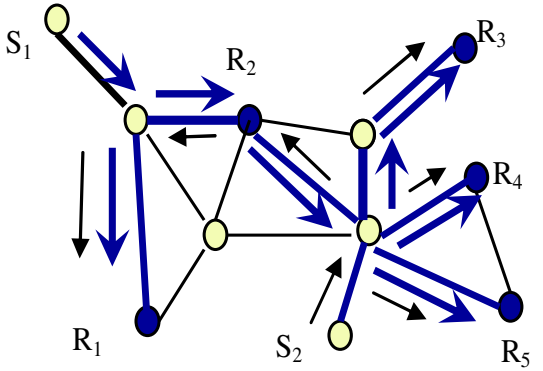
- **Tree types used in mc routing protocols**



Unidirectional tree
One tree per source
S_1 rooted tree (SPT)
S_2 rooted tree (SPT)
Optimised for source specific mc communication



Unidirectional Shared (by all sources) **tree**
Components:
Shared tree
Data path S1→RP
Data path S2 → RP



Bidirectional Shared Tree
Distribution of S1 data
Distribution of S2 data

2. Multicast Routing Algorithms

- **Tree types used in mc routing protocols (previous slide)**
- 1.Unidirectional tree case. We have one tree per each source. In our example two tree exist rooted the sources S1 and S2.
- 2.Unidirectional Shared tree is a shared resource for all sources. The tree has a central core (root). It is named also Rendez-Vous Point (RP) in some protocols. The shared tree is unidirectional in the sense that all data packets are distributed towards receivers starting from RP. Therefore each source has to first send its data to RP which in its turn distributes the packet on the shared tree.
- 3. Bidirectional shared tree. In this case all sources can use every part of the tree they want to send packets to the receivers.
-

2. Multicast Routing Algorithms

- **Properties of a good multicast tree**
  - **High priority properties**
    - Low cost (minimum cost is desired)
      - (source –destination, or total cost)
    - Low delay between (S,D) pairs
    - Scalability
      - reasonable tree computation time for large networks
      - number of trees supported by the network nodes
  - **Medium priority properties**
    - Support of dynamic multicast groups (join, leave of members)
    - Survivability (in case of node or link failures)
  - **Low priority properties**
    - Fairness
      - w.r.t. QoS offered to different nodes
      - fairly divide the effort of packet multiplication between nodes

2. Multicast Routing Algorithms

- **Multicast Algorithms Taxonomy, [3]**
- **Mc routing process builds a multicast tree *T*, optimising some objective functions.**
- **Additionally: in a QoS related context, a set of constraints (e.g. end-to-end (E2E) delay bound, jitter, bandwidth, loss or combinations of them) have to be met**
- **Criteria of classification: *constraints* and *optimisations* required**
- **The resulting mc tree must provide**
  - reachability from source(s) to a set of destinations
  - paths satisfying the constraints (QoS related bounds)
- **A large set of combinations may exist between requirements to fulfill:**
  - **Optimisation**
    - **Paths (least cost path)**
    - **Tree as a whole – e.g. minimum cost tree**
  - **Constraints (one or more)**
    - **On links**
    - **On paths**
    - **On the whole tree**
- **Important issue : complexity of calculus**
  - polynomial time complexity (PC)
  - **or NP-complete complexity (NP-C).**

2. Multicast Routing Algorithms

- **Multicast Algorithms Taxonomy**
- **Constraints**
  - to a link (e.g bandwidth, available buffer, etc.).
  - to a path or to the whole tree,
- **Tree constraints can be** expressed by using metrics (m = metric)
  - ***additive*** (e.g. E2E delay on every path from source to destination)
    - ***m(u,v) = m(u,i) + m(i,j) + …m(pv), for a path P(u,i,j, ..v)***
    - Sum of the costs on all edges of the tree
  - ***multiplicative*** (e.g the probability that a packet will reach the destination, being given the loss probability on each link)
    - ***m(u,v) = m(u,i) *m(i,j) * …m(pv), for a path P(u,i,j, ..v)***
  - ***concave*** (e.g. minimum bandwidth on a chain of links on a path)
    - ***m(u,v) = Min{ m(u,i), m(i,j), …m(pv)}, for a path P(u,i,j, ..v)***

2. Multicast Routing Algorithms

- **Multicast Algorithms Taxonomy**
- ***A.Constraint Types***
- **1.Link-constraints problems:**
  - a. Single - each link has a constraint, e.g., bandwidth-constrained
  - b. multiple- constraints on each link
    - e.g. link- problem: bandwidth and buffer-constraints.
- ***2. Tree-constraints***
  - a. Single - imposed to paths of the tree or to the whole tree
    - (e.g., path delay constraint)
  - b. Multiple-constraints problem:
    - (e.g., delay-constraint + inter-receiver-delay-jitter-constraints).
- ***3. Combined link- and tree-constrained***
- (e.g., delay for the tree and bandwidth for links -constraints).

- ***B. Optimisation problem***
  - ***1. Link/path optimization***
    - e.g., maximization of the link bandwidth of a path
  - ***2.Tree optimization***
    - e.g., minimization of the total cost of a mc tree = *Steiner tree* problem.

2. Multicast Routing Algorithms

- **Multicast Algorithms Taxonomy**
- ***C. Combined constraints and optimisation required***
  - 1. *link-constrained* and *link/path optimization*
    - e.g., bandwidth-constrained links, buffer optimization problem for the path
  - ***2.Link-constrained*** and ***tree optimization***
    - e.g., the bandwidth on links+ constrained Steiner tree problem).
  - **3.** ***Tree-constrained*** and ***link/path optimization*** routing problem
    - e.g., the delay-constrained for the tree and bandwidth optimization problem for the path
  - 4. ***Tree-constrained*** and ***tree optimization***
    - e.g., the delay-constrained and Steiner tree problem
  - ***5. Link constraint*** and ***tree constrained*** and ***tree optimization***
    - e.g.: min guaranteed bandwidth for every link
    - delay-constrained for every path
    - overall tree optimization problem

2. Multicast Routing Algorithms

- **Multicast Algorithms Taxonomy**
- **Notations:** link optimisation –LO; tree optimisation –TO; link constraints –LC; tree constraints –TC
- **Notation Examples**
  - LO/PC - *Link Optimisation/Polynomial Complexity*
  - LC/PC- *Link Constraint/Polynomial Complexity*
  - *NP – non polynomial complete*
- **What solutions exist for the above problems?**
- Note: Wang and Crowcroft [10] :
  - finding a path with ≥ 2 independent additive and/or multiplicative constraints in any possible combination is NP-complete.
  - the only tractable combinations : *concave constraint* and the other could be an *additive/multiplicative* constraint.

2. Multicast Routing Algorithms

- **Multicast Algorithms Taxonomy**
- **Results of analysis [3] :**
- ***A1.a, A1.b - Link-constraints* (Single or multiple) *problems***
  - **tractable, by removing links that not meeting the constraint**
- ***A2.a – Single constraint on Tree* - PC**
- ***A2.b – Multiple constraints- on tree* - NP-complete**
- ***A3 - Combined link- and tree-constrained problem***
- **Remove some links and reduce the problem to A2.a - PC**

- ***B1. Link/path optimization* – shown as solvable - PC**
- ***B2 Tree optimization: Steiner tree* problem - NP-complete**

- ***C1. Link-constrained* and *link/path optimization***
- **Reducible to B1 if the links not meeting the constraints are removed – PC**
- ***C2. Link-constrained and tree optimization* - Reducible to B2, NP-complete**

- ***C3. Tree-constrained* and *link/path optimization***
  - **Shown as polynomial time solvable - PC**
- ***C4. Tree-constrained* and *tree optimization* NP-complete**
- **C5. *Link constraint* and *tree constrained* and *tree optimization* -NP-complete**

2. MulticastRouting Algorithms

- **Multicast Algorithms Taxonomy**
- **Results of analysis [3] :**

Link/path **Tree**		**Single Constraint**	**Two or more Constraints**	**Link/path Optimisation**	**Link Constraints and optimisation**
		A1.a e.g. Bw_{min}, Buf_{min} PC	**A1.b** e.g. Bw_{min} and Buf_{min} PC	**B1** e.g. B_{max} PC	**C1** e.g. Buf_{min} and BW_{max} PC
Single Tree Constraint	**A2.a** e.g. D_{max} PC	**A3** e.g.T: D_{max} L: Bw_{min} for **PC**		**C3.** e.g., T: D_{max} L: Bw_{max} PC	
Two or more Tree Constraints	**A2.b** e.g., D_{max}, J_{max} NP-complete				
Tree optimisation	***B2-Steiner tree*** problem **NP-complete**	*C2.* Reducible to B2, so it is **NP-complete**			
Tree constraints and optimisation	**C4** e.g., Dmax and Steiner tree problem **NP-complete**	C5. e.g. L: B_{min} T:D_{max} T:$Cost_{min}$ **NP-complete**			

2. Multicast Routing Algorithms

- **Simple Examples of Multicast Routing Problems 1/3**
- Network example

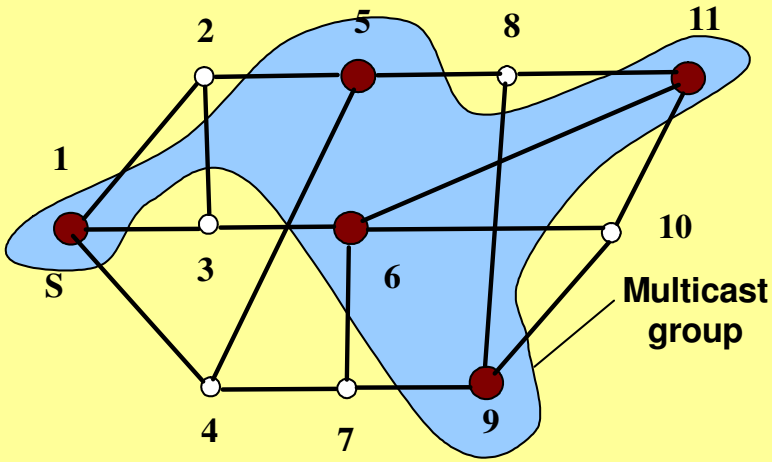


Fig . 2.4-1 Network graph
M= {1,5,6,9,11}
e.g. Cost/link =1

Shortest Path Tree

SPT - minimizes the sum of the weights on the links along each individual path from the source to a receiver in the mc group

Examples

(Link weight =1) ⇒ least-hop tree

(Link weight = delay) ⇒ tree is a least-delay tree

Algorithms: both are exact and run in polynomial time

Bellman-Ford and Dijkstra (most well known)

SP algorithms may solve TC problems (e.g., delay-constrained).

2. Multicast Routing Algorithms

- **Simple Examples of Multicast Routing Problems 1/3**
- (previous slide)
- This slide shows a WAN network example graph
- Each node is a router having local connected LANs.
- The nodes belonging to multicast group are in the set M ={1,5,6,9,11}.
- Mc group is characterised by a common IP group address.
- The source is not mandatory belonging to the multicast group
- The costs of the links in this example are considered equal to unity.

- The Figure 2.4-2 shows a Shortest Path Tree, unidirectional, rooted in Node 1 as a source. The tree can be found for instance by using the Dijkstra algorithm.
- The total cost of the tree is 8.
- If we consider the costs proportional to delay-per-link then this tree offers the best solution for having minimum delay between Source Node 1 and nodes 5,6,9,11 as destinations.
- The average delay value is $D_{Sav} = 2.5$.
- The maximum delay is $D_{max} = 3$

2. Multicast Routing Algorithms

- **Simple Examples of Multicast Routing Problems 2/3**
- Network example

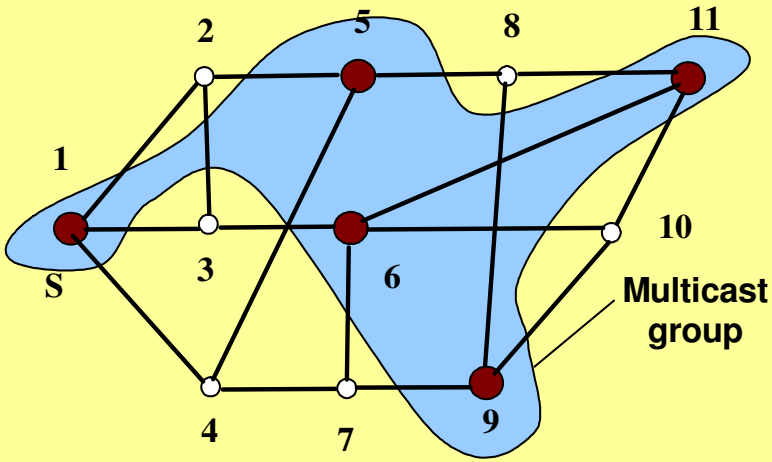


Fig . 2.4-1 Network graph
M= {1,5,6,9,11}
e.g. Cost/link =1

M= {1,5,6,9,11}
Cost/link =1

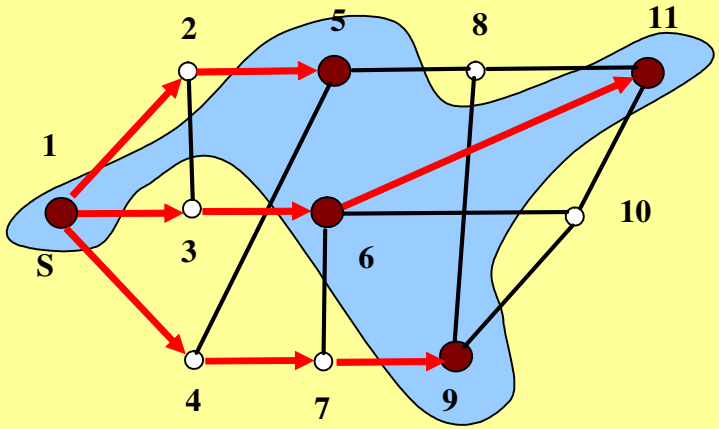


Fig . 2.4-2
Shortest Path Tree (SPT), Source Specific Tree
C_{SPT} = 8, D_{SPT}max = 3, D_{SPT}av =2.67
Note: this is not the optimum tree from point of view total cost!
Find another solution !

2. Multicast Routing Algorithms

- **Simple Examples of Multicast Routing Problems 3/3**

Steiner Tree Problem in Networks (SPN) – tree optimisation

Note that this is an NP-complete problem

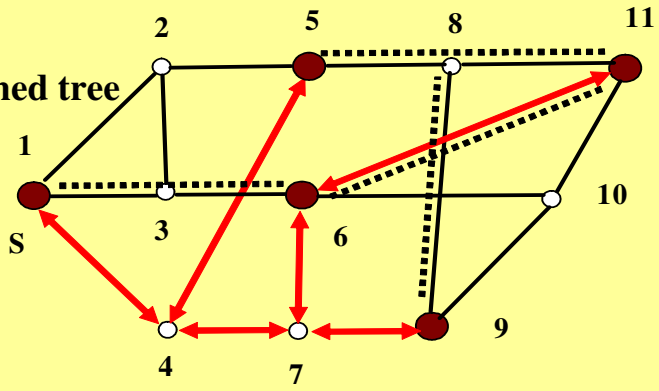
- Given: **G = (V,E)** – undirected
- M = multicast group included in V
- c_{uv} = cost of link (u,v); u,v ∈ V
- **Required: T= (V_T, E_T), which spans M so that** $\sum_{(u,v)\in E_T} c_{uv} = \min imum$

- Steiner nodes = nodes u,v ∈ V_T which do not belong to M
- ST is not unique (—— or .....)

Example

- **M = {1,5,6,9,11}**
- **Steiner nodes= {4,7}- red tree or or {3,8} for dashed tree**
- **C_{ST} = 6- red tree**
- **D_{ST}av = (2 + 3 + 4 + 3)/4 = 12/4 = 3 – red tree**

- **Comparison SPT - ST ——**
  - **C_{SPT} = 8 C_{ST} = 6**
  - **D_{SPT}max = 3 D_{ST}max=4**
  - **D_{SPT}av =2.67 D_{ST}av= 3**



M= {1,5,6,9,11}
Cost/link =1

Group Shared Tree
Example of Steiner Tree

2. Multicast Routing Algorithms

- **Simple Examples of Multicast Routing Problems 3/3**
- **(previous slide)**
- The Steiner Problem in Networks is a classical optimisation problem, for finding a minimum total cost tree (a shared tree) on an undirected graph.

- The SPN is non-polynomial complete problem (NP-complete); this means that there is no possible in general case to find a solution after a polynomial time proportional effort.

- Since the graph G is undirected, it models a bidirectional linked network; therefore the Steiner tree is a group-shared multicast tree – that means that every node in the tree can be source or destination.

- Each link (u,v) is assigned a cost c_{uv}. The Steiner tree has a minimum cost which is the sum of all branches costs.

2. Multicast Routing Algorithms

- **Summary of Multicast Routing Algorithms , [3] :**
- **Algorithms to build the multicast trees**
  - with/without constraints and/or optimisation.
- **see *Wang, Hou-[3],* for a comprehensive comparative presentation studies**
  - **Shortest Path Constrained Tree: *Dijkstra* Centralised**
  - **Minimum Spanning Optimised Tree: *Prim* - Centralised, *Gallager* Distributed**
  - **Optimised Steiner Tree: *KMB Heuristic, Takahashi, Maxemchuk-* Centralised, *Bauer* - Distributed**
  - **Delay Constrained Steiner Tree with tree optimisation:**
    - ***Zhu, Kompella's Haberman, Bauer-* Centralized**
    - ***Kompella , Jia* – Distributed**
  - **Maximum Bandwidth Tree *: Sacham-* Centralised**
  - **Delay-jitter constrained: *Rouskas* - Centralised**
  - **Bandwidth-delay constrained: *Chen* – Distributed**
  - **The list is not exhaustive**

2. Multicast Routing Algorithms

- **Summary of Multicast Routing Algorithms , [3] :**
- **Algorithms to build the multicast trees**
  - They can be differentiated by
    - initiator (source/receiver), complexity, tree type
    - problem solved : tree constraints, bandwidth constraints and/or delay constraints
- Algorithms are incorporated in different protocols depending on their fitness to the protocol design philosophy.
- ***Shortest Path Tree***
- SPT - minimizes the sum of the weights on the links *along each individual path* from the *source to a receiver* in the mc group
  - Examples
    - (Link weight =1) $\Rightarrow$ least-hop tree
    - (Link weight = delay) $\Rightarrow$ tree is a least-delay tree
  - **Algorithms:** both are exact and run in polynomial time
  - Bellman-Ford and Dijkstra (most well known)
  - SP algorithms may solve Tree-Constrained problems (e.g., delay-constrained).

2. Multicast Routing Algorithms

- **Summary of Multicast Routing Algorithms , [3] :**
- ***Minimum Spanning Tree (MSpT)***
  - tree spanning all the group members; minimizes the total tree weight
  - **Algorithms:**
    - **Centralized algorithm :** Prim, [11]
      - the tree T building starts from an arbitrary root node
      - grows until the tree spans all the nodes in the network
      - each step: a least-cost edge linking an off-tree node to the partial tree is added to T
    - The algorithm is greedy: the tree is augmented with an edge that contributes the minimum amount possible to the tree's total cost.
    - **Distributed version** : Gallager et al. [12].
    - MSpT algorithms – can be used for tree optimization problems.

2. Multicast Routing Algorithms

• Summary of Multicast Routing Algorithms

- ***Steiner Tree***
  - Minimize the total multicast tree cost ; NP-complete [13], [14]
  - Mc group includes all nodes in the network ⇒ (Steiner tree = MSpT)
  - Unconstrained Steiner tree algorithms - used for tree optimization problems.
  - **Surveys on exact and heuristic algorithms:**
    - Winter [13] and Hwang [14], Bauer [15] and Salama [16]
  - Example of heuristik: KMB
- ***Constrained Steiner Tree***
  - Constraints examples: delay, delay jitter, or a combination.
  - NP-complete ⇒ Heuristic algorithms exist.
  - Variants:
    - Centralised algorithms, source-initiated. [17], [18]
    - Distributed algorithms, [19], [20]

- ***Heuristic*** *Greek: "Εύρίσκω", **"find"** or **"discover"**) refers to experience-based techniques for problem solving, learning, and discovery. If an exhaustive search is impractical, heuristic methods are used to speed up the process of finding a satisfactory solution. Examples of this method include using a rule of thumb, an educated guess, an intuitive judgment, or common sense.*

2. Multicast Routing Algorithms

- **Summary of Multicast Routing Algorithms**
- **Multicast tree**: source rooted; **Problem solved:** tree optimisation
- **Algorithm type:** centralised/distributed (C/D)
- **Initiator:** source (S) or receiver (R)

	Algorithm	**Type (C/D)**	**Initiator**	**Complexity**
Shortest Path Tree	Dijkstra	C	S	O (|E|logV)
Min Spanning Tree	Prim	C	S	O (|E|logV)
	Gallager	D	R	$\|V\|\log_2\|V\|$ (1)
Steiner Tree	Kou	C	S	$O(\|M\|\|V\|^2)$
	Takahashi	C	S	$O(\|M\|\|V\|^2)$
	Bauer	D	R	$O(D\|M\|)$(2)
	Maxemchuk	C	S	$O(\|M\|\|V\|^2)$

- (1) Message complexity: $O(|V|\log 2|V| + |E|)$.
- (2) D is the one-way trip time over the longest path between two nodes in the network or the diameter of the network. Message complexity: $O(|M||V|)$.

2. Multicast Routing Algorithms

- **Summary of Multicast Routing Algorithms**

	Algorithm	Type (C/D)	Initiator	Complexity	Problem solved
Constrained Steiner Tree	Zhu	C	S	$O(k\|V\|^3\log\|V\|)$	DC-TO
	Kompella	C	S	$O(\|V\|^3\Delta)$ (3)	DC-TO
	Haberman	C	S	$O(k\|M\|\|V\|^4)$ (4)	DC/JC-TO
	Kompella	D	S	$O(\|V\|^3)$ (5)	DC-TO
	Jia	D	S/R	$O(2\|M\|)$ (6)	DC-TO
	Bauer	C	R	$O(\|M\|\|V\|^2)$	DC-TO
Max Bandwidth Tree	Shacaham	C	S	$O(\|E\|\log\|V\|)$	LO
Miscellaneous	Rouskas	C	S	$O(k\|M\|\|V\|^4)$ (7)	BC-DC
	Chen	D	S/R	$O(\|M\|\|E\|)$ (8)	DC-JC

2. Multicast Routing Algorithms

- **Summary of Multicast Routing Algorithms**
- **NOTES**
- (3) Δ is the delay requirement. The time complexity is polynomial if Δ is a bounded integer.
- (4) k is the number of paths in the initial least-cost path tree; l is the number of paths tried when adding a multicast group member.
- (5) Message complexity: $O(|V|3)$.
- (6) Message complexity: $O(2 \cdot |M|)$.
- (7) k and l are constants in the algorithm. A larger k or l results in a higher probability of finding a feasible tree and a higher overhead.
- (8) Message complexity: $O(|E|)$.

2. MulticastRouting Algorithms

- **References for Multicast Routing Algorithms**
- [1] Aaron Striegel, G. Manimaran, “A Survey of QoS Multicasting Issues” IEEE Communications Magazine, June 2002, pp. 82-87.
- [2] A. Striegel, G. Manimaran “A Scalable Approach for DiffServ Multicasting”, …
- [3] Bin Wang and Jennifer C. Hou, “Multicast Routing and Its QoS Extension: Problems, Algorithms, and Protocols”, IEEE Networking Magazine, 2000.
- [4] Jinquan Dai, Hung Keng Pung and Touchai Angchuan, “QROUTE: An Integrated Framework for QoS-Guaranteed Multicast”, Proceedings 27th Conference on Local Computer Networks, Tampa, Florida, 6-8 Nov 2002, pg 90-99
- [4] Jinquan Dai, Hung Keng Pung and Touchai Angchuan, “QROUTE: An Integrated Framework for QoS-Guaranteed Multicast”, Proceedings 27th Conference on Local Computer Networks, Tampa, Florida, 6-8 Nov 2002, pg 90-99 .
- [5] C. Donahoo, H.Zegura, “Core Migration for Dynamic”, Multicast Routing,” *Proc. ICCCN*, 1995.
- [6] R. Sriram, G. Marimaran, and C. Siva Ram Murthy, “Preferred Link-Based Delay-Constrained Least Cost Routing in Wide Area Networks,” Comp. Commun., vol. 21, no. 18, 1998, pp. 1655–69.
- [7] S. Chen, K. Nahrstedt, and Y. Shavitt, “A QoS-Aware Multicast Routing Protocol,” IEEE INFOCOM, 2000, pp. 1594–1603.
- [8] G. Manimaran, H. Shankar Rahul, and C. Siva Ram Murthy, “A New Distributed Route Selection Approach for Channel Establishment in Real-Time Networks,” IEEE/ACM Trans. Net., vol. 7, no. 5, Oct. 1999, pp. 698–709.
- [9] M. Faloutsos, A. Banerrjea, and R.Pankaj. QoSMIC: Quality of Service Sensitive Multicast Internet Protocol. In ACM SIG-COMM, Sep. 1998.
- [10] Z. Wang and J. Crowcroft, “QoS Routing for Supporting Resource Reservation,” *IEEE JSAC*, Sept. 1996.

2. MulticastRouting Algorithms

- **References for Multicast Routing Algorithms**

- [11] T. H. Cormen, C. E. Leiserson, and R. L. Rivest, *Introduction to Algorithms*, MIT Press, 1997.
- [12] R. Gallager, P. Humblet, and P. Spira, "A Distributed Algorithm for Minimum-Weight spanning trees," *ACM Trans. Programming Lang. and Sys.*, Jan. 1983, pp. 66–77.
- [13] P. Winter, "Steiner Problem in Networks: A Survey," *Networks*, 1987, pp.129–67.
- [14] F. K. Hwang, Steiner Tree Problems, *Networks*, 1992, pp. 55–89.
- [15] F. Bauer, "Multicast Routing in Point-to-Point Networks Under Constraints," Ph.D. dissertation, UC Santa Cruz, 1996.
- [16] H. F. Salama, "Multicast Routing for Real-Time Communication on High Speed Networks," Ph.D. dissertation, NC State Univ., Raleigh, 1996.
- [17] Q. Zhu, M. Parsa, and J. Garcia-Luna-Aceves, "A Source-Based Algorithm for Delay-Constrained Minimal-Cost Multicasting," *Proc. IEEE INFOCOM '95*, 1995, pp. 377–84
- [18] V. P. Kompella, J. C. Pasquale, and G. C. Polyzo, "Multicast Routing for Multimedia Communication," *IEEE/ACM Trans. Net.*, 1993, pp. 286–92.
- [19] V. P. Kompella, J. C. Pasquale, and G. C. Polyzo, "Two Distributed Algorithms for Multicasting Multimedia Information," *Proc. ICCCN '93*, 1993, pp. 343–49.
- [20] X. Jia, "A Distributed Algorithm of Delay-Bounded Multicast Routing for Multimedia Applications in Wide Area Networks, *IEEE/ACM Trans. Net.*,
- Dec. 1998, pp. 828–37

2. IP Multicast Routing Algorithms

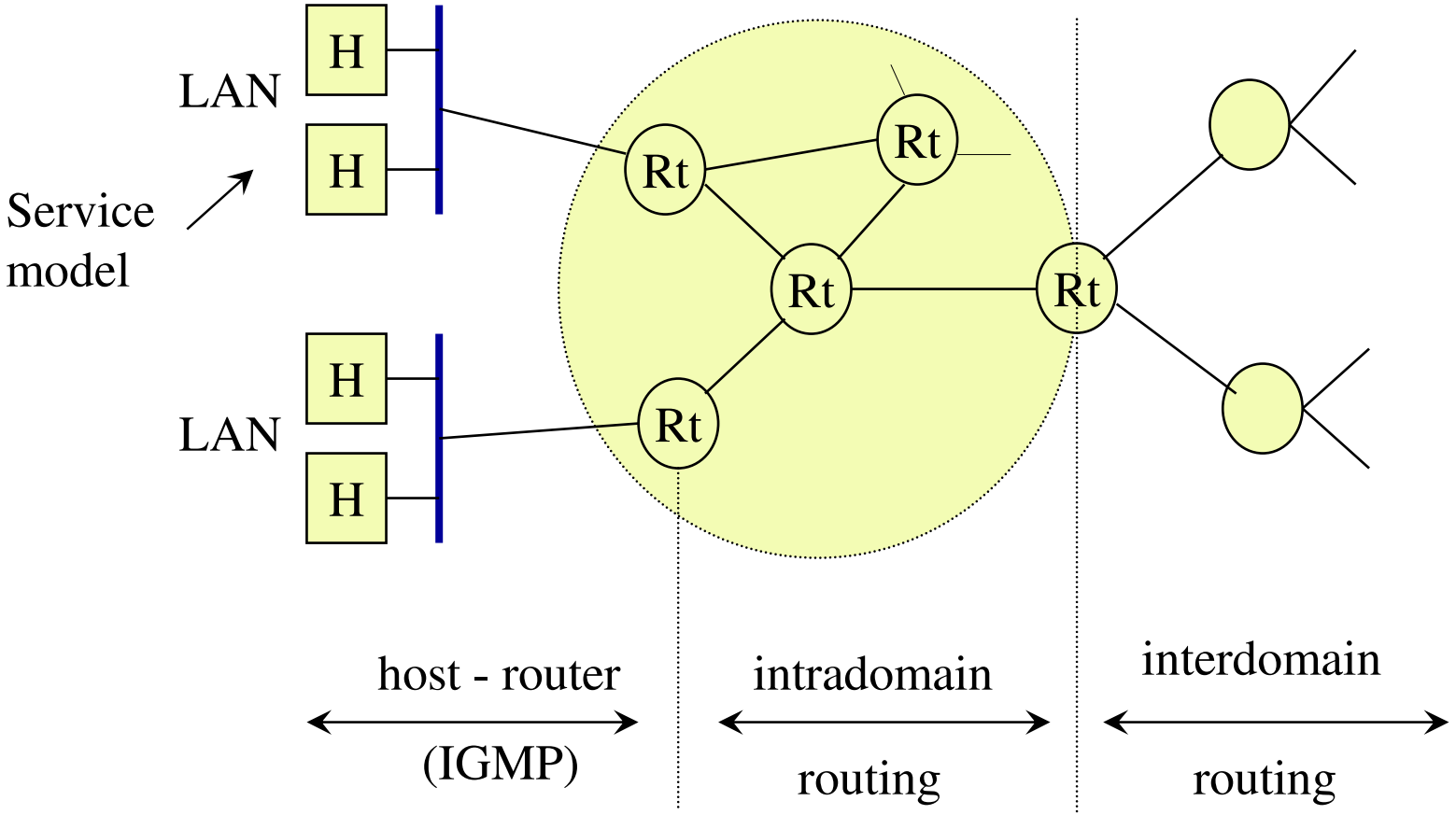
- **References for MulticastRouting Algorithms**
- [21] B. K. Haberman and G. Rouskas, “Cost, Delay, and Delay Variation Conscious Multicast Routing,” Tech. rep. TR-97-03, NC State Univ., 1997.
- [22] F. Bauer and A. Verma, “Degree-Constrained Multicasting in Point-to-Point, Networks,” *Proc. IEEE INFOCOM ’95*, 1995.
- [23] N. Shacham, “Multipoint Communication by Hierarchically Encoded Data,” *Proc. IEEE INFOCOM ’92*, May 1992, pp. 2107–14.
- [24] R. N. Rouskas and I. Baldine, “Multicast Routing with End-to-End Delay and Delay Variation Constraints,” *IEEE JSAC*, Apr. 1997, pp. 346–56.
- [25] S. Chen and K. Nahrstedt, “Distributed QoS Routing in High-Speed Networks Based on Selective Probing,” Tech. rep., CSD, UIUC, 1998.
- [26] M. Imase and B. Waxman, “Dynamic Steiner Tree Problem,” *SIAM J. Disc. Math*, Aug. 1991, pp. 369–84.
- [27] F. Bauer and A. Varma, “ARIES: A Rearrangeable Inexpensive Edge-Based On-Line Steiner Algorithm,” *IEEE JSAC*, Apr. 1997, pp 382–97.
- [28] P. Narvaez, K. Siu, and H. Tzeng, “New Dynamic SPT Algorithm Based on a Ball-and-String Model,” *IEEE INFOCOM ’99*, New York, Mar. 1999.

Contents

- 1. *Multicast General Overview*
- 2. *Multicast Routing Algorithms*
- 3. IP Level Intradomain Multicast
- 4. IP Multicast Networks (1)
- 5. Inter-domain IP level multicast protocols
- 6. Multicast Transport Protocols
- 7. Overlay Multicast
- 8. Multicast in wireless and mobile environment
- 9. Security issues in multicast communication
- 10. Multicast based applications and services
- 11. Open issues in IP networks multicast

3. IP Level Intra-domain Multicast

- **Components of IP multicast architecture**



3. IP Level Intra-domain Multicast

- **IP Multicast Service Model (RFC 1112)**
- **Multicast groups**
  - there is a range of group addresses (Class D in IPv.4)
  - each group is identified by an IP address (belonging to class D)
  - (∀) number of members in a group
  - (∀) location of a member
  - dynamic groups (join, leave of members at any time- without negotiating this with a centralised group management entity)
  - a group can have one or more sources
  - in principle any host may be sender/receiver
  - applications using IP mc are working on UDP and not TCP (best effort !)
  - IP mc model = "Any source multicast" – the source is not mandatory belonging to the group

3. IP Level Intra-domain Multicast

- **IP Multicast Service Model (RFC 1112)(cont'd)**
- Receivers (Rec)
  - must register to a group address
  - can make join/leave actions
- Senders (sources) (S)
  - may not be members of group, can also be receivers
  - open groups: S sends without knowing the members
  - sending application has to :
    - specify outgoing network I/F, TTL
    - enable / disable loopback (if S is also R)
- Conclusions on initial IP multicast model:
  - No business model of IP multicast to control the group
  - Problems with group address assignment
  - Need further development or solve the group management at higher layers

3. IP Level Intra-domain Multicast

- **IP Multicast Addressing Issues**
- **IP Multicast Group Addresses (IPv4)**
  - ***224.0.0.0–239.255.255.255***
  - Class "D" Address Space
    - High order bits of 1st Octet = "1110"
- **Reserved Link-local Addresses**
  - ***224.0.0.0–224.0.0.255, sent with TTL = 1***
  - Examples:
    - 224.0.0.1 All systems on this subnet
    - 224.0.0.2 All routers on this subnet
    - 224.0.0.4 DVMRP routers
    - 224.0.0.5 OSPF routers
    - 224.0.0.13 PIMv2 Routers

3. IP Level Intra-domain Multicast

- **IP Multicast Addressing Issues**
- Administratively Scoped Addresses
  - ***239.0.0.0–239.255.255.255***
  - Private address space
    - Similar to RFC1918 unicast addresses
    - Not used for global Internet traffic
    - 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

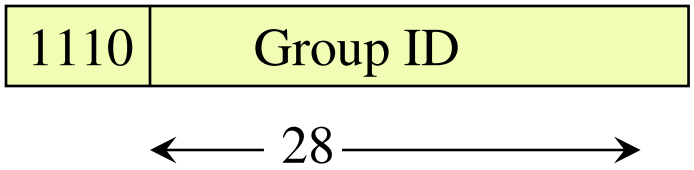
3. IP Level Intra-domain Multicast

- **IP Multicast Addressing Issues**
- **Dynamic assignment** of Group Address
  - *sdr* application used historically to get addresses
  - Sessions/groups announced over well-known multicast groups
  - Address collisions detected and resolved at session creation time
  - Scalability problems
- **Present dynamic techniques**
  - **Multicast Address Set-Claim (MASC)**
    - Hierarchical, dynamic address allocation scheme
    - Extremely complex garbage-collection problem.
    - Long ways off
  - **MADCAP**
    - Similar to DHCP, need application and host stack support
- **Static Group Address Assignment**
  - Temporary method to meet immediate needs
  - Group range: 233.0.0.0 - 233.255.255.255
    - Middle two octets gets the AS number
    - Remaining low-order octet used for group assignment
  - Defined in IETF draft
    - "draft-ietf-mboned-glop-addressing-00.txt"

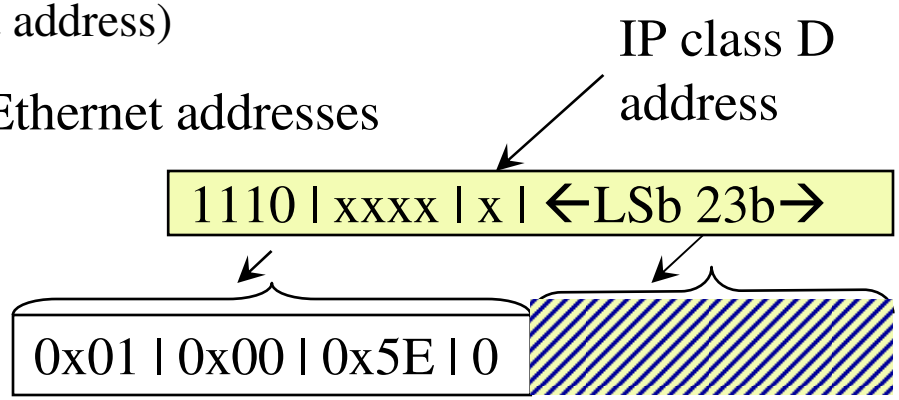
3. IP Level Intra-domain Multicast

- **Data Link and IP group addressing**
- Class D IP v.4 address – designed for multicast

224.0.0.0 – 239.255.255.255



- A host R_i can join group G_j, and later leave G_j
- Data Link layer- multicast transfer solutions:
  - 1. broadcast Dst = 0x ff ff ff ff ff ff
    - filtering all at IP layer
  - 2. multicast (on DL multicast address)
- Mapping IP mc addresses to Ethernet addresses



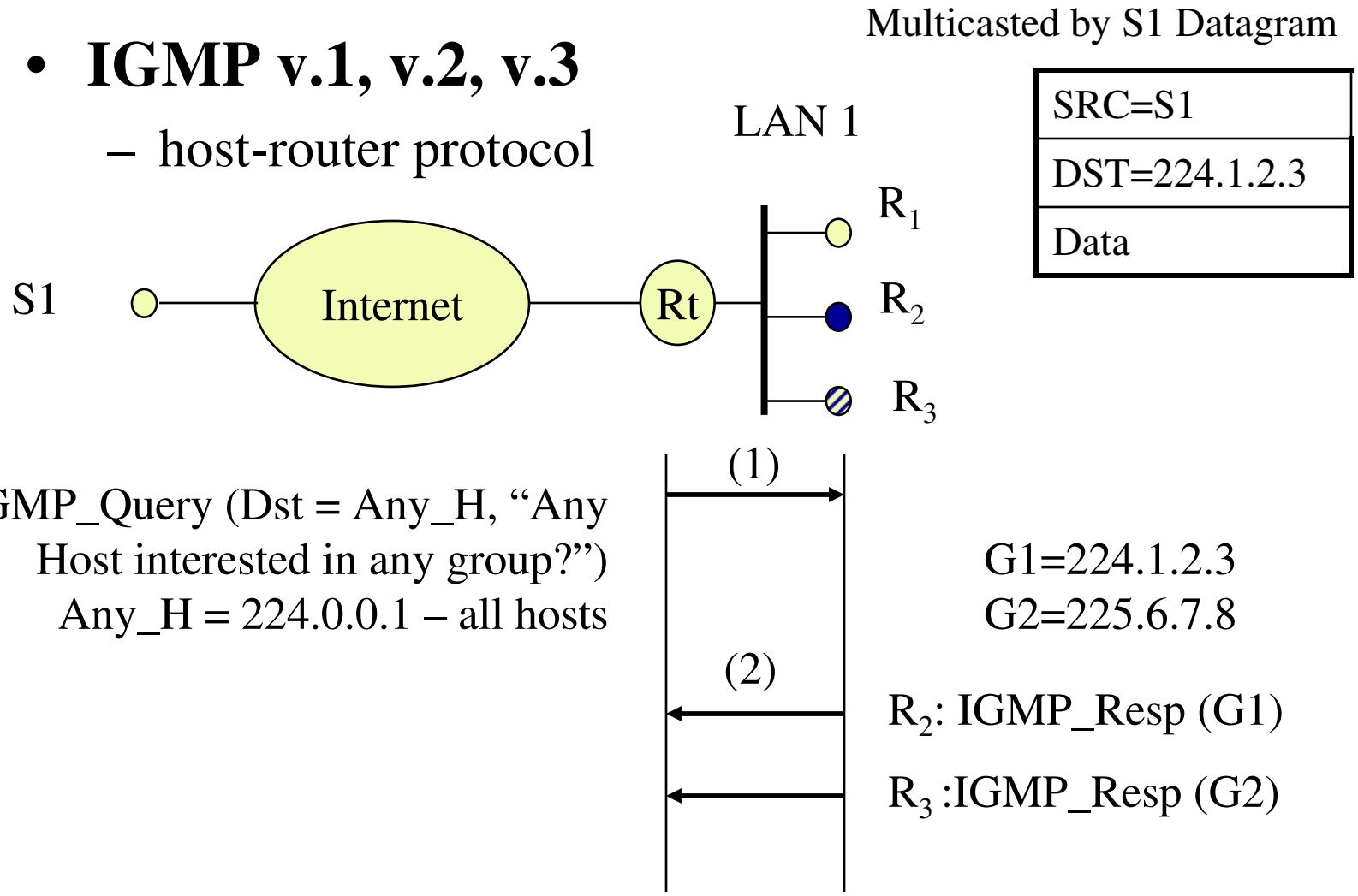
3. IP Level Intra-domain Multicast

- **Data Link and IP group addressing (previous slide)**
- In the set of IP v.4 addresses a range is reserved for multicasting, that is 224.0.0.0 – 239.255.255.255. Each address is associated to multicast group.
- A host R_i can join group G_j, and later leave G_j.

- At data link layer (on shared medium) the multicast transfer can be done in two manners:
  - - by using the broadcast address Dst = 0x ff ff ff ff ff ff. In this case the filtering of packets is done at IP layer. This solution is processing power consuming because the DL layers of all hosts on a LAN are crossed by the mc packets that have to be filtered at IP layer and dropped in the hosts which are not destinations
  - by using a multicast address at DL layer. In this case the filtering of not-desired frames is performed at DL layer, avoiding overloading the IP level.
  - The figure presents the mapping between an Ethernet group address to an IP level group address. Note that the mapping is not 1-to-1, therefore several IP group addresses are mapped on the same DL group address.
  - In fact we have : 32-to-1 mapping (not 1-to-1!), hence not complete filtering at DL layer.

3. IP Level Intra-domain Multicast

- **IGMP v.1, v.2, v.3**
  - host-router protocol



3. IP Level Intra-domain Multicast

- **IGMP v.1, v.2, v.3**
- Internet Group Management Protocol (IGMP) is a host-router protocol used by the hosts to subscribe to or leave fromn a certain group
- It has been specified in three versions v.1, v.2, v.3.
- on each link – one router is elected as querrier
- it periodical sends IGMP_Querry to all systems (224.0.0.1 is a broadcast address mening all_hosts)
- on reception, each R_i set a random timer (0 – 10s)
- it responds at time – out returning a membership report to group G, with TTL = 1)
- suppose that R_i and R_k are both interested in group G_x and R_k response to Rt querry is issued earlier than response of R_i
- if R_j observes an IGMP_Resp$_k$ (G_x) then it suppresses its own response, because is redundant (RT has already registered that it has some members of the group Gx on LAN1
- Rt time – outs non responding groups
- if Rt will receive in a future a datagram destined to G_x group then it will distribute this datagram to LAN1 (as long as at least one member of G_x exists in LAN1)
-

3. IP Level Intra-domain Multicast

- **IGMP v.1, v.2, v.3**
- **Join to a group**
  - Router sends periodic Queries to 224.0.0.1 (all hosts on the subnet)
  - One member per group per subnet reports
  - Other members suppress reports (normally one report message per group present – for one query)
  - usual query time period : 60 – 90sec
  - to decrease latency : at first join of a host, it sends one report (not wait for a query)
- **Leave from group (IGMP1)**
  - Host leaves group without announcing it
  - Router sends several General Queries (e.g. 3 times, 60 secs apart)
  - If no more existing group members then no more IGMP report for the group is received
  - Group times out (Worst case delay ~= 3 minutes)
- **IGMP v.2**
  - a host does ***explicitly*** inform its router when it leaves a group (reduce leave latency)
  - (in Version 1, a receiver wanting to "leave", will stop responding to queries)
  - standard querier election method
  - currently – is the most used standard
  - widely implemented

3. IP Level Intra-domain Multicast

- **IGMP v.3**
  - RFC 3376/2002:enable filtering in hosts :
    - selection of senders to listen to
    - all senders but not a specific set
  - additional protocol to inform a source that no one is listening
  - backward compatibility IGMP v.3 => IGMP v.2, IGMP v.1
  - IGMP v.1: RFC 1112
  - IGMP v.2: RFC 2236
  - IGMP v.3: RFC 3376
    - Version 3 adds support for "source filtering“: a system may report interest in receiving packets
      - *only* from specific source addresses, as required to support Source-Specific Multicast [SSM],
      - or from *all but* specific source addresses, sent to a particular multicast address
    - **IGMP snooping**

3. IP Level Intra-domain Multicast

- **IGMP snooping**
- By using this technique IGMP v.3 avoids broadcast in LANs
- The bridges / switches look inside received multicast frames to detect :
    - IGMP Responses to learn directions in which hosts belonging to group reside
    - IGMP Querries, DVMRP probes, MOSPF Hellos, PIM Hellos etc – to learn directions in which multicast routers reside
- The bridges / switches multicast data packets only to necessary directions
- The SNOOPing has some problems
  - does not work for non IP multicast
  - stops working if new protocol is deployed
  - Contribute to lowering the performance because one have to look inside every multicast frame

3. IP Level Intra-domain Multicast

- **IP Level Multicast Routing Protocols presentation**
  - Dense Mode Protocols (DMP)
  - Sparse Mode Protocols (SMP)
- **Differences in these protocols :**
  - mainly in the type of mc routing trees they build
- DVMRP, MOSPF, PIM –DM build mc spanning trees that are SPT from each source
- PIM-SM, CBT, OCBT, HIP
  - build mc spanning trees that are SP from a known central core (*rendezvous point* - RP),
  - where all sources in the session share the same spanning tree.
  - (additionally PIM-SM can build optionally a source-rooted SPT
- PIM SM - *unidirectional shared* tree
  - packets are sent first to the core,
  - core then sends packets down the multicast spanning tree to all participants of the session.
- CBT, OCBT, BGMP, HIP build *bidirectional shared* trees: packets from each source are disseminated along the tree starting from any point

3. IP Level Intra-domain Multicast

- **IP Level Multicast Routing Protocols presentation**
  - Dense Mode Protocols (DMP)
  - Sparse Mode Protocols (SMP)
- **Intradomain protocols**
  - **DVMRP** - Distance Vector Multicast Routing Protocol
    - » based on unicast RIP, builds own routing table
    - » unidirectional shortest path tree (SPT) – per source
    - » broadcast and prune approach, dense mode
  - **MOSPF** – Multicast Open Shortest Path First
    - » link state based, uses SPT – per source
    - » broadcast membership

3. IP Level Intra-domain Multicast

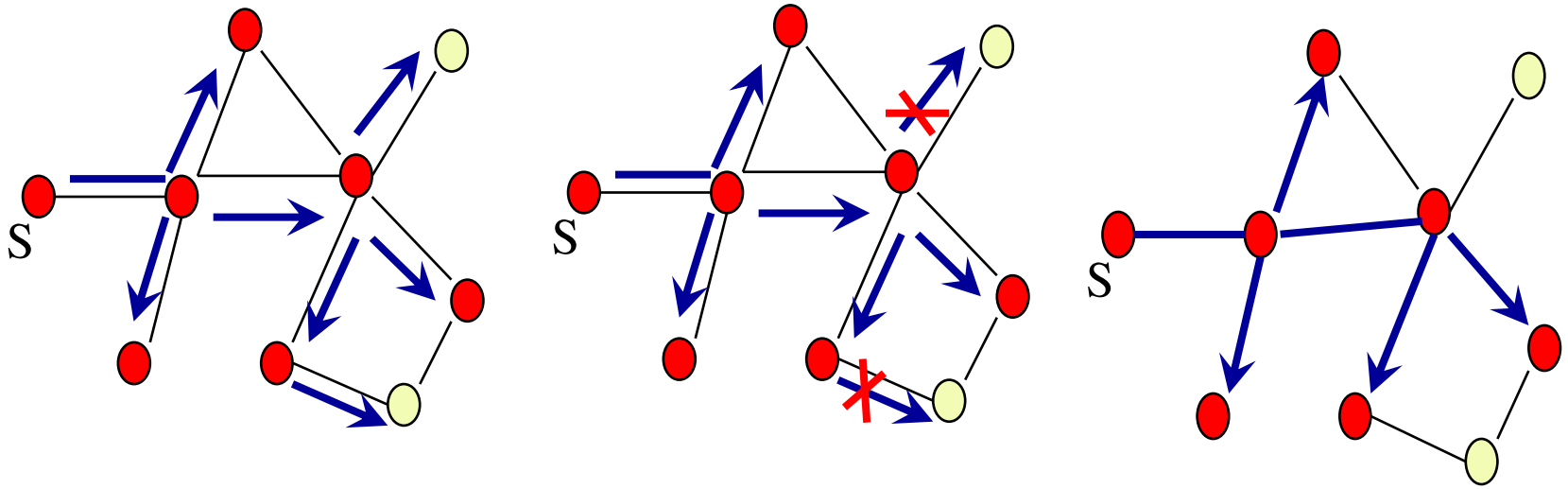
- **IP Level Multicast Routing Protocols Presentation**
- **PIM - Protocol Independent Multicast**
  - ***can operate on any unicast routing protocol***
  - Dense Mode / Sparse Mode (DM / SM) variants
  - uses a center based tree or SPT
  - **– PIM - DM (similar to DVMRP)**
    - broadcast tree and then prune some branches
    - shortest path tree – unidirectional, per source
    - uses unicast routing table
  - **– PIM – SM**
    - uses explicit join of members (receivers) to setup a shared tree rooted in a centre named ***Rendez-Vous Point (RP)***
    - source sends to Rendez – Vous Point (RP)
    - RP distribute packets on a shared unidirectional tree
    - can switch on a unidirectional per source tree
    - uses unicast routing table

3. IP Level Intra-domain Multicast

- **IP Level Multicast Routing Protocols Presentation**
  - **CBT - Core Based Tree**
    - uses a single bidirectional shared tree for a group
    - scalable, efficient bandwidth utilization
  - **OCBT - Ordered CBT**
- **Inter-domain Protocols**
  - *Currently developed (1999) Near-Term solutions*
    - MBGP- Multi-protocol Extension of BGP4
    - PIM-SM used as inter-domain multicast routing protocol
    - MSDP – Multicast Source Discovery Protocol
  - *Long-Term Proposal – for Internet-wide inter-domain multicast*
    - BGMP – Border Gateway Multicast Protocol
    - MASC – Multicast Address Set Claim
  - **BGMP - Border Gateway Multicast Protocol**
    - Exchange multicast reachability between Autonomous Systems (AS)
    - Uses center – based tree – bidirectional
    - Uses TCP as transport protocol
  - **SSM –Source Specific Multicast-** simplifies inter-domain multicast

3. IP Level Intra-domain Multicast

- **Tree types used in mc routing protocols**
- Multicast tree building – approaches
  - a. Broadcast & Prune & Graft- useful for "dense mode"
    - By default all nodes are considered to be interested in multicast distribution
    - Those which are not, should announce this by "prune " messages
    - Time out for "prune" states ⇒ the mc tree is automatically extended over nodes that are no longer sending "prune"

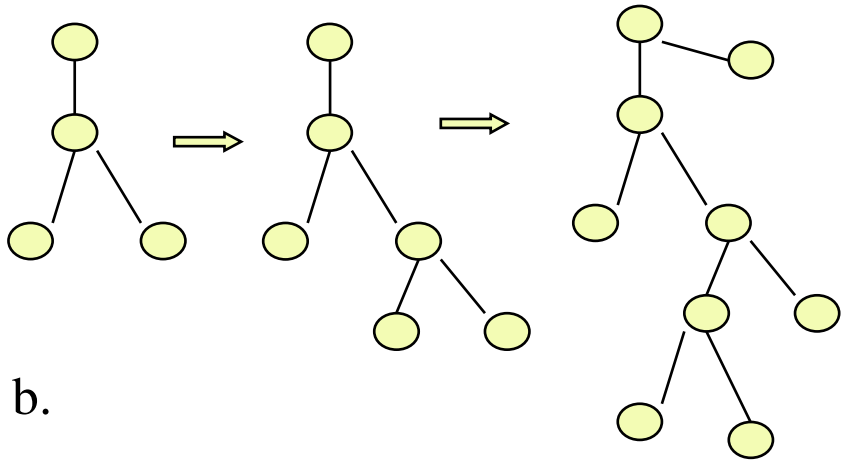


3. IP Level Intra-domain Multicast

- **Tree types used in mc routing protocols**
- Multicast tree building – approaches
  - a. Broadcast & Prune & Graft- useful for "dense mode"
- We sumarise the two basic methods to build a multicast tree:
- - build a broadcast tree and then prune undesired branches
- - explicit join of routers having group members to a centered tree.
- The figure in the previous slide illustrates the first method. After building the broadcast tree (by using for instance the RPF algorithm) this is pruned by removing some branches.
- A graft operation is used by some protocols to add new branches to the tree

3. IP Level Intra-domain Multicast

- **Tree types used in mc routing protocols**
- Multicast tree building – approach- useful in sparse mode
  - b. Explicit Join / Leave (increase the tree - at request - with new branches)
    - By default no node is supposed to be interested in mc distribution
    - Those interested should make a Join request and refresh this requests
    - The tree is constructed by adding necessary branches only
    - The tree is reduced automatically for those branches whose nodes no longer make refresh of join
    - Special mechanisms are required to restore the tree in case of leave or failures.



b.

3. IP Level Intra-domain Multicast

- **Distance Vector Multicast Routing Protocol (DVMRP)**
- old : RFC 1075, 1988
- extension of RIP
- dense mode protocol
- based on SPT rooted at source
- broadcast (by flooding) to build Reverse SPT (RSPT)
- adjust the tree by pruning
- unicast tunneling technique to cross the non DVMRP capable routers
- widespread use on Internet, (Mbone- after 1990- no longer in use))
- DVMRP uses RPM (Reverse Path Multicasting) – simple broadcast protocol
    - Shortest Path Tree to the source
    - Assigns each communication link a metric and a threshold

3. IP Level Intra-domain Multicast

- **Distance Vector Multicast Routing Protocol (DVMRP) (previous slide)**
  - *Threshold* = minimum TTL (Time to live) a multicast packet needs to be forwarded onto a given link.
    - Example:
    - TTL threshold Scope
    - 0 Restricted to the same host
    - 1 Restricted to the same sub-network
    - 15 Restricted to the same site
    - 32 Restricted to the same region
    - 127 Worldwide
    - 255 Unrestricted

3. IP Level Intra-domain Multicast

- **DVMRP- principles**

- Simple **broadcast** tree algorithm: ***Reverse Path Forwarding (RPF)***
- Assumption: links on the graph are bi-directional
- Configuration:
  - One source (S), all other nodes-receivers, **each node knows the shortest path to S**
- A simple protocol ***Reverse Path Multicast (RPM)*** can be based on RPF
- (RPM) = SPT having S as root and uses flood packets F sent by S:
  - ***if (R_j receives F on I / F_k) $\cap$ (I / $F_k \in$ SPT – up-tree) then R_j retransmits F on all other I/F_s => broadcast***
    - ***Otherwise: drop***
  - ***Multiple copies of the same packet can be sent over a link***

- RPF Problem: not all nodes are interested in multicast=> pruning of the tree will be added

3. IP Level Intra-domain Multicast

- **DVMRP- principles – details**
- *Reverse Path Forwarding (RPF).*
- Let we have in a graph one source (S) and all other nodes are receivers. Every node knows the shortest path to S- for instance after running a unicast SPT search algorithm.

- A simple protocol named *Reverse Path Multicast (RPM)* can be based on RPF.
- (RPM) is based on SPT having S as root and uses flood packets F sent by S:
  - if R_j *receives F on* I/F_k and I/F_k *belongs to SPT* – on the up-tree towards S then R_j retransmits F on all other interfaces that is a broadcast. Note that Multiple copies of the same packet can be sent over a link
- The RPF main problem is that not all nodes are interested in multicast. Therefore some pruning of the tree will be added to this algorithm to cut the unused branches.

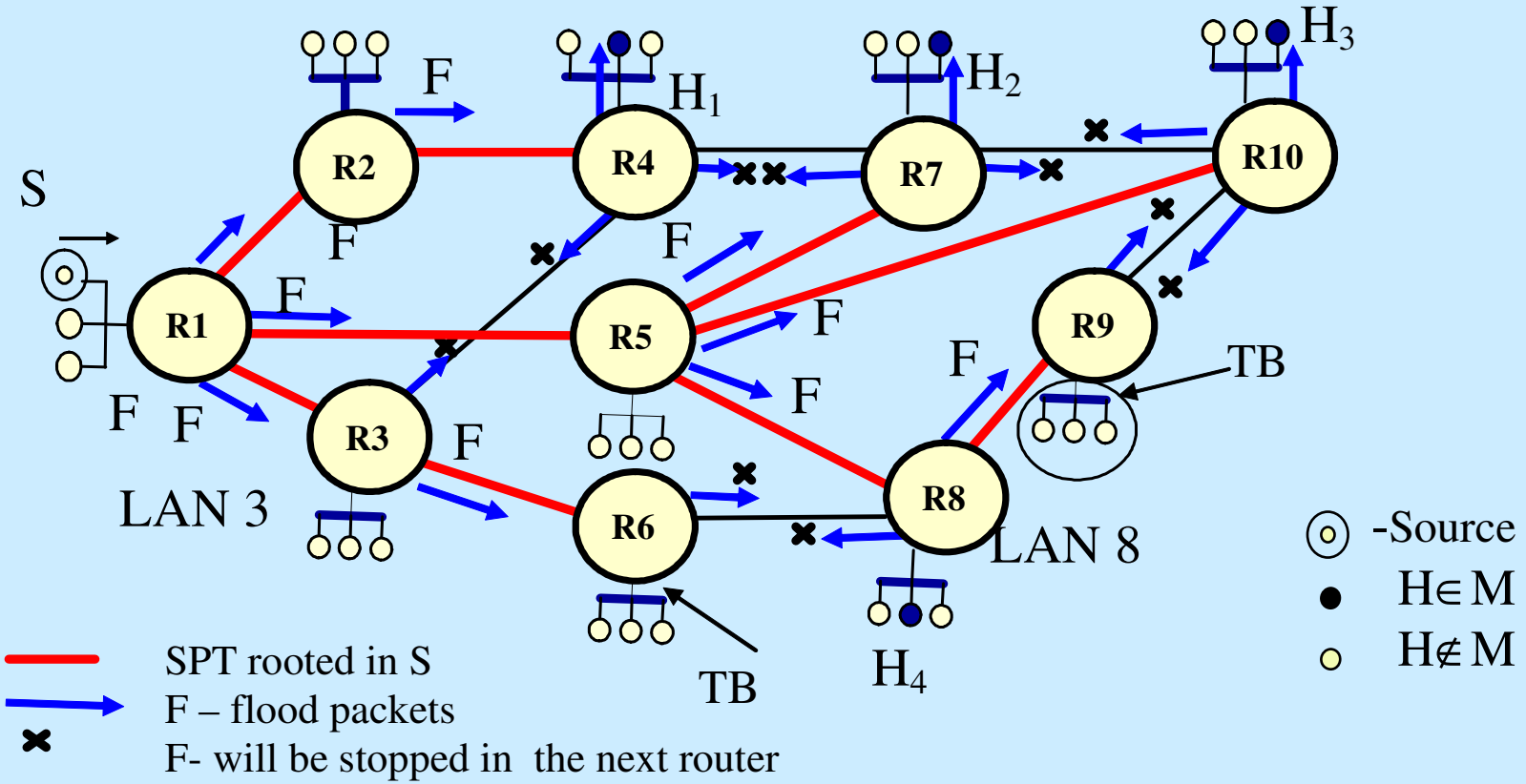
3. IP Level Intra-domain Multicast

- **DVMRP- principles (cont'd)**
  - It uses RPF
  - The RPF algorithm takes advantage of the existing unicast routing table to look up routing state information and perform the following tasks:
  - When a multicast packet is received, save the source's address *S* and the incoming interface identifier *I*.
    - If *I* is the interface used to forward a unicast packet back to the source *S* (RPF check), then:
      - Forward the packet on all interfaces except *I*.
      - Else, the packet is discarded.
  - DVMRP guarantees the *minimum length path* end-to-end delivery, since the packets follow the shortest path from source to destination
  - Furthermore, the RPF algorithm is *robust* regarding routing loops.
    - However, transient loops can still occur during unicast routing table updates.

3. IP Level Intra-domain Multicast

- **DVMRP- principles**

  Reverse Path Forwarding (RPF) + Truncated Broadcast (TB): Example



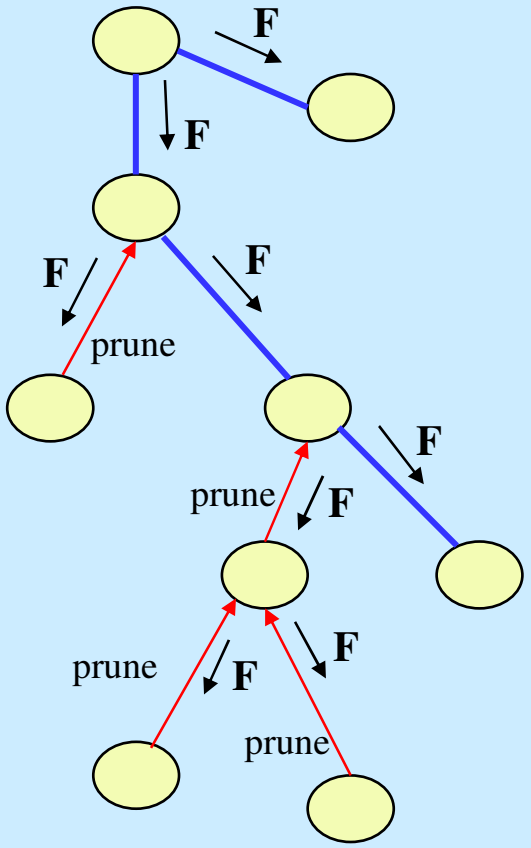
3. IP Level Intra-domain Multicast

- **DVMRP- principles**

  Reverse Path Forwarding (RPF) + Truncated Broadcast (TB): Example (previous slide)
- The figure shows an example of Reverse Path Forwarding principle.
- The members of the group M={ H1, H2, H3, H4} are connected to Routers R4, R7, R8 and R10.
- The flood packets F are issued by S. Every router Ri receiving an F packet analyses if the incoming interface is on the up-tree (SPT) towards the source and:
- if yes, then Ri forward the F packet on all other own interfaces
- - else the F packet will be discarded (please note the * mark on the figure)
- The red lines mark the final Shortest Path Broadcast tree rooted in the Source S.
- Note that some leaf routers do not have members on their local connected LANs (e.g. R6, R9)- information got via IGMP. Therefore these routers do not broadcast the F packets on their subnets. This form of broadcast is named Truncated Broadcast (TB).
- TB reduce the traffic in the leaf subnets but not in the core network. A further action – of reducing the tree branches- named *pruning*

3. IP Level Intra-domain Multicast

- **DVMRP principles (cont'd)**
- RPM uses a modified form of Reverse Path Forwarding (RPF):
  - the set of routers and the corresponding child links = spanning tree (RSPT = Reverse Shortest Path Spanning Tree). RSPT - broadcast tree
  - each mc router knows (via IGMP or statically), if its sub-networks have or not members of the group
  - **a leaf mc router not having members on its subnet sends a *prune* message to its parent**
  - **additionally** a leaf router can send prune on ALL its interfaces, except for the one situated on RSPT to S
  - When applicable, a flag is set for interface *I* indicating that the interface has been pruned (prune state).
  - DVMRP uses no special control messages to advertise the source, but its identity is obtained when receiving the first flooded data packet.
  - *Security* aspects (e.g., which source is entitled to send to which receivers) and constrained (QoS) and policy routing have not been foreseen for DVMRP.

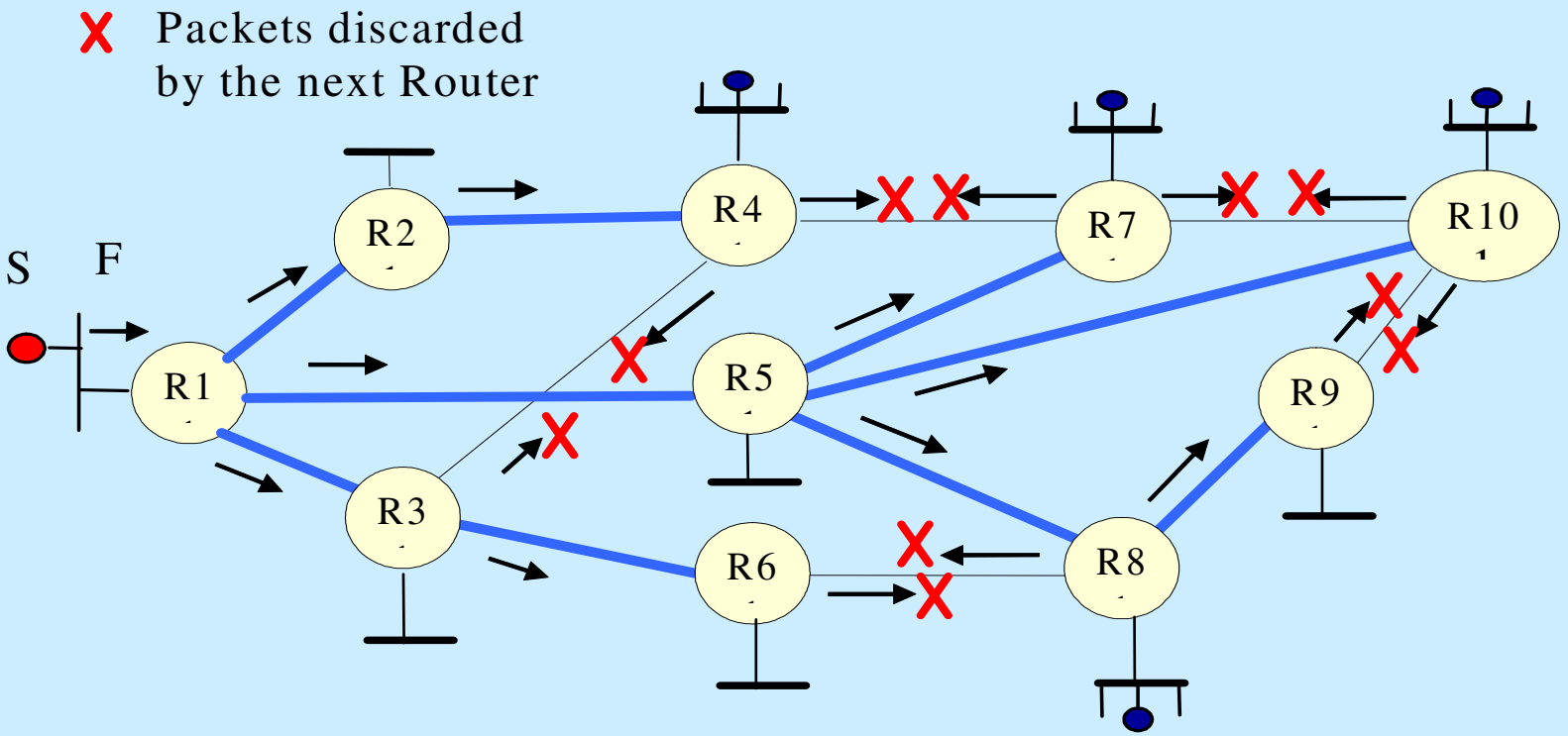


3. IP Level Intra-domain Multicast

- **DVMRP principles (cont'd)**
- Modified RPM steps:
  - a multicast F packet is sent by the source S
  - a router R_k- receives F and makes an RPF check
  - router R_k – broadcast F on all other interfaces including its own sub-nets if it has members there (this means TB mode)
  - a leaf router R_j (on a specific tree) not having members on its subnet sends *prune* to its father
  - a leaf R can send prune on all its interfaces, except for the one on RSPT to S
  - when an intermediate router gets *prune* messages through all the outgoing interfaces then it sends *prune* message upward to its father node
  - ***the result of flood/prune is the final multicast tree***
  - **the flood/prune process has to be periodic (soft state of routers)**
  - if a new member appear a *cancellation_of_prune* message is sent by the respective router
  - aging the prune messages is used: increasing age at each router + discard if age > limit

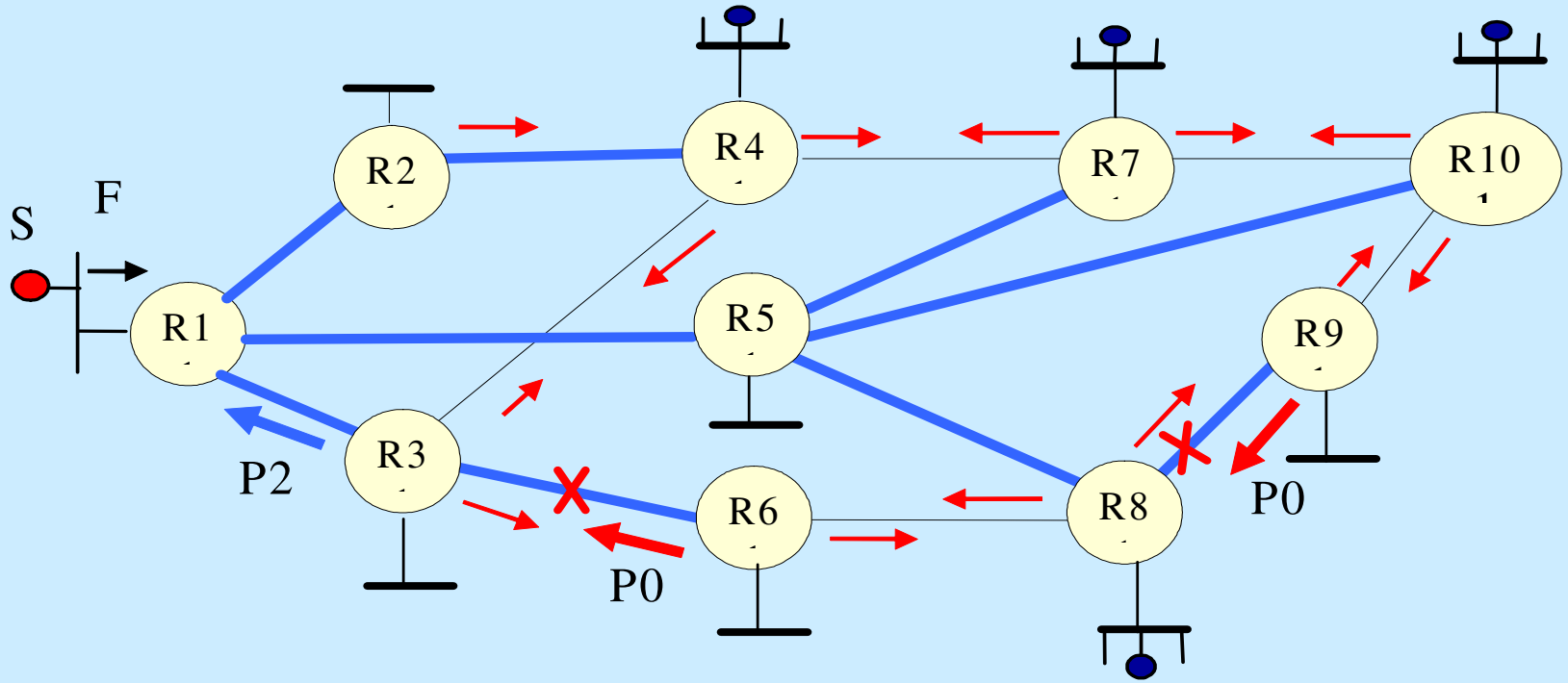
3. IP Level Intra-domain Multicast

- **DVMRP –Example: Flooding**
- The source issues the Flood packets F.
- All Rs make the RPF check and broadcast the F packets. They ignore the F packets that do not come fro the SPT interface towards S.
- The states (S,G) are setup in routers.
- The broadcast tree is marked in blue line



3. IP Level Intra-domain Multicast

- **DVMRP –Example: prune messages**



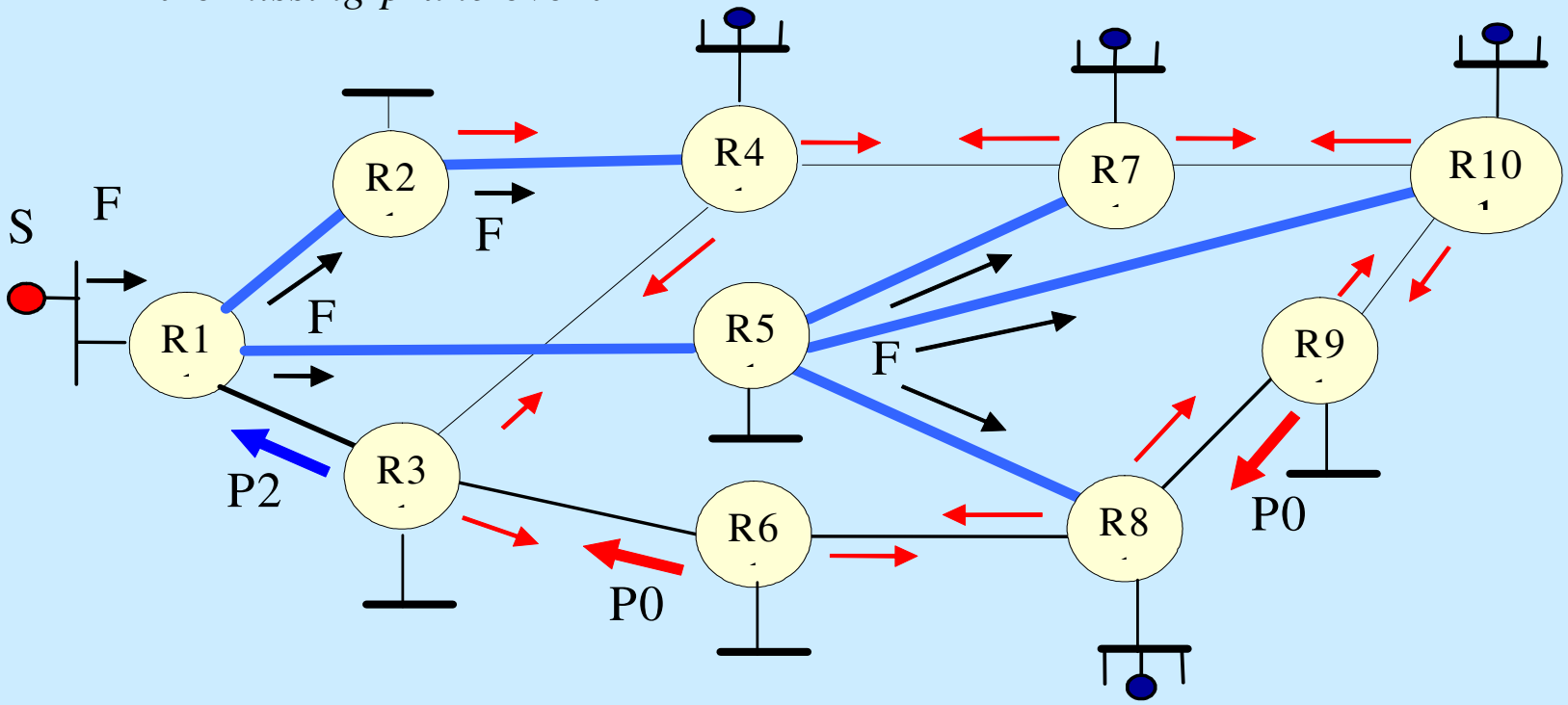
P0 (S, G) – *prune* sent up-stream by a leaf R (has no members on its subnet)

P1(S,G) - *prune* sent to I/Fs different from that ∈ uptree to the S

P2(S,G) - *prune* sent up-stream by a router which receives *prune* from all its downstream interfaces (it has no local members neither should be transit router)

3. IP Level Intra-domain Multicast

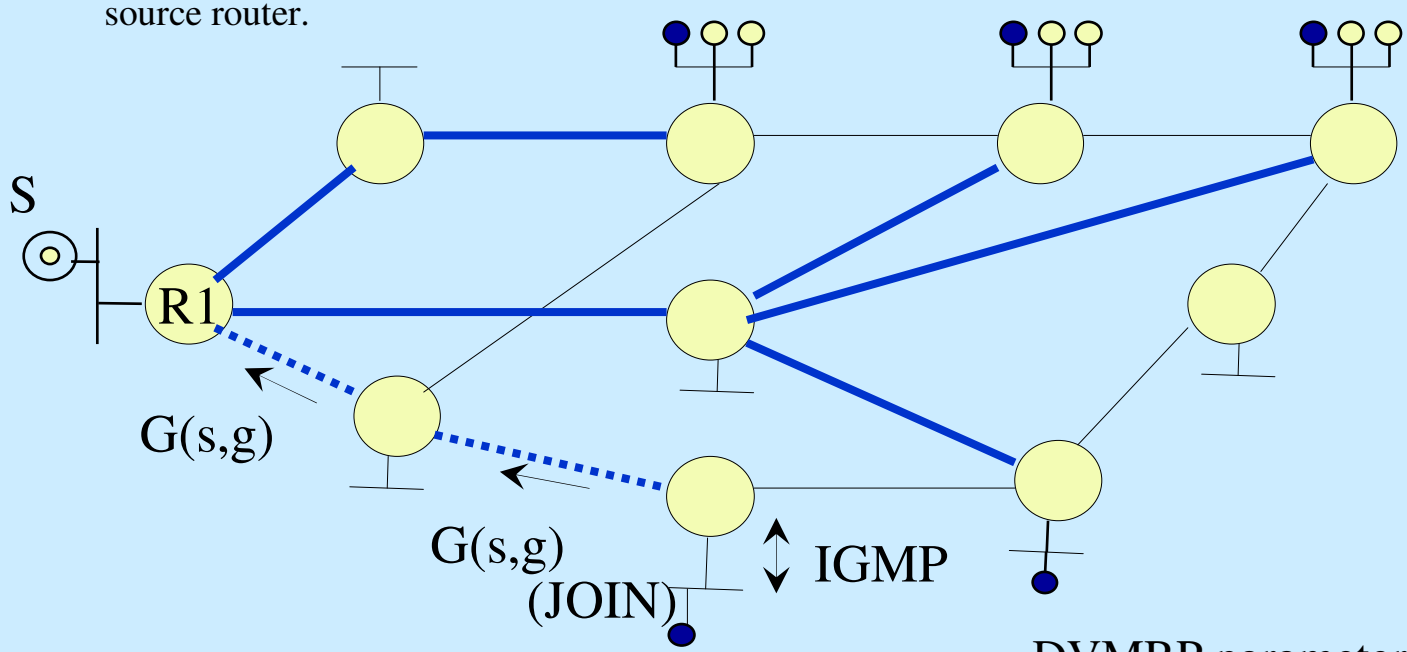
- **DVMRP – Example: prune messages**
  - Note the tree is aggressive: if no longer *prune* msg is received on a branch then the tree is automatically extended in that direction, by the node observing the *missing-prune* event



Final multicast tree after pruning

3. IP Level Intra-domain Multicast

- **DVMRP:** Grafting the tree (new branches added to the tree)
  - The *graft* messages reduce the connection delay time for a new host to the RSPT tree.
  - The *flood* and *prune* process is periodical but a host can connect faster by using graft messages:
  - - the host joins the group via IGMP
  - - the elected router for multicast can send a graft (S,G) message on the I/F which offers the shortest path towards source S
  - - the message is relayed up to the closest router already connected in RSPT orup to the source router.



3. IP Level Intra-domain Multicast

- **DVMRP**
  - (+) easy to implement when compared, for instance, with MOSPF, described later.
  - (+) computational complexity is low (RPF check for every packet and maintaining "prune" timers at every node for every active source and downstream interface).
  - (-) assumes that routes between every two nodes are *symmetric* and of equal cost and tunnels can be used when these assumptions do not apply
  - (-) not scalable for sparse groups and large networks
  - DVMRP's *deployment* is mainly bounded to the Mbone.
  - DVMRP is available in public domain (m-routed): it is accessible to all who want to participate in *Mbone* multicast sessions.

3. IP Level Intra-domain Multicast

- **Multicast Open Shortest Path (MOSPF)**
- MOSPF is an extension of OSPF
  - Dense mode protocol
  - Constructs a source rooted SPT
  - **MOSPF builds a multicast forwarding tree on demand for each (S,G), pair**
    - **it uses group membership obtained from IGMP**
    - **and unicast routing info obtained from OSPF**
  - supports hierarchical routing; hosts are partitioned in ASes
  - MOSPF is defined in RFC-1584 and *depends* on OSPF, RFC-1583
    - to construct the unicast routing table
  - (+) OSPF can use different types of a single link state metric (e.g., delay, number of traversed hops) to express the cost of a path.
  - MOSPF complements OSPF's routing database with a **new type of "link**
  - **state advertisement" records: the group memberships.**

3. IP Level Intra-domain Multicast

- **Multicast Open Shortest Path (MOSPF)**
  - MOSPF routers can essentially perform the *RPF check* and join and prune computations ***locally***
    - Given that every MOSPF router has **complete information about the routing topology and receivers' locations.**
  - Thus, on-tree routers can build *source-rooted trees –SPTs* **without having to flood the first datagram of each of the sources.**
  - The *unidirectional* tree is built **on-demand when the first datagram from a source reaches an MOSPF router**
  - Thus, routers that are not part of the tree **do not perform any computation for the group**
    - **because they will nor receive data mc packets** (they do not "know" about this tree)

3. IP Level Intra-domain Multicast

- **Multicast Open Shortest Path (MOSPF)**
- **Intra-area routing**
  - Runs in a single OSPF area and supports mc
  - (source and the multicast receivers are in the same OSPF area, or AS = area)
  - Each MOSPF router maintains a **local group DB** (list of directly attached group members)
  - Each subnetwork has a designated router (DR)
- DR :
  - sends IGMP host membership queries and listens to reports
  - propagates in the area, this group membership Info **to all other routers** by using ***group membership link state advertisement (LSA)***
  - SR-SPT is built based on the **router LSAs** and **network LSAs** in the MOSPF link state database
  - SR-SPT + router's local group database info are used to build a forwarding table (cache) at each router **for each (S,G) pair**.
    - This table is used to forward subsequent datagrams.

3. IP Level Intra-domain Multicast

- **Multicast Open Shortest Path (MOSPF)**
  Complexity Issues and Design Decisions
  - **MOSPF requires *heavy computation* for each (S,G)combination**
  - Considering that in a routing domain
    - there are as many potential sources as the number of hosts
    - and that the number of groups is likely to grow with the size of the routing domain (also referred to as "autonomous system" in MOSPF),
  - then the number of computations that follow any routing update is likely to grow at the $O(N^2)$,
    - N is the number of nodes in the network
    - The best possible case for *Dijkstra*'s computations is of the order of $O(N.\log N)$.
  - **Solution to improve *scalability* :** on demand tree computation
    - I.e. the tree will be ***calculated only when the first packet from a source S to a group G is received*.**
  - After that, the ***group membership information*** is used to prune the branches of the tree that do not lead to any group member.
  - **Finally, the multicast packet is forwarded to those outgoing interfaces that belong to the pruned multicast tree (source-based tree).**

3. IP Level Intra-domain Multicast

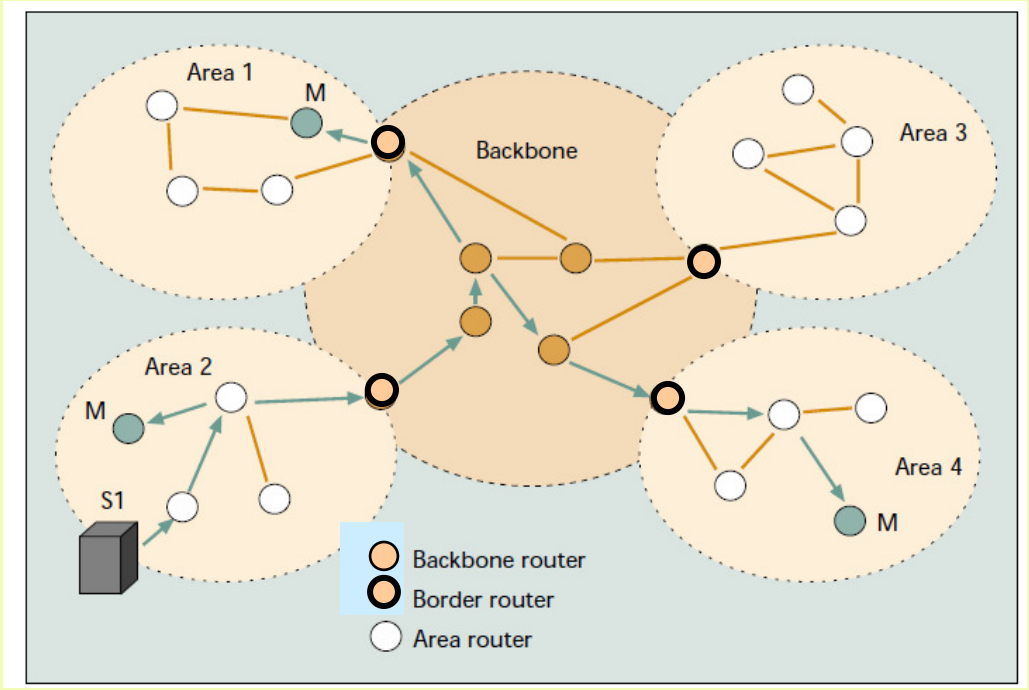
- **Multicast Open Shortest Path (MOSPF)- details**

  Summary of Intra-Area routing

  - **Prerequisits**
    - OSPF allows a AS to be split into areas.
    - The OSPF link state **DB provides the complete map of an area at each router**.
    - By adding a new type of link state advertisement "**Group-Membership-LSA"** the info about the **location of members of multicast groups** can be obtained and put in the database.
  - The trees are constructed **on demand** (when a router receives the first mc. datagram of a (S,G) pair
  - From OSPF link state information, **SR-SPT is constructed (Dijkstra) for this (S,G)** pair
  - Then, **group membership info is used to prune** unnecessary links
  - Since all area routers have the complete information on area topology and group memberships => **all the routers will build with the same tree** for a given (S,G)
    - as long as source and all group members are in the same area.
  - A router knows its predecessor and successors for each (S,G) tree
  - At each router the **"forwarding cache"** is created (separate entry for each (S, G) pair), containing info:
    - on which I/F the packets are expected to be received
    - on which I/Fs the packets should be forwarded.
  - Unlike DVMRP, the first packet need not to be flooded in an area.

3. IP Level Intra-domain Multicast

- **Multicast Open Shortest Path (MOSPF) -details**
  - **Inter-area routing**
    - **divide the routing domain in routing areas inter-connected via a backbone area**
    - The number of routers per area is limited to a max.
    - multicasting between areas is always done via the backbone area



Example of mc tree for source S1 visualizing MOSPF division in areas connected via a BB area.
Border routers (BR) advertise the existence of members in their respective areas to the BB area.
The BRs of the area for which there are members of the group will, then, extend the tree to reach the new member.

3. IP Level Intra-domain Multicast

- **Multicast Open Shortest Path (MOSPF)**
  - **Inter-area Routing**
  - There are a number of special cases that make the SPT computation in MOSPF more complex
    - **A. Distributing membership info between areas**
    - **B. There can be ambiguity if having more than one path with equal cost.**

      Note that in order to avoid routing loops, all routers should construct locally the same shortest path tree

3. IP Level Intra-domain Multicast

- **Multicast Open Shortest Path (MOSPF)**
  - **Inter-Area Routing- solutions for previous A, B problems**
  - **A.** BRs advertise to the BB **the presence of at least one member** in their respective area.
  - This limits the number of group membership advertisements (LSA) to **one per group**.
  - For *inter-operating* with other protocols, there are **external routers** (border routers of the AS).
  - The external routers should not advertise internally all the groups that have been defined on the whole Internet
  - The solution is to consider, as default, that **external routers are members of all the groups,** and thus part of the source-based trees is computed in the backbone.
  - **B. The second issue** is solved by giving privilege to broadcast networks as well as paths serving multiple members.

3. IP Level Intra-domain Multicast

- **Multicast Open Shortest Path (MOSPF)**
- ***Inter-area routing***
  - *src* and *dest* are in different OSPF areas
  - the forwarding cache is built from the local group DB and the datagram SPT
  - for inter-area multicast, MOSPF uses a **subset of the area's area border router (ABR)** as ***inter-area multicast forwarders (IAMF)***
    - responsible for fwd. group membership and multicast datagrams between different OSPF areas
  - These IAMFs summarize their attached area's group membership info to the BB by originating new group membership LSAs.
  - **New concept:** *wildcard multicast receiver (WMR)* - to permit forwarding of mc traffic between areas
    - **WMR receives all mc traffic generated in an area, regardless of the multicast group membership**
  - For a non-backbone area, an IAMF works as a WMR so that all the mc traffic can be forwarded to backbone and other non-backbone areas.
  - The BB has the complete picture of group membership of different areas.

3. IP Level Intra-domain Multicast

- **Multicast Open Shortest Path (MOSPF)**
- ***Inter-area routing summary***
  - A **subset of the area border routers (ABRs)** are elected to function as **"inter-area multicast forwarders"(IAMF)**
  - IAMFs forward a **summarized version of group membership** information *of their attached areas* **to the BB** area using a new type of group membership LSAs.
  - This information *is not flooded into non-backbone areas.*
  - MOSPF has the concept of "**wild-card multicast receiver"**
  - **WMRs receive all the multicast messages originated in their areas.**
  - All IAMFs in non-backbone areas function as WMRs guaranteeing that all mc messages originated in a non-backbone area reaches a IAMF and can be forwarded to the BB area if it is necessary.
  - **BB has complete information about group memberships in different areas =>** multicast packets can be forwarded to the appropriate areas in AS.

3. IP Level Intra-domain Multicast

- **Multicast Open Shortest Path (MOSPF)**
- ***Inter-AS routing***
- source and/or some of the destination multicast group members are in different ASes
- Solution: similar to that of inter-area routing.
- Some of the AS Boundary Routers (ASBRs) are configures as "inter-AS multicast forwarders“
- MOSPF assumes that inter-AS multicast forwarders construct RPB trees for forwarding multicast messages.
- Inter-AS multicast forwarders are wildcard multicast receivers in their attached areas
  - guaranteeing that these routers remain on all multicast delivery trees and receive all multicast datagrams
- While forward path is used inside an AS, paths to external sources are found by using reverse-path source-based trees.

3. IP Level Intra-domain Multicast

- **Multicast Open Shortest Path (MOSPF)**
- Multicast group membership dynamics
  - MOSPF advertises changes in the set of receivers to all the nodes of the area
  - This will trigger an update of the routing state at every on-tree node, for each of the sources of the group.
  - If a new source becomes active, its adjacent router just needs to calculate the shortest path tree rooted at the new source, since it has updated information on the set of receivers.
  - (-) Given the above, one can conclude that MOSPF is slow to react when there is a high degree of dynamics in the set of receivers and incurs a high control message overhead in order to advertise membership changes
  - Moreover, it maintains a routing state entry per every (S,G), even if the source is just transmitting sporadically.

3. IP Level Intra-domain Multicast

- **Multicast Open Shortest Path (MOSPF)**
- Conclusions/critics
  - (-) MOSPF is *not scalable* for domains with a large number of nodes.
  - (+) The two-level hierarchy (areas connected to a BB area) has been one of the steps taken in order to overcome that
  - However, the hierarchy does not provide much value for multicast routing since there is no connection between group members and routing areas.
  - Because of all this, MOSPF has *not* been *widely deployed*.
  - MOSPF does not support tunnels nor any feature for *incremental deployment*.

3. IP Level Intra-domain Multicast

- **Protocol Independent Multicast**
  - Family of IP multicast routing protocols
  - PIM does not include its own topology discovery mechanism
  - It uses routing info supplied by other traditional routing protocols : RIP, OSPF, BGP, MDSP (Multicast Source Discovery Protocol)
  - **Variants:**
    - **PIM Sparse Mode (PIM-SM) RFC 4601:**
      - explicitly builds unidirectional shared trees rooted at a rendezvous point (RP) per group
      - optionally creates SPT per source (depending on traffic conditions)
      - scales fairly well for wide-area usage
    - **PIM Dense Mode (PIM-DM) RFC 3973:**
      - uses dense multicast routing (similar to DVMRP)
      - it implicitly builds SPTs by flooding multicast traffic domain wide, and then pruning branches where no receivers are present
      - straightforward to implement
      - poor scaling properties.

3. IP Level Intra-domain Multicast

- **Protocol Independent Multicast**
- Variants (cont'd)
  - **Bidirectional PIM (RFC 5015):**
    - explicitly builds *shared bi-directional trees*.
    - It never builds a SPT , so may have longer end-to-end delays than PIM-SM
    - scales well because it needs no source-specific state.
  - **PIM source-specific multicast (PIM-SSM) RFC 3569**
    - builds SR-SPTs (rooted in *just one* source)
    - offer a *more secure and scalable model* for a limited amount of applications (mostly broadcasting of content)
    - an IP datagram is transmitted by S to an SSM dest. address G
    - receivers can receive this datagram by subscribing to channel (S,G).

- **In practice:**
  - PIM-SM : widest deployment
  - PIM-SM **is commonly used in IPTV systems** for routing multicast streams between LANs, VLANs, Subnets

3. IP Level Intra-domain Multicast

- **Protocol Independent Multicast –Dense Mode (PIM-DM)**
- **PIM-DM - similar to DVMRP**
- It assumes that when S starts sending, all downstream Hs want to receive mc.
- Initially, mc. datagrams are flooded to all areas of the network
- PIM-DM uses RPF to prevent looping while flooding
- If some network areas do not have group members, PIM-DM will prune off the forwarding branch by instantiating *prune state*.
- *Prune state* has a finite lifetime: when that lifetime expires, data
- will again be forwarded down the previously pruned branch.
- *Prune state* is associated with an *(S,G) pair.*
  - When a new member for a G appears in a pruned area, a router can "graft" toward the S for the group, thereby turning the pruned branch back into a forwarding branch.

3. IP Level Intra-domain Multicast

- **Protocol Independent Multicast –Dense Mode (PIM-DM)**
  - The broadcast of datagrams followed by pruning of unwanted branches is referred to as a *flood and prune cycle* (typical of DM protocols).
  - To minimize repeated flooding of datagrams and subsequent pruning associated with an (S,G) pair, PIM-DM uses a state refresh msg.
  - This message is sent by the router(s) directly connected to S and is propagated throughout the network.
    - When received by a router on its RPF interface, the state refresh message causes an existing prune state to be refreshed.

3. IP Level Intra-domain Multicast

- **Protocol Independent Multicast –Dense Mode (PIM-DM)**
- Compared with mc.routing protocols with *built-in topology discovery mechanisms* (e.g., DVMRP)
- PIM-DM has a *simplified design* and is not *hard-wired into a specific topology discovery protocol*

- However, this simplification does incur more overhead by causing flooding and pruning to occur on some links that could be avoided if sufficient topology information were available;
  - i.e., to decide whether an I/F leads to any downstream members of a particular group.
  - Additional overhead is chosen in favor of the simplification and flexibility gained by not depending on a specific topology discovery protocol.

3. IP Level Intra-domain Multicast

- **Protocol Independent Multicast –Dense Mode (PIM-DM)**
- PIM-DM - DVMRP - two major differences.
  - 1. PIM (both DM and SM) uses the unicast routing table to perform RPF checks
    - DVMRP maintains its own routing table
    - PIM uses whatever unicast table is available
    - PIM needs an unicast routing table to exist *independent on how is built*
  - 2. DVMRP tries to avoid sending packets to neighbors who will then generate prune messages based on a failed RPF check.
    - The set of of outgoing I/Fs built by a DVMRP router include only its children (on the tree SPT to source)
    - PIM-DM is simpler: packets are forwarded on all outgoing interfaces.
    - Unnecessary packets are often forwarded to routers which must then generate prune messages because of the resulting RPF failure.

3. IP Level Intra-domain Multicast

- **Protocol Independent Multicast – Sparse Mode**
- **(PIM - SM)**
- Design goals
  - sparse mode regions (LANs, WANs)
  - low latency data distribution
  - maintain IP multicast model (receiver initiated group membership)
  - host model unchanged
  - independent of unicast routing protocol
  - unidirectional shared tree or can switch to source tree
  - soft state mechanism
  - interoperability (intradomain, interdomain)
  - robustness
  - scalability
  - remove the CBT shortcomings (traffic congestion and latency)
- RFC 2362 – Initial Standard for PIM – SM- 1998 June

3. IP Level Intra-domain Multicast

- **Protocol Independent Multicast – Sparse Mode**
- **(PIM - SM) – details**
- **Protocol Independent Multicast – Sparse Mode (PIM - SM) –**important and rather used multicast protocol.
- *Sparse mode* means that the number of networks having members of the group is much less than the total number of networks. The group members are widely distributed in different regions. In such conditions the overhead of broadcast & prune becomes quite significant and is not acceptable.
- The host model remains unchanged: PIM – SM is a router-to-router protocol.
- PIM-SM is independent of the used unicast routing protocol in the sense that PIM make use of unicast routing table no matter how it was created.
- PIM-SM can use either a unidirectional shared tree or a source routed tree – which assures a low latency for applications requiring this feature.
- PIM-SM is based on a soft state mechanism, that means:
  - unless refreshed, the router's state confirmation expires
  - adaptability to changes in network topology
- PIM-SM interoperability means that if deployed for inter-domain, then PIM-SM should interoperate with other traditional multicast routing protocols (DVMRP, MOSPF, etc.).
- Robustness characteristic requires that no single point of failure should exist.
- Scalability means that the overhead imposed by control messages for building the mc-tree should be less than an acceptable percentage of the link bandwidth, no matter the group dimension.

3. IP Level Intra-domain Multicast

- **Protocol Independent Multicast – Sparse Mode**
- (PIM - SM)
- PIM-SM version 2 was originally specified in RFC 2117
- Revised in RFC 2362, both Experimental RFCs
- RFC 4601- august 2006: correct a number of deficiencies that have been
- identified with the way PIM-SM was previously specified, and to bring PIM-SM onto the IETF Standards Track

- As far as possible, this document specifies the same protocol as RFC 2362 and only diverges from the behavior intended by RFC 2362 when the previously specified behavior was clearly incorrect.

- Routers implemented according to the RFC 4601- will be able to interoperate successfully with routers implemented according to RFC 2362

3. IP Level Intra-domain Multicast

- **Protocol Independent Multicast – Sparse Mode (PIM - SM)**
- **PIM-SM General Description**
  - much more widely used than Core Based Tree (CBT)
    - CBT creates a shared bi-directional tree
  - similar to PIM-DM w.r.t routing decisions (based on existent underlying unicast RT)
  - tree construction mechanism different for PIM-SM/DM
- PIM-SM's tree construction similar to that used by CBT
  - A core = *Rendezvous Point (RP)* must be configured (basic goal of RP: "meeting place" for sources and receivers)
  - different groups may use different routers for RPs
  - a group can only have a single RP
  - Info about RPs, and the mappings of mc groups to RPs, must be discovered by all routers
  - RP discovery is done using a *bootstrap protocol (BSP)*
    - RP discovery mechanism is not included in the PIM-SMv1 spec
    - a vendor implementations of PIM-SMv1 has its own RP discovery mechanism.
    - PIMSMv2 includes BSP in its spec.
- **Function of the BSP**
  - RP discovery,
  - provide robustness in case of RP failure.
  - BSP can select an alternate RP if the primary RP fails

3. IP Level Intra-domain Multicast

- **Protocol Independent Multicast – Sparse Mode (PIM - SM)**
- **PIM-SM General Description**
- **Receivers JOIN**
  - They send explicit *join messages* to the RP
  - Fwd. state is created in each router along the path from the receiver to the RP
  - One shared tree (reverse SPT), rooted at the RP per group
  - *Join* messages follow a reverse path from receivers to the RP.
- **Data Transfer**
  - Each S sends mc data packets, encapsulated in unicast packets, to the RP.
  - RP receives one of these *register packets*, and may do several actions:
  - **1. if the RP has forwarding state for the group** (i.e.,there are receivers who have joined the group)
    - then the encapsulation is stripped off the packet
    - and it is sent in mc mode on the shared tree
    - if the RP does not have forwarding state for the group
    - it sends a *register-stop message* to the S
    - this avoids wasting bandwidth between the source and the RP
  - 2. the RP **may** wish to send a join message toward S.
  - mc fwd. state is setup between S and the RP
  - then RP can receive S traffic as mc (avoid encapsulation)
  - **A receiver may switch on a tree SPT-type, S rooted.**

3. IP Level Intra-domain Multicast

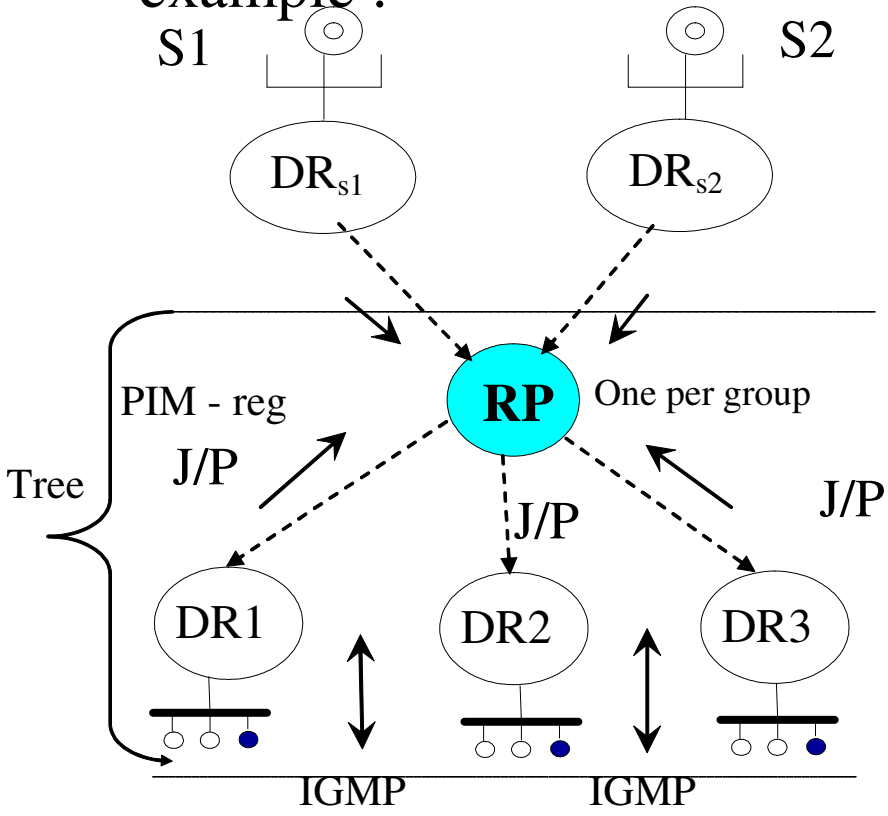
- **Protocol Independent Multicast – Sparse Mode**
- **Components and functions**
  - ***Designated Routers (DR)-*** act on behalf of hosts
  - ***Rendez-Vous Points (RP)-*** configured as a root for a shared unidirectional tree
  - ***Last Hop Routers (LHR)-*** last router on the RP-rooted Shared Tree (RPT)
  - ***Boot Strap Routers (BSR)*** dynamically elected router ; builds the set of RPs
- **Phases of operation**
  1. Creating the PIM framework
  2. Build Shared Tree
  3. Multicast data forwarding
  4. Stop encapsulation
  5. Switch to source SPT
  6. Prune Shared Tree

3. IP Level Intra-domain Multicast

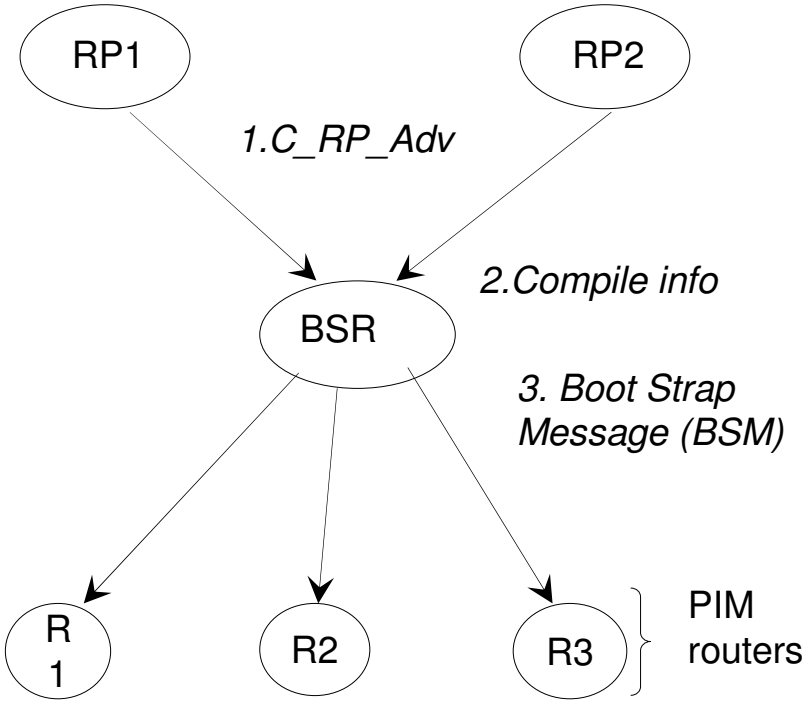
- **Protocol Independent Multicast – Sparse Mode**
- **Components and functions- details**
- *Designated Router* (DR) is a router which will act on behalf of hosts w.r.t. PIM DR is elected by a simple election process.
- *Rendez-Vous Point* (RP) is a router configured as a root for a shared unidirectional tree containing all receivers of a group. There is only one RP per group. But we can have several RPs in a network (1 / group).
- *Last Hop Router* (LHR) is the last router on the RP-rooted Shared Tree (RPT) before a LAN containing hosts that belongs to M (M = multicast group). LHR may be the same as DR but not necessarily.
- *Bootstrap Router* (BSR) – dynamically elected router which constructs the set of RPs and distributes their identities among the PIM routers.

3. IP Level Intra-domain Multicast

- **PIM – SM**
- PIM components - example :



1. Creating PIM framework

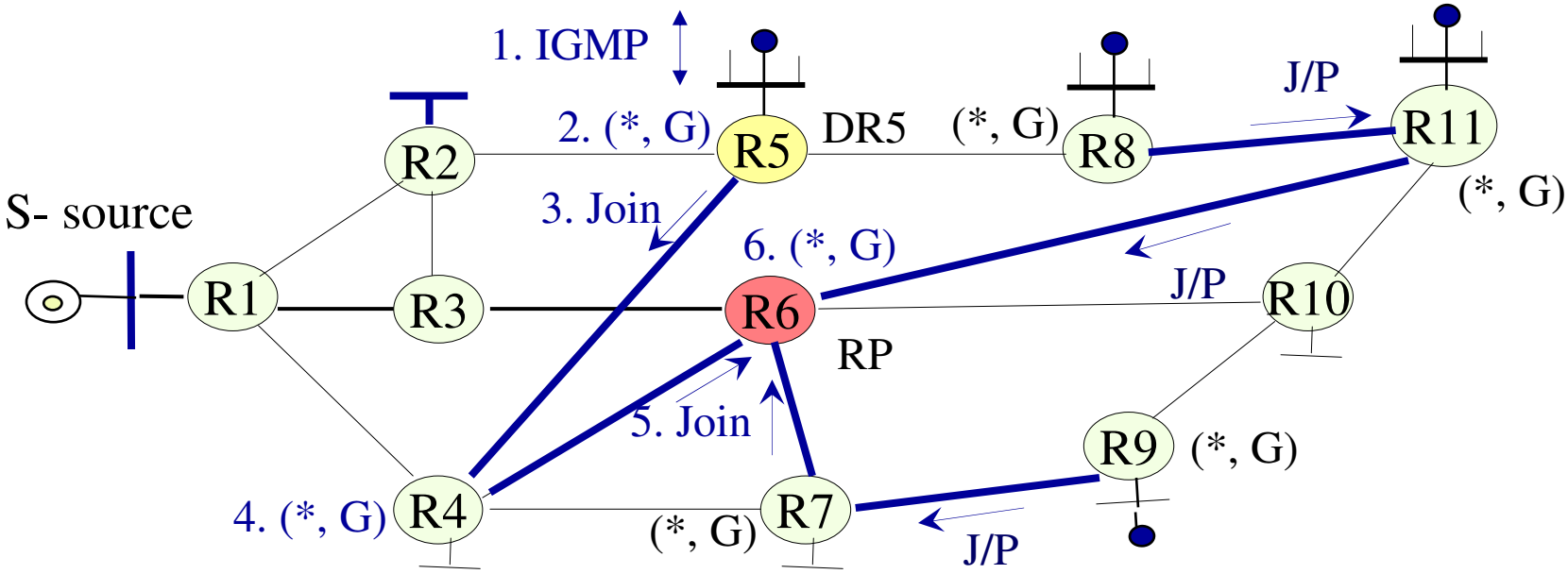


3. IP Level Intra-domain Multicast

- **Protocol Independent Multicast – Sparse Mode**
- **Components and functions- details**
- Left side figure: the comonents of PIM-SM: Designated Routers, Rendez-Vous Point , Sources and Receivers.
- The receivers register to their DRs by IGMP.
- The DRs of the receivers *join* the RP. The current adopted PIM-SM version suppose that there are **one RP per group**.
- The right figure illustrate the principle of PIM framework creation. The *Boot Strap Router* BSR is a dynamically elected router within a PIM domain, responsible for constructing the set of RPs and *distributing the set of RPs identities to the routers within the domain*.
- Thr RPs send *Candidate_RP_Advertisments* to the BSR.
- Based on this, BSR compiles the set of RPs.
- BSR distribute the RP set information to the routers using bootstrap messages.

3. IP Level Intra-domain Multicast

- **PIM – SM**
- **Phase 2 : Building RP rooted shared tree (RPT)**
- J/P = **Join**/Prune message- sent periodically as long as the respective router belongs to G
- (*,G) state – non source specific
- Hypothesis: all routers PIM-capable



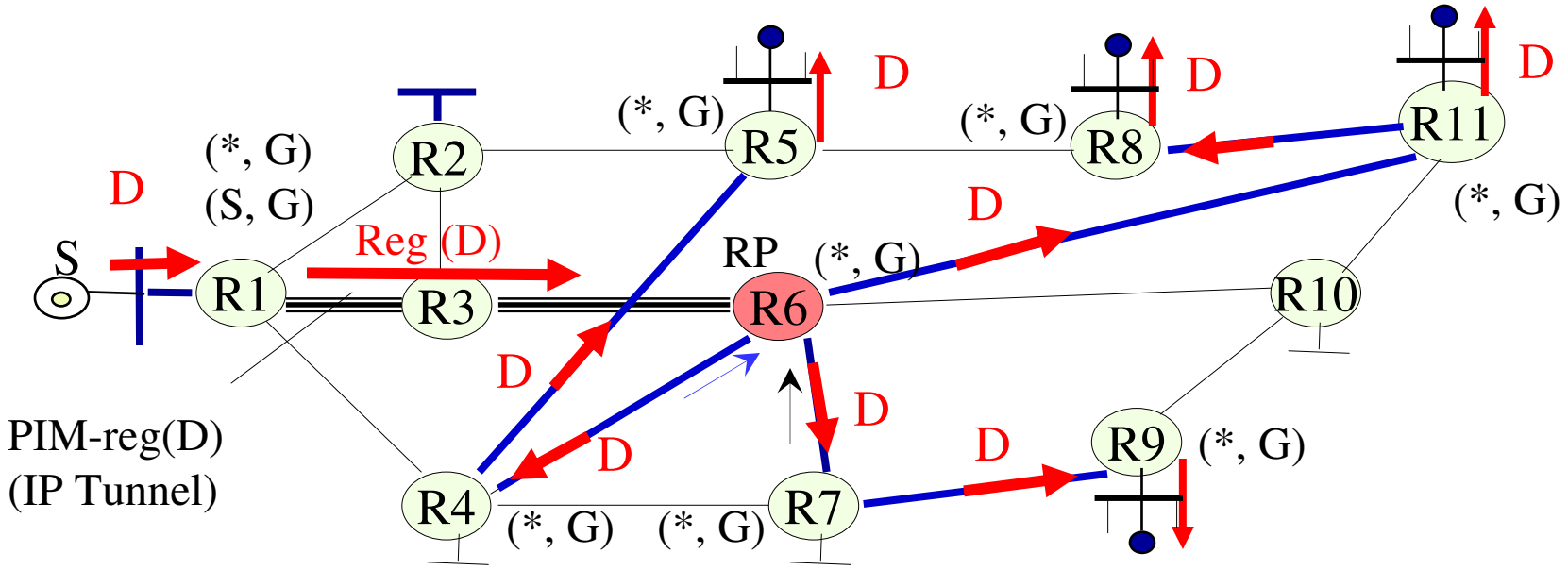
Result: Unidirectional Shared Tree rooted in RP

3. IP Level Intra-domain Multicast

- **Protocol Independent Multicast – Sparse Mode**
- **Phase 2: Building the RP Tree**
- a multicast receiver (Rec) expresses its interest in receiving traffic destined for a multicast group. (using **IGMP** or other mechanisms)
- One of the receiver's local routers is elected as the Designated Router (DR) for that subnet.
- On receiving the receiver's expression of interest, the DR then sends a *PIM Join* message towards the RP for that mc group, denoted ***(*,G) Join*** because it joins group G for all sources that may send info to that group.
- (*,G) Join travels hop-by-hop towards the RP for the group; in each router it passes through, a ***mc tree state (*, G) is instantiated.***
- Eventually the ***(*,G)*** *Join* either reaches the RP, or reaches a router that already has ***(*,G) Join*** state for that group.
- The receivers joined to the group form a distribution shared unidirectional tree (RPT) for group G , rooted at the RP
- **Join messages are resent periodically so long as the receiver remains in the group.**
- When all receivers on a leaf-network leave the group, the DR will send a PIM *(*,G) Prune* message towards the RP for that mc group.
- if the Prune message is not sent for any reason, the state will eventually time out.

3. IP Level Intra-domain Multicast

- **PIM – SM**
- Phase 3 : Data multicasting on shared unidirectional tree (RPT):
  - mc packets encapsulated (unicast) and tunnelled to RP (S → RP)
  - native multicast data packets distributed on RPT (RP → Receivers)



Unidirectional Shared Tree

IP Tunnel (DR_{S1} – RP) – for multicast packets encapsulated and sent by S_1 to RP

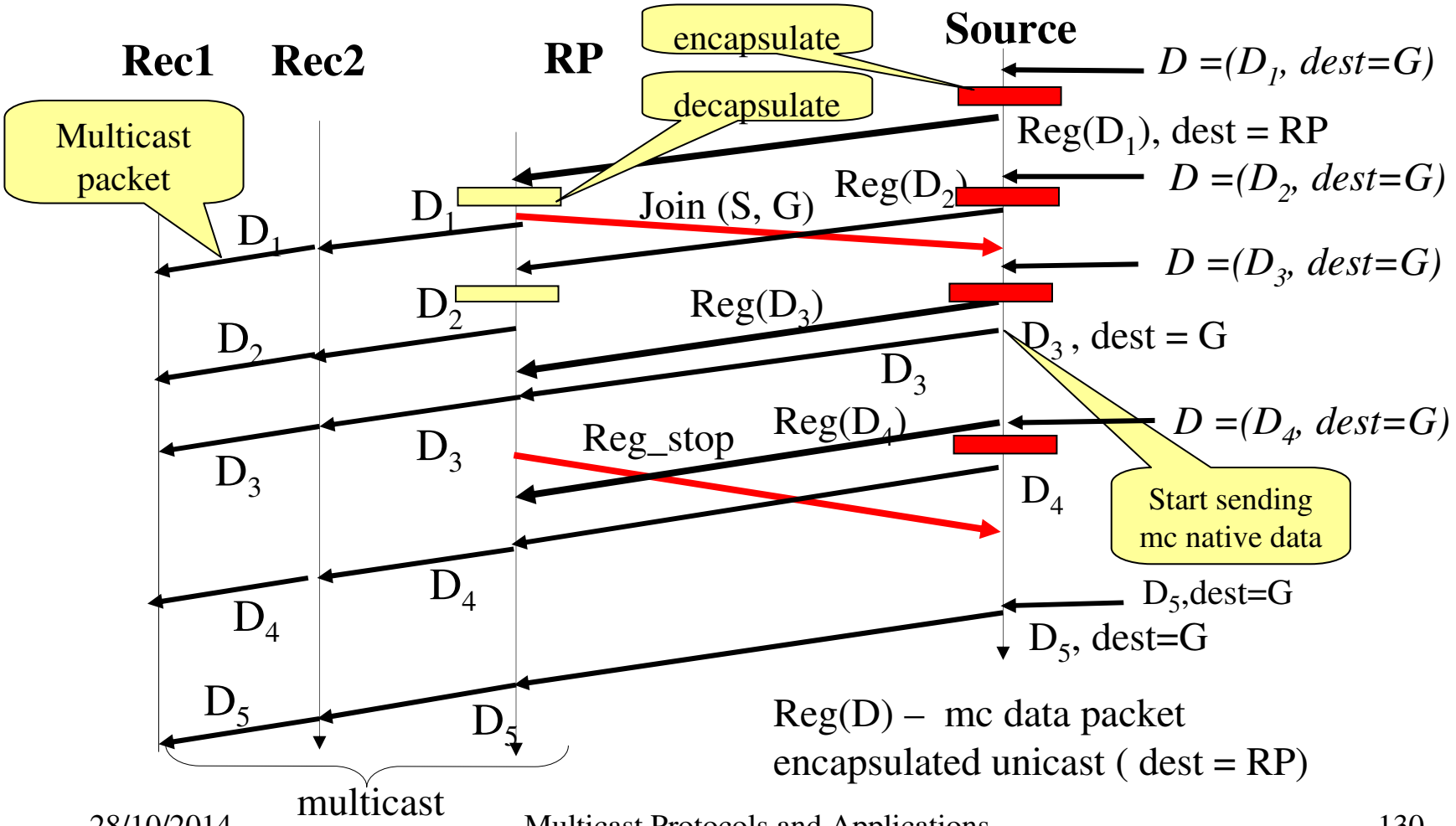
3. IP Level Intra-domain Multicast

- **PIM – SM**
- Phase 3 : Data multicasting on shared unidirectional tree (RPT): details
- A multicast data sender just starts sending data destined for a multicast group.
- The sender's local router (DR) takes data packets, **unicast-encapsulates** them, and sends them directly to the RP.
- The RP receives the encapsulated packets, **decapsulates them**, and forwards them onto RPT(to mc_group_address)
- The packets then **follow the (*,G) mc tree state** in the routers on the RPT, being replicated wherever the RPT branches, and eventually reaching all the receivers for that mc group.
- **The process of encapsulating data packets to the RP is called registering.**
- The encapsulation packets are known as ***PIM Register packets*** **Reg(D).**
-
- The multicast traffic is flowing encapsulated to the RP, and then natively over the RP tree to the multicast receivers.

3.30 IP Multicast Protocols

•PIM – SM

• Phase 4: Stop encapsulation – message sequence example



3.30 IP Multicast Protocols

•PIM – SM

• Phase 4: Stop encapsulation – details

Register-encapsulation of data packets is inefficient for two reasons:
- the are expensive operations for a router (depends on having appropriate hardware)
- the path length S-RP-Rec can be much longer than a direct path S-Rec. For some applications, this increased latency is undesirable.

Therefore RP will normally choose to switch to native forwarding.

Sub-phase 4.1 RP joins the source

Receiving a register-encapsulated data packet from S, RP normally initiates an *(S,G) source-specific Join* towards S, which travels hop-by-hop, instantiating (S,G) mc tree state in the routers along the path to S.

(S,G) mc tree state is used only to forward packets for group G, if they come from source S.

Eventually the *Join(S,G)* message reaches the source S's subnet or a router that already has (S,G) multicast tree state.

3.30 IP Multicast Protocols

•PIM – SM

• Phase 4: Stop encapsulation – details

Sub-phase 4.2 Native mc data sent by S to RP

Then S data packets start to flow following the (S,G) tree state towards the RP. If along the path towards the RP, they reach routers with (*,G) state, then the data packets can short-cut onto the RP tree at this point (**possible, as they are native mc addressed packets)**

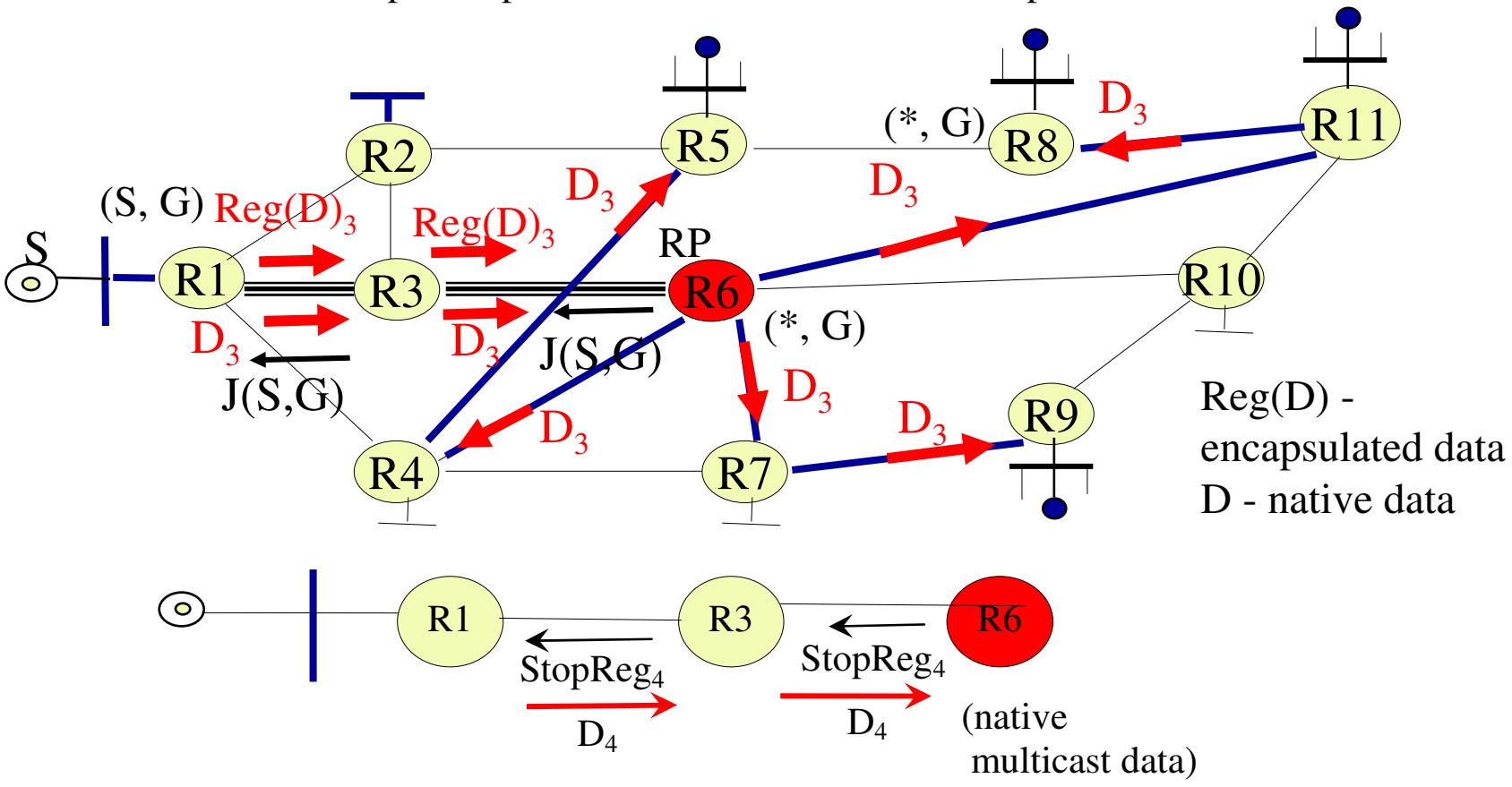
While RP is in the process of joining the source-specific tree for S, the data packets will continue being encapsulated to the RP. When packets from S also start to arrive natively at RP, it will be receiving two copies of each packet. Then RP starts to discard the encapsulated copies and it sends a *RegisterStop* message back to S's DR to prevent the DR unnecessarily encapsulating the packets.

At the end of phase 4, traffic will be flowing natively from S along a source-specific tree to the RP, and from there along the shared tree to he receivers. Where the two trees intersect, traffic may transfer from the source-specific tree to the RP tree, and so avoid taking a long path via the RP.

A sender may start sending before or after a receiver joins the group, and thus phase four may happen before the RPT to the receiver is built.

3. IP Level Intra-domain Multicast

- **PIM – SM**
- Phase 4 : RP join to the S for a group
- Sender DR stops encapsulation of source multicast data packets



3. IP Level Intra-domain Multicast

- **PIM – SM**
- **Phase 4 : RP join to the S for a group- details**
- The previous figure is an example associated to the previous message sequence chart diagram, showing the RP joining to the source for the group G. and requesting to DR of the source to stop encapsulation of multicast data packets.
- The sequence of RP actions has been presented in the previous slide, that are:

1. Joining to the source S for the group G (J(S,G) message.
2. Request to DR of the source to stop encapsulation of data packets.

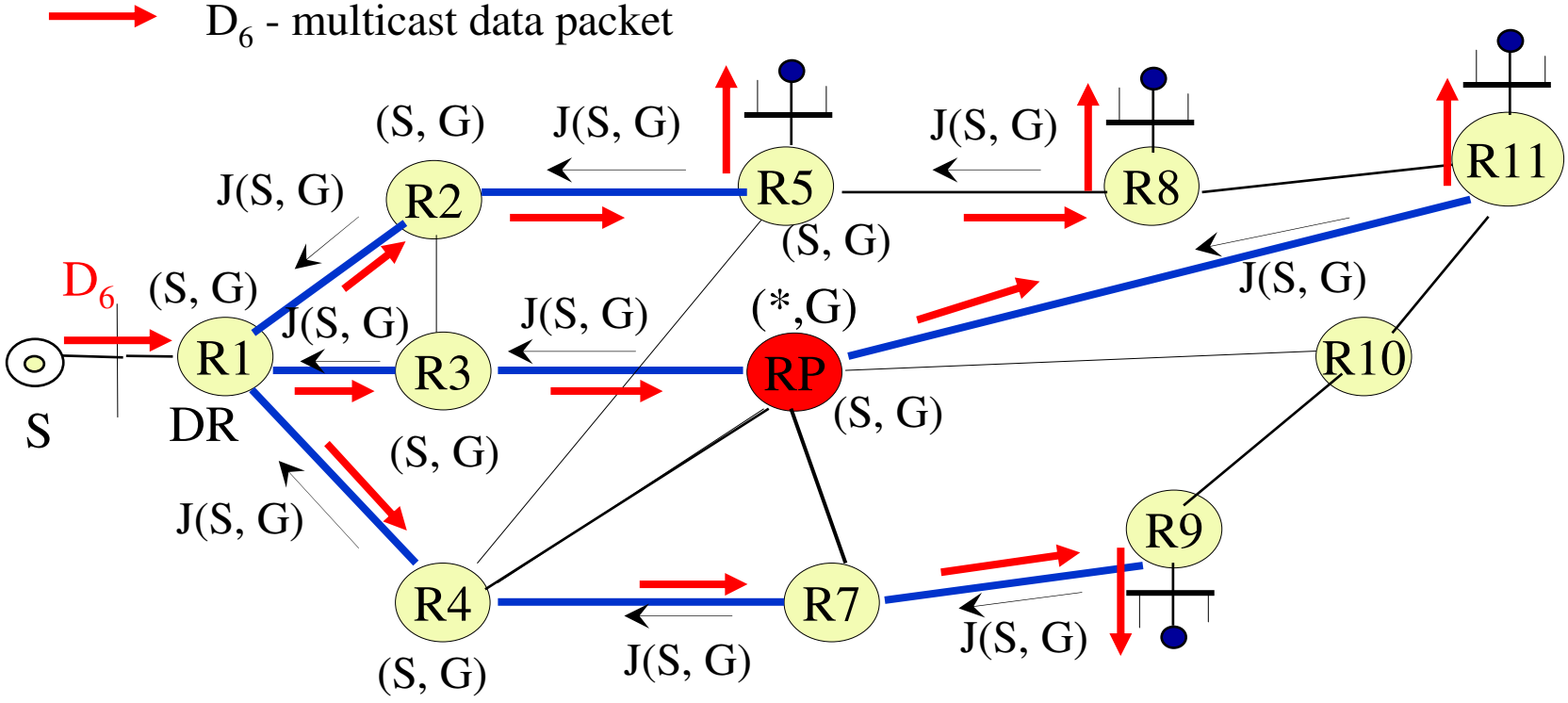
3. IP Level Intra-domain Multicast

•PIM – SM

- **Phase 5: Switch to Shortest Path Tree:**

  5.1 The DR router of a receiver joins the source: J(S,G) on SPT

  5.2 The data are multicasted by S onto SPT to receivers



3. IP Level Intra-domain Multicast

•PIM – SM

• Phase 5: Switch to Shortest Path Tree - details

5.1 The DR router of a receiver joins the source: J(S,G) on SPT

5.2 The data are multicasted by S onto SPT to receivers

Explanation:

•The *encapsulation stop* does not completely optimize the forwarding paths. For some receivers the route via the RP is longer when compared with the SPT from the source to the receiver.

•That is why, a router on the receiver's LAN, typically the DR, may ***optionally initiate*** a transfer from the ***shared tree*** to a ***source-specific shortest-path tree (SPT).*** It issues an *(S,G) Join* towards S.

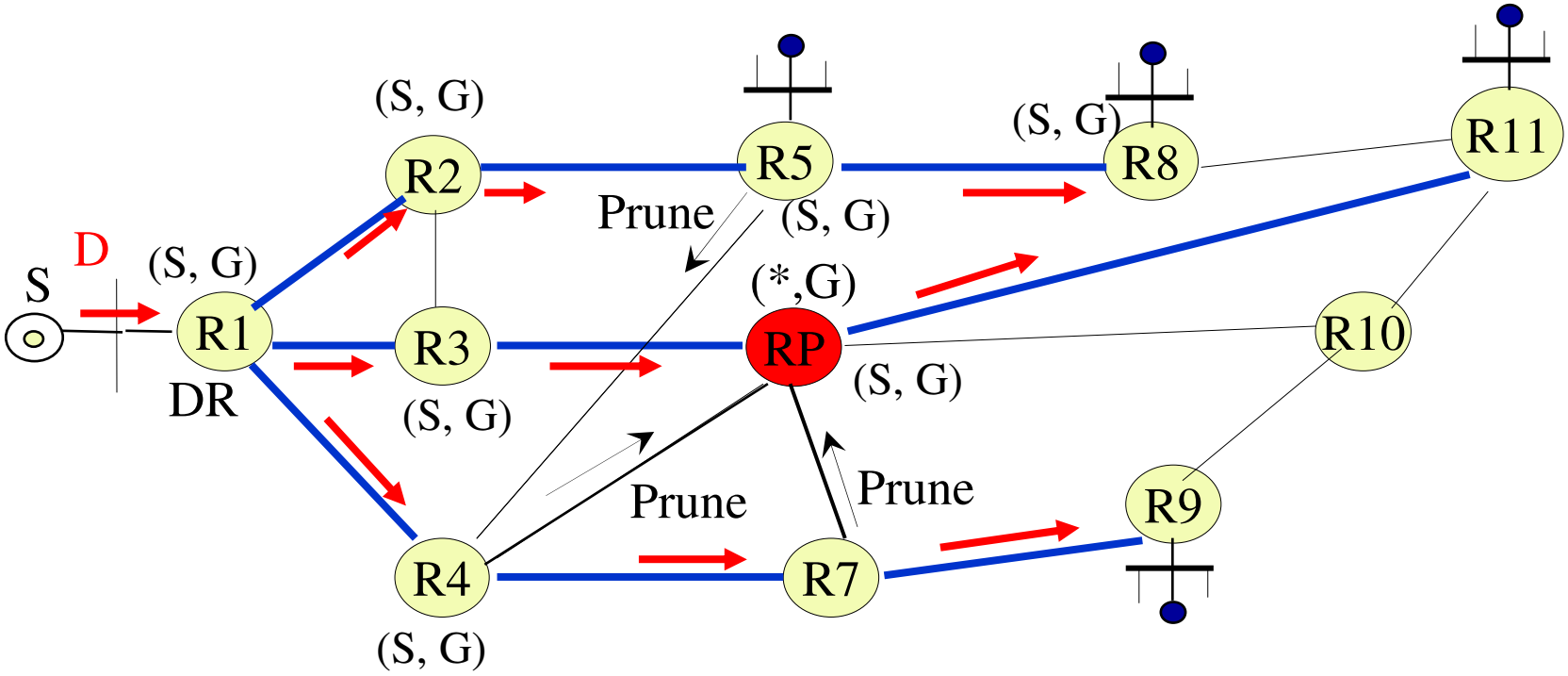
•This instantiates state in the routers along the path to S. Eventually this join either reaches the subnet of S, or reaches a router that already has (S,G) state.

•When this happens, data packets from S start to flow following the (S,G) state until they reach the receiver.

3. IP Level Intra-domain Multicast

•PIM – SM

- **Phase 6:Pruning the RPT:** *Prune (S, G, rpt)* messages
- D - Data packet



3. IP Level Intra-domain Multicast

•PIM – SM

• **Phase 6:Pruning the RPT:** *Prune (S, G, rpt)* messages

• D - Data packet

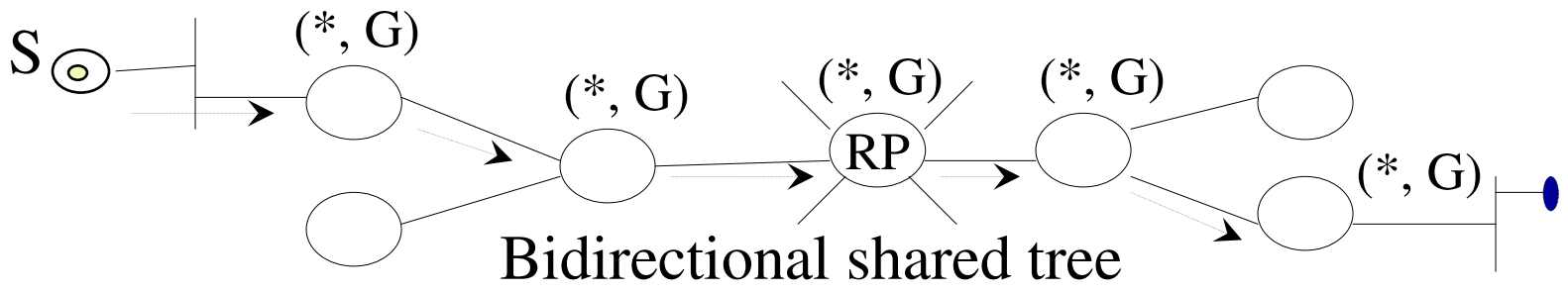
•The receiver (or a router upstream of the receiver) will be receiving two copies of the data - one from the SPT and one from the RPT. When the first traffic starts to arrive from the SPT, the DR or upstream router starts to drop the packets for G from S that arrive via the RP tree. In addition, it sends an (S,G) Prune message towards the RP. This is known as an *(S,G, rpt) Prune*.

•The Prune message travels hop-by-hop, instantiating state along the path towards the RP indicating that traffic from S for G should NOT be forwarded in this direction. The prune is propagated until it reaches the RP or a router that still needs the traffic from S for other receivers.

•By now, the receiver will be receiving traffic from S along the shortest-path tree between the receiver and S. In addition, the RP is receiving the traffic from S, but this traffic is no longer reaching the receiver along the RP tree. As far as the receiver is concerned, this is the final distribution tree.

3. IP Level Intra-domain Multicast

- **Bidirectional PIM (RFC 5015)**
- efficient many-to-many communication within individual PIM domain
- traffic routed along a bidirectional tree rooted in RP
- derived from PIM – SM
- no registering process as in PIM – SM
- eliminates any source specific state



3. IP Level Intra-domain Multicast

- **Bidirectional PIM (RFC 5015)- details**
- Variant of PIM sparse-Mode.
- It builds **bi-directional shared trees** connecting multicast sources and receivers.
- It dispenses with both encapsulation and source state
  - by allowing packets to be natively forwarded from a source to the RP using shared tree state.
- Bi-directional trees are built using a fail-safe **Designated Forwarder (DF)** election mechanism operating on each link of a mc topology.
- With the assistance of the DF, mc data is natively forwarded from sources to the RP and hence along the shared tree to receivers without requiring source-specific state.
  - The DF election takes place at RP discovery time and provides a default route to the RP thus eliminating the requirement for data-driven protocol events.

- The main differences between Bidir PIM and sparse-mode PIM are:
- Bidir PIM uses a **single shared tree** for for all the sources of a multicast group.
  - This reduces state requirements on a router.
  - The drawback is that it may produce suboptimal paths from sources to receivers.

3. IP Level Intra-domain Multicast

- **Bidirectional PIM (RFC 5015)- details**

- In Bidir PIM, packets traveling from a source to the RP, are natively forwarded on the shared tree. In contrast sparse-mode PIM uses unicast encapsulation or source-specific state.
- In Bidir PIM, sender-only branches do not need to keep group state. Data from the source can be natively forwarded towards the RP using RP-specific forwarding state.
- The Bidir Designated Forwarder (DF) assumes all the responsibilities of the sparse-mode DR.
- With PIM-SM, when forwarding packets using shared-tree (*,G) state, a directly-connected-source check has to be made on every packet.
  - This is done to determine if the packet was originated by a source which is directly connected to the router. For a connected source, source-specific state has to be created to register packets to the RP and prune the source off the shared tree.

- With Bidir PIM directly connected sources do not need any special handling. The DF for the RP of the group the source is sending to, seamlessly picks-up and forwards upstream traveling packets.