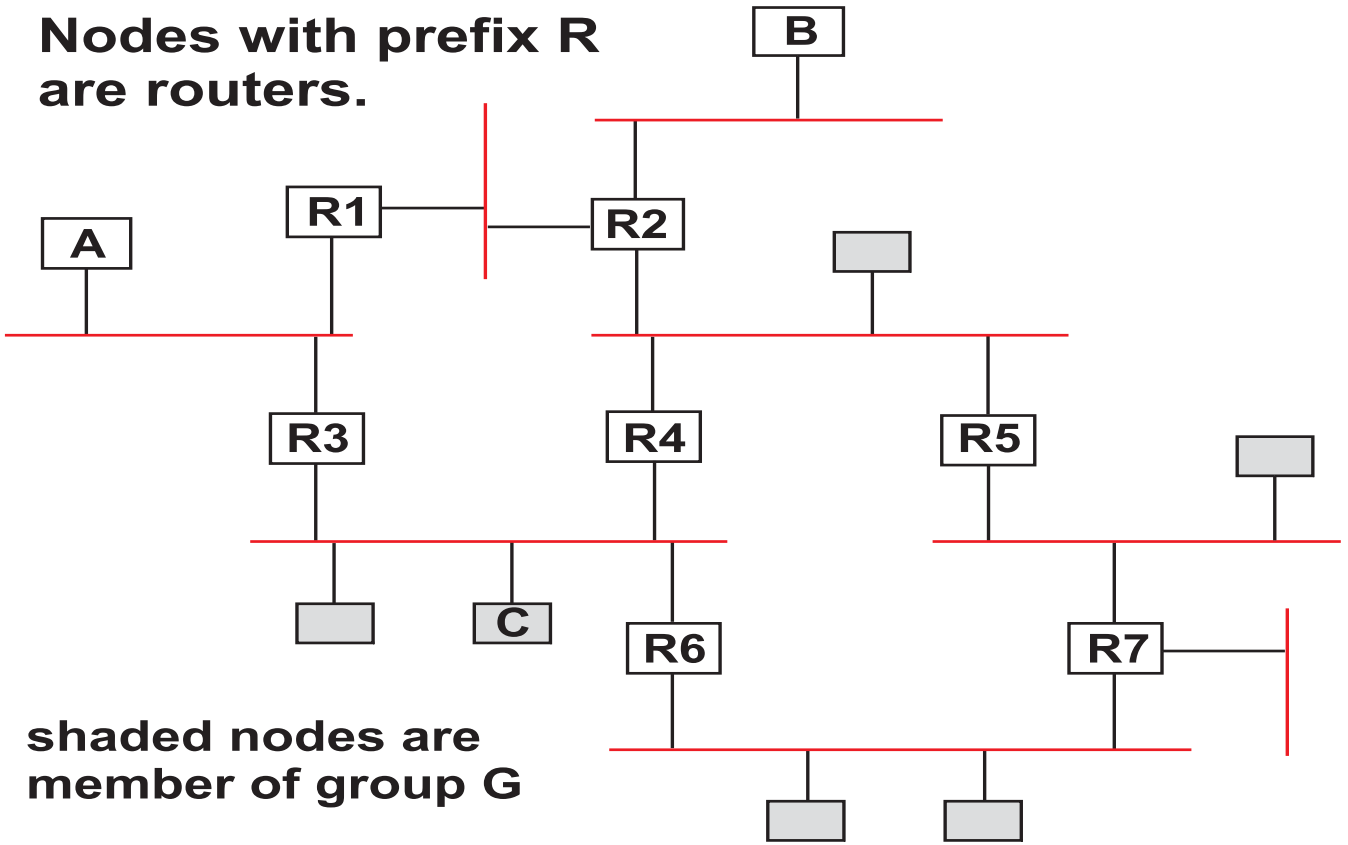
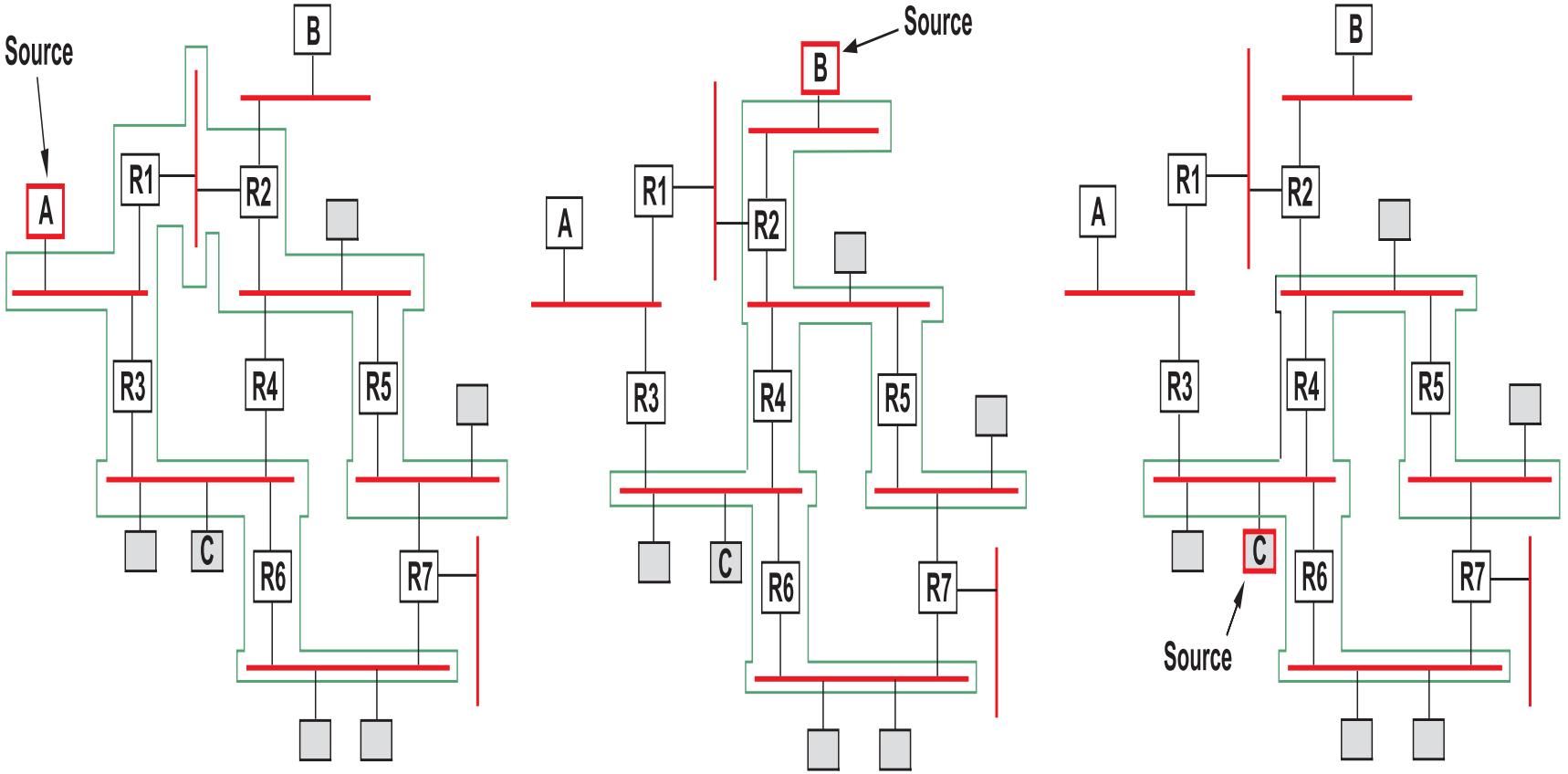
Internet Multicast Routing

- Distribute information from a source to multiple destinations (multicast group) — seminar, meetings, distance learning, van multicast services.
- MBONE (Internet Multicast BackBone) is an example of its usage.



Examples of Shortest-path Multicast Tree



- Source does not have to be a member of the group, e.g. A and B.
- LANs without member do not have to be involved in multicasting.
- Members can join and leave without synchronization or negotiation.

Multicast Routing in Internet and Extended LANs [Deering 88]

Multicast Routing provides two benefits to distributed applications:

- For sending same information to multiple destinations, multicast is more efficient than unicast.
  —reduce transmission overhead on the sender and the network
  —reduce the time it takes for all destinations to receive the information.
- Query or locate one or more address unknown or moving hosts.
  Multicast serves as a simple, robust alternative to
  —configuration files
  —name servers or other binding mechanisms.

How are multicast capability presented to the users?

- UDP broadcast sockets (use software filtering on the receiving hosts)
- process groups in Stanford V Kernel
- NetBIOS multicast datagrams in MS-DOS.

New Ethernet and other network conforming IEEE 802 standard will have multicast capability.

Link-layer bridges, such as DEC LANBridge 100 and Vitalink TransLAN, extend LAN performance and functionality (multicast) across multiple networks.

Network Layer Multicast Routing

Extend the existing point-to-point routing algorithms

- distance-vector routing
- link-state routing

to provide LAN-style multicasting across datagrams-based Internetworks.

Properties of LAN-style multicasting

- Group addressing
  - — The sender needs not know the membership of the group and need not be a member of a group.
  - — There is no restriction on the number or location of hosts in a group.
  - — A host can join and leave a group at will (no synchronization or negotiation with other member of the group is needed.)

  ##What are the drawbacks of using group addressing?
- High probability of delivery

  —the probability of damaging, duplicate, or misordering of multicast packets in a LAN is low but not zero, end-to-end protocols are used to recover from such events.
- Low delay—low *join latency* (time to update a local address filter(s))

TTL (Time To Live) field in Multicast Msg

By using a very small TTL value, a sender may limit the "scope" of a multicast packet to reach nearby group members only.

—It is called "expanding ring searching" in Boggs' dissertation "internet broadcasting".

Assumed Environment for Internetwork Multicasting

- Multi-access networks (LAN, satellite networks)
- interconnected in an arbitrary topology by packet switching nodes (bridges or routers)
- Additional point-to-point links (fibers or X.25 links) connect switching nodes.
- Assume the globally-unique internetwork multicast addresses can be mapped to corresponding LAN multicast addresses (one on one basis).

Single-Spanning-Tree Multicast Routing

Radia Perlman's distributed spanning tree algorithm is run by bridges to compute a spanning tree that is used to restrict packet traffic and avoid looping.

- — Packet is forward to all incident branch of the tree except the arrival one.
- — The packet is delivered to a segment exactly once.
- — Only forward multicast message to branches led to group members.

  If group members periodically send membership-report msgs with group address in src field,

  then the bridge can apply learning algorithm to know which branches led to group members.

Example

Assume B is all-bridge group which consists of all bridges.

Each host of Group G report its membership to bridges by sending a membership reporting packet with source address G, destination address B.

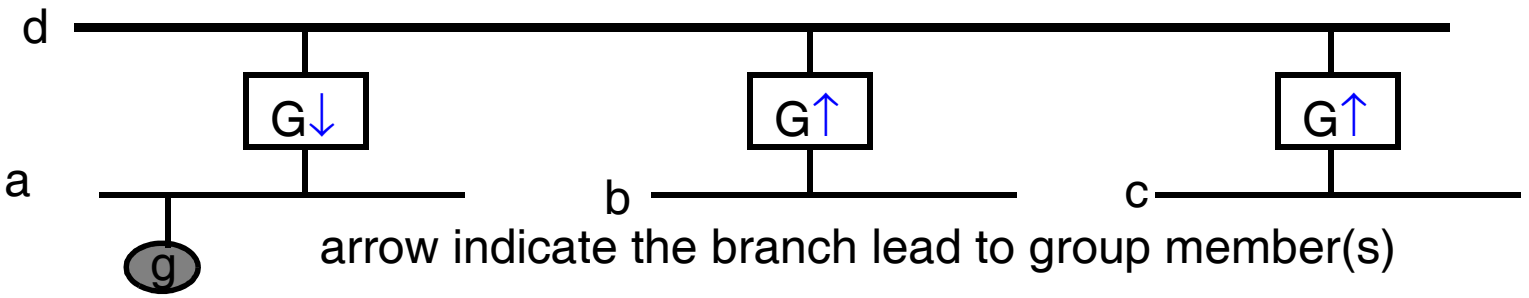


Figure 1. bridged LAN with one group member at LAN a

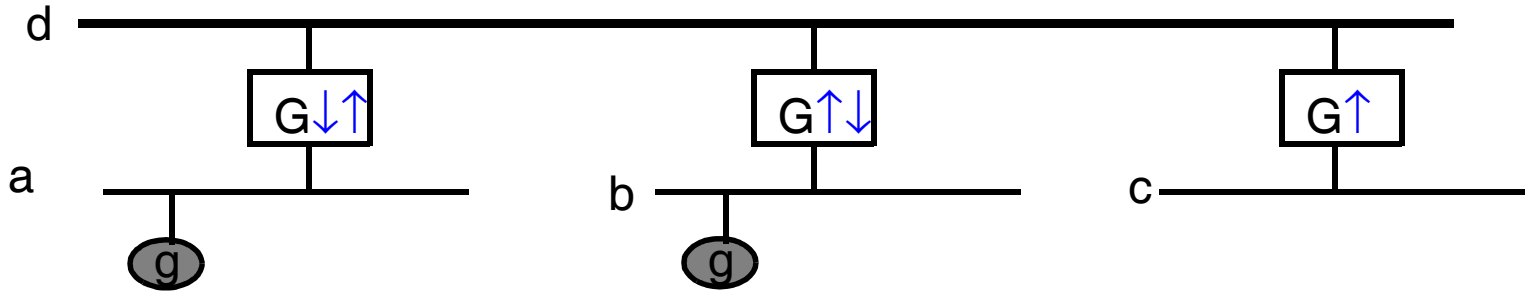


Figure 2. second group member join G at LAN b.

Extra space in the look-up table of learning bridges

- unicast entry (address, outgoing-branch, age)
- multicast entry (address, (outgoing-branch, age), (outgoing-branch, age),...)

The age fields in these different entries are interpreted differently.

- unicast msg is forward to all outgoing branches (except the incoming one) if age field expired.
- multicast msg is forward to non-expired branches. Those expired branches are considered not longer having group members.

Overhead of membership reporting

Let the membership reporting interval be T_{report},

- The expiry threshold for bridge table entries, T_{expire}, should be a multiple of T_{report} to avoid rare msg loss. Longer T_{expire} is not particular serious.
- To avoid low join latency (due to msg loss), issue several membership reports in close session when one first joins the group.
- If spanning tree changes, revert back to broadcasting style forwarding for a period of T_{expire}.

To reduce membership reporting traffic: T_{report} can be Longer (minutes rather seconds) or avoid multiple reports from the same LAN:

- Change the destination address of membership-report msg from all-B to G.
- The bridge then replace destination address field value from G to all-B.why?
- It allows other group members in the same LAN to recognize the report msg (same group address in src/dst fields) and to suppress their own reporting.
- Whenever observe the membership-report msg, reset the time-out with T_{report}+random value?
- The first time-out host will send membership-report msg.
- This reduces the reporting to one per group per LAN per Treport.

Cost of Single-Spanning Tree Multicast Routing

Example. A typical extended LAN with 10 segments, each host belongs to 5 groups, each segment has member of 20 different groups, there are 50 groups in total, and the membership reporting interval is T_{report} 200 seconds. Then,

- The host will send/receive one membership report packet every 40 seconds.
- A leaf segment generates one membership report packet every 10 seconds.
- On a non-leaf segment, the reporting traffic is one membership packet every second. The message includes the 20 local reporting messages and the reporting messages sent among bridges, 20+9*20=200 msgs in 200 seconds?
- The storage overhead is 50 group address entries per bridge.

The bridge multicast routing algorithm requires the hosts be modified to send membership reports. For non-conforming hosts, the connected bridges should allow to be inserted manually with membership information.

For non-conforming bridge, IP-tunnel is used to relay the multicast msg.

Distance-Vector Multicast Routing

The Ford-Fulkerson or Bellman-Ford algorithm has been used in old Arpanet and Xerox PUP Internet routing protocol. It is currently used in Xerox Network Systems Internetwork routers, some DARPA Internet core gateways and Berkeley Unix's *routed* internetwork routing process.

The routing table entry is
(Destination Distance Next-hop-address Next-hop-link, age)

Periodically, (Destination distance) fields, called distance vector, of the table are sent to neighbors

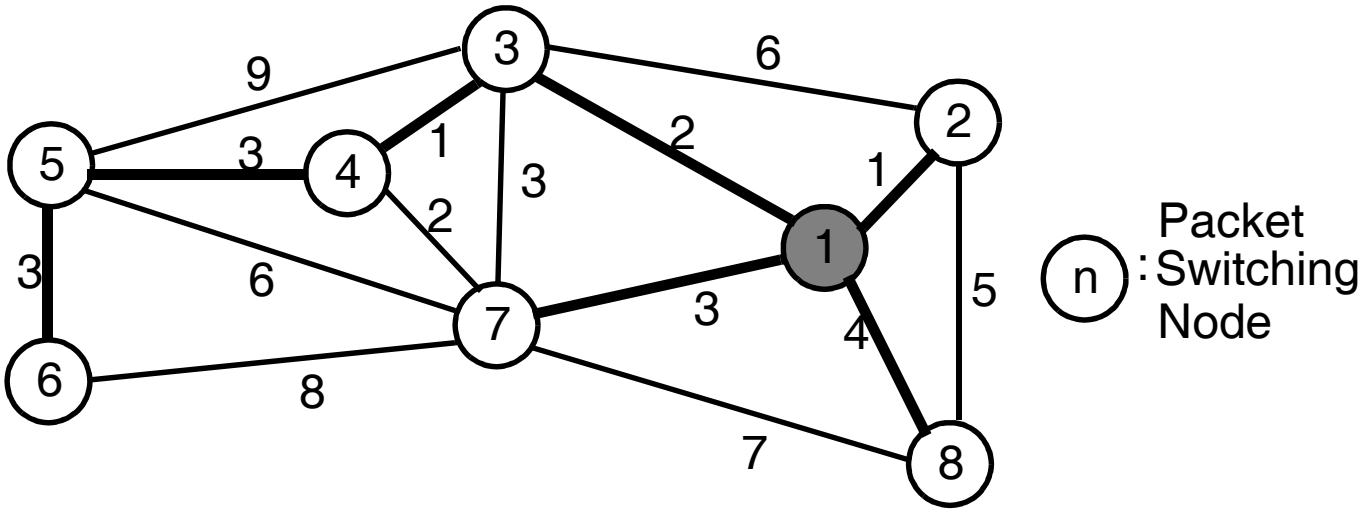
Based on neighbors' distance vectors a node decides the new routing table.

How to implement multicast routing in this environment?

Shortest-Path Multicast Tree

To meet the goal of low delay multicasting (as opposed to low cost multicasting in multicast path finding [chow91]), the multicast packet should traverse along the shortest-path multicast tree where each source-destination path is a shortest path. Observation:

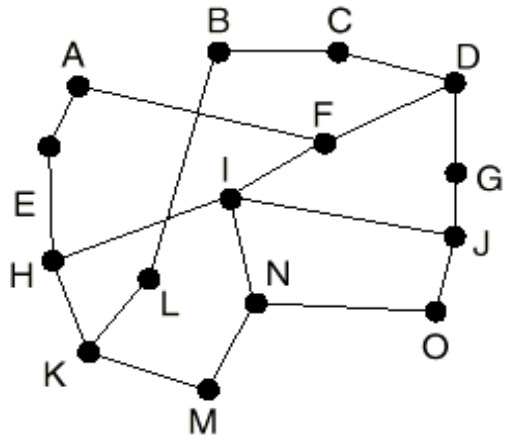
- A shortest-path multicast tree is a subset of a shortest-path broadcast tree rooted at the same source node.
- Is the minimum spanning tree a shortest-path broadcast tree?

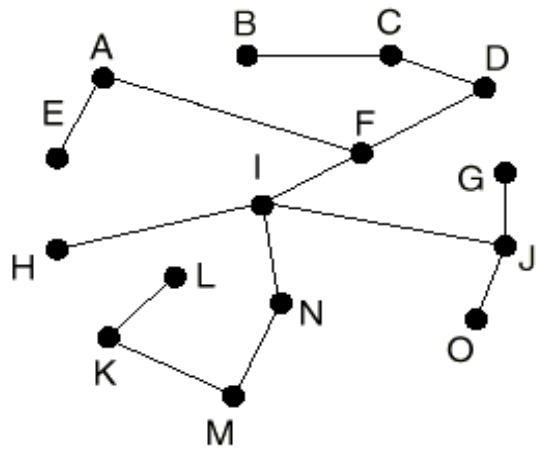


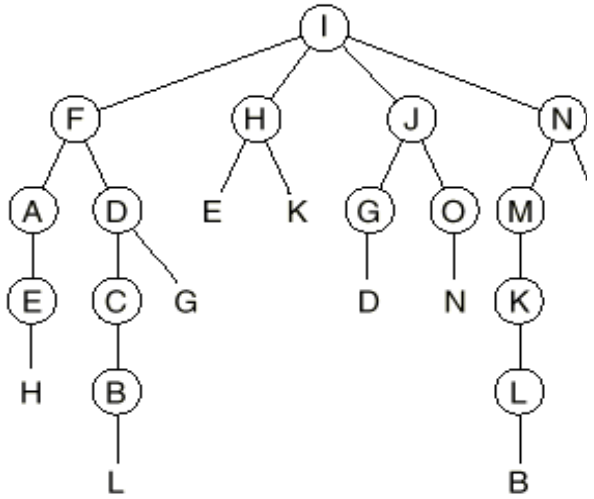
Reverse Path Forwarding

Dalal and Metcalfe's reverse path forwarding algorithm is a broadcast algorithm where the broadcast message will reach all the nodes along shortest paths.

- A router selective broadcasts a broadcast msg from a source, S, if it arrives on the link where the router sends point-to-point msg to S.
  Note that the distance vector routing entries to S form a shortest path tree rooted at S.
- If the broadcast msg arrives from other links, discard it.
  This reduces the # of broadcast msg copies.







Reverse Path Flooding (RPF)

- Use the reverse path forwarding technique
- extend to allow multicast addresses as destinations
- Hosts discard msgs belong to other group.

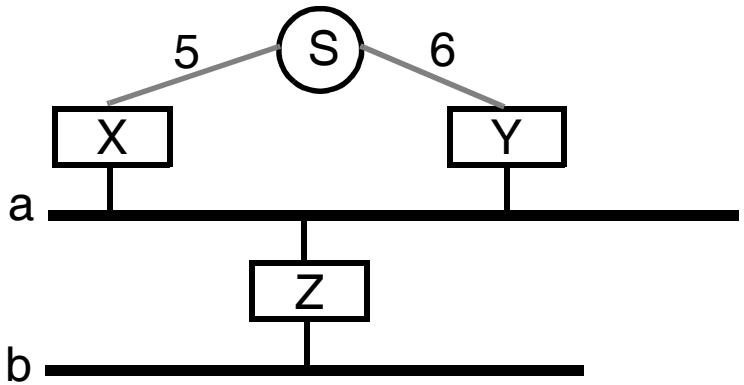
but

- a packet may appear in a link more than once, up to the number of routers that shared the link (It could be a multi-access link).
  -> instead of sending each router a msg, send a local multicast msg.
- a link may receive multiple copies of the same multicast msg, when it shares multiple routers.

Reverse Path Broadcasting (RPB)

To eliminate the duplicate broadcast packets generated by RFP, each router must be able to identify the child links in the shortest reverse-path tree rooted at any given source S. This information can be obtained by

1. [Dalal] Each router periodically sends "You are my next hop to these destinations."
2. [Deering] Select a single parent router for each link to each possible source, S.
   - A router can independently decide if it is the parent of a link by observing the distance vectors in the routing packets over that link.
   - Need new field, child, in the routing table entry. Child field is a bitmap where a bit is set if a corresponding link is a child link for the broadcast that is originated at *destination.*

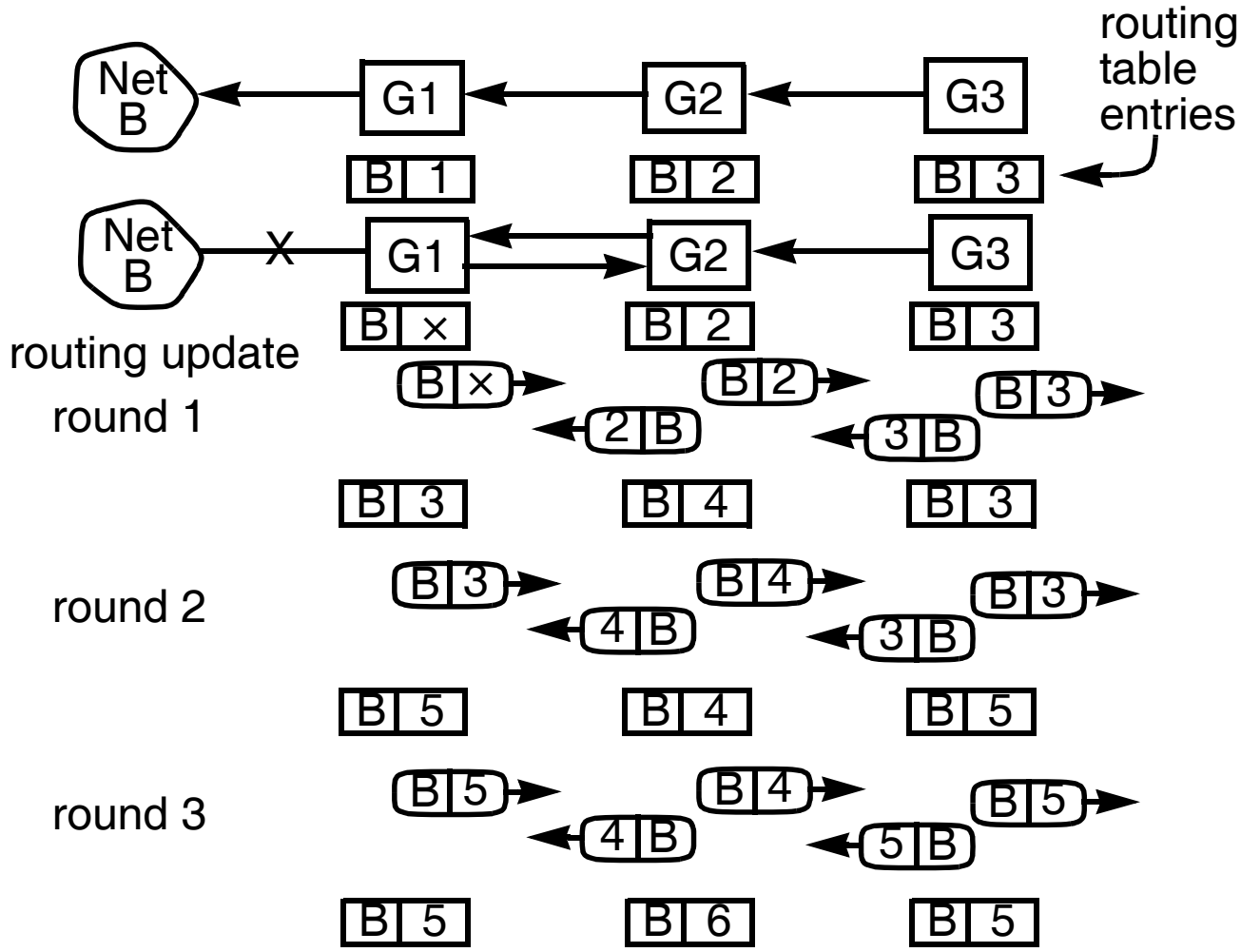


Truncated Reverse Path Broadcasting (TRPB)

Two basic approaches to use the RFP or RFB algorithms for multicast purpose:

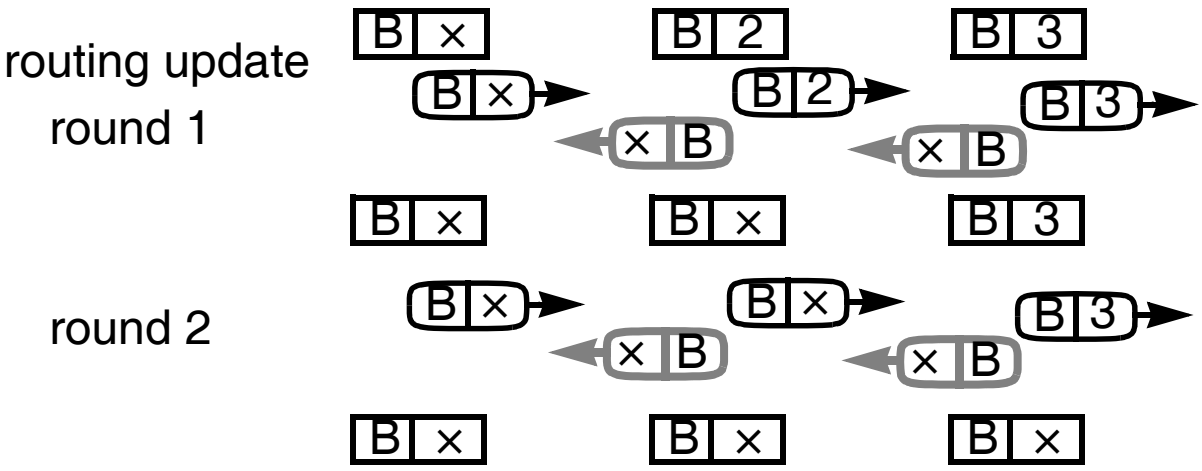
1. Hosts filter out unwanted messages.
2. Prune branches without a group member from the broadcast tree.
   - Let group members send membership report back up the broadcast tree to each source S periodically.
     —Branches without membership reports are deleted.
     —This has to be done for each S and each group → very expensive.
   - Prune leaf networks that do not have group members—TRPB
     A leaf network is identified by letting each router sends periodically "This link is my next hop to these destinations" msg along each link.
     The parent router will then know if the link is connected to a leaf network.
     Two ways to implement this:
     (1) Use routing information provided in *split horizon* technique.
     (See next viewgraph.)
     (2) Add extra bit to each (destination, distance) pair in the routing packet to indicate if this is the link over which the destination can be reached.
     A bit-map field, leaves is added to each entry, identifying which of the children links are leaf links.

The Slow Convergence Problem



Good news travels quickly; bad news travels slowly

Split Horizon Update



Truncated Reverse Path Broadcasting (TRPB)

The group membership mechanism is the same as the “alternative” technique mentioned in spanning tree multicasting algorithm.

The routers keep a list for each link on which groups is on the link.

The multicast packet from S to G will be forwarded to each child links for S except leaf links which have no members of G.

The overhead of TRPB:

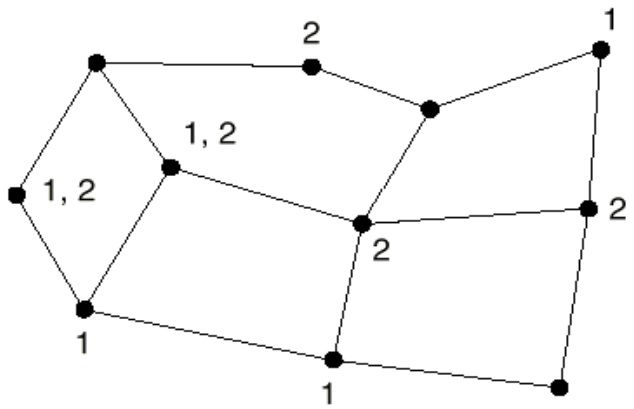
- For each routing entry, besides the children bitmap, add a leave bitmap which identifies leaf links.
- A bandwidth cost on each link of one membership report per group per reporting interval.
- There is no bandwidth cost for conveying the next hop information in the routing packet if split horizon technique is used.

Reverse Path Multicast (RPM)

This algorithm is based on TRPB and prunes the non-leaf no-group-member branches based on "non-membership reports," NMRs, from downstream routers.

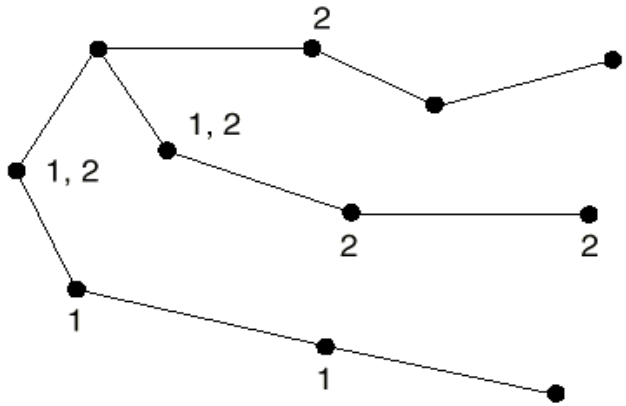
- The first multicast packet traverses along the shortest path broadcast tree.
- On receiving the multicast msg (S→G), a router whose all children links are leaf-links and no group member of G will send NMR back up the tree.
- For a non-leaf router, if all its child routers send up NMR and if its child links also have no members, it in turn sends an NMR back to its parent router.
- The subsequent multicast packets are blocked from travelling down those branches that send up NMRs.
- The NMR has an age field. When an NMR reaches T_{maxage}, it is discarded. The path pruned by this NMR will then rejoin the multicast tree.
- When new member appear on those pruned branches, a cancellation msg will send up by the router.
- For a router R, let n be the number of multicast sources active within T_{maxage}; g be the average number of group they send; r be the number of adjacent routers. The number of NMRs that R must store in worst case is ngr.

Example of Pruned multicast trees



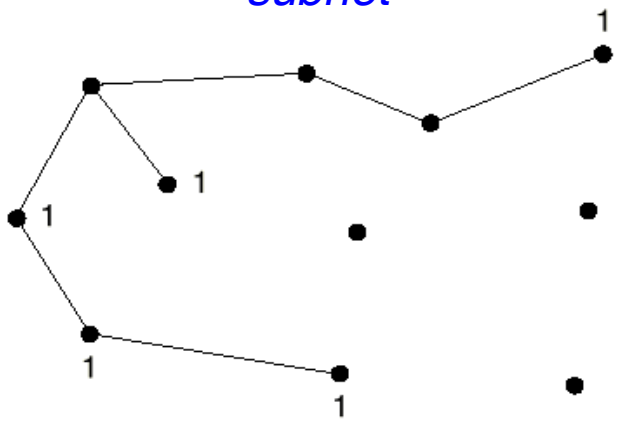
(a)

subnet



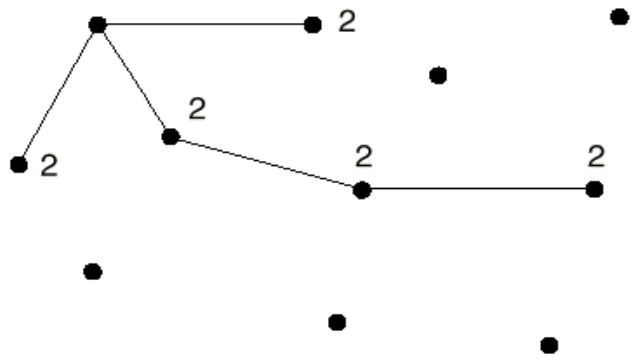
(b)

spanning tree of left most router



(c)

multicast tree for group1



(d)

multicast tree for group 2

Link-State Multicast Routing

It is based on the link-state routing algorithm used in Arpanet:

- When the state of a link changes, the routers attached to the link broadcast the information to all routers in the internet.
- With the complete topology info, each router computes the shortest path spanning tree, SPT, routed at the router using Dijkstra's algorithm.

Extend the link-state routing algorithm by

- include "groups on the link" in the link-state update msg.
- when there is a group change on the link, broadcast the link-state update.
- Given full knowledge on groups on each link, each router compute the shortest path multicast tree from any source to any group.
- reduce the redundant msgs by having a designated router for each links.

Compute the shortest path multicast tree on demand:

- each router keeps a cache of multicast routing records of the form:
  (source, subtree, (group, link-ttls)$^{+}$)
  where subtree is a list of all descendent links of the router on SPT routed S,
  link-ttls is a vector of time-to-live value, one for each descendent link, specifying the minimum TTL required to reach the nearest member.

Link-State Multicast Routing

Compute the shortest path multicast tree on demand:

- Only compute the multicast tree when info is not in the cache.
- Whenever the topology changes, all cache records are discarded.
- When group changes, the corresponding (group, link-ttls) fields are removed.

The major costs of the algorithm are

- in the memory required to store the cache record and the processing time required to compute the multicast trees.
  Assume that most multicast packets only traverse a small percentage of the routers, this algorithm will require less storage than the RPM algorithm.
- The first multicast packet will experience a long delay since each router needs to compute the multicast tree before forward the packet.

What are the implementation status of IDMR?

- A good tutorial in Internet-draft: draft-ietf-mboned-intro-multicast-00.txt (1/97)
  You can find it in ~cs622/i_draft
- Class D address
  "1110" followed by 28-bit multicast group ID.
  range from 224.0.0.0 to 239.255.255

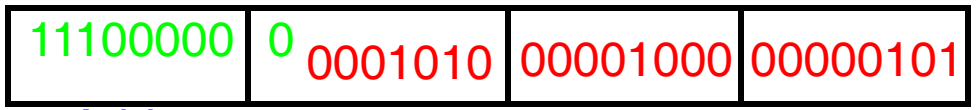
Table 1: Reserved Multicast Address

Group	Address	
all systems on this subnet	224.0.0.1	
all routers on this subnet	224.0.0.2	
all DVMRP routers	224.0.0.4	Use distance vector MR
all OSPF routers	224.0.0.5	Use Multicast ext. of OSPF (new link state routing)
all OSPF designated router	224.0.0.6	one per LAN segment
all RIP2 routers	224.0.0.9	Routing Info. Protocol
all PIM routers	224.0.0.13	Protocol Independent Multicast

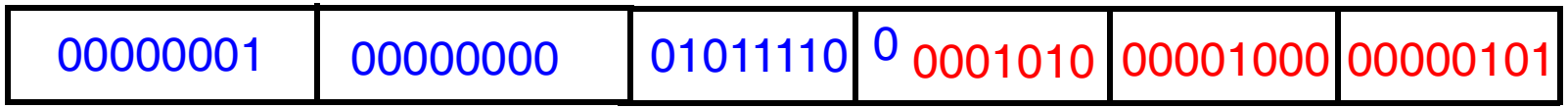
Mapping Class D to IEEE 802 MAC Address

- Internet Assigned Numbers Authority (IANA) were allocated with a reserved portion of IEEE-802 MAC layer multicast address space. (23 bits)

Class D IP address



IEEE 802 MAC Layer Multicast Address



- 5 bits of Class D IP address not mapped -> possible multiple to one mapping.

Internet Group Management Protocol (IGMP)

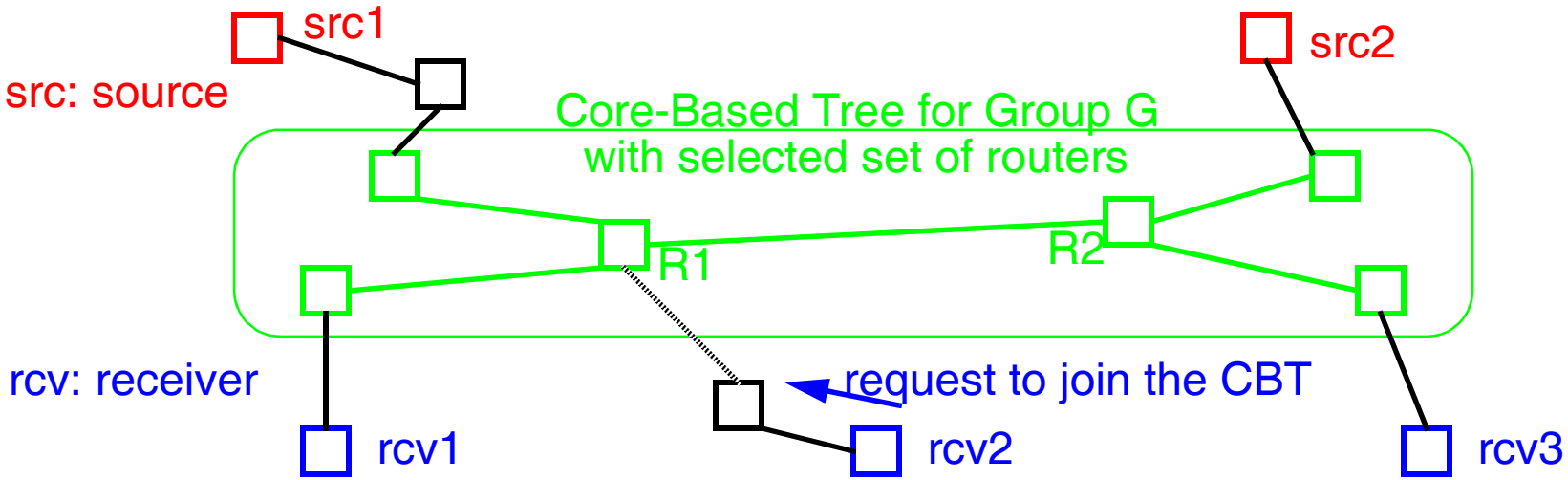
- IGMP runs between hosts and immediately-neighboring multicast routers.
- Hosts use IGMP to request the delivery of msgs for to multicast group.
- Routers use IGMP to query periodically the LAN for existence of groups.
- One of the routers on the LAN is elected to be the “querier”.
- Version 1 IGMP (RFC-1112):
  Query msg to “all-hosts group” (224.0.0.1) and TTL=1 (in IP header)
  (not forwarded to other routers)
  Hosts respond with Host Membership Report. (dest addr=G)
- Each host starts a randomly-chosen report delay timer for each group membership.
- Version 2 IGMP (draft-ietf-idmr-idmp-v2-05.txt) on version3.8 IP multicast code: define procedure for electing querier router
  - new Group-Specific Query msg for specific group.
  - new Leave Group msg for the last host to report its leave (early detection)
  - router responds by sending Group-Specific Query msg.
  - if no response and this is a leaf net, the interface is removed from delivery tree.

IGMP Version 3

- Add Group-Source Report Msgs so a host can elect to receive traffic from certain sources.
- Two different types: Inclusive Group-Source Report and Exclusive Group-Source Report.
- New Group-Source Leave msg allow to leave certain (source, group) pairs.
- IGMP only concerns with forwarding from routers to directed attached subnets.
- A Multicast Routing Protocol is responsible for constructing the multicast delivery tree.

Three classes of MR techniques

- Simpledminded flooding or spanning tree
- Source-Based Tree (SBT) techniques
  - RPB
  - TRPB
  - RPM
- “Shared-Tree” Techniques
  - Core-Based Tree (CBT)
  - Protocol Independent Multicast (sparse/dense mode) (PIM)



- rcv2 through R2 may be shorter to source 2. It becomes a semi-optimal tree.

Limiting the Scope of Multicasting

- Use TTL to limit the scope of multicasting
  - 1 for the same subnet
  - 15 for the same site
  - 63 for the same region?
  - 127 for the worldwide
  - 255 unrestricted
- Use Multicast address to limit the scope.
  - 239.0.0.0 to 239.255.255.255 are reserved for “administratively scoped applications”, e.g., multicast within a corporation.

Provide Integrated Service (IS) over Internet

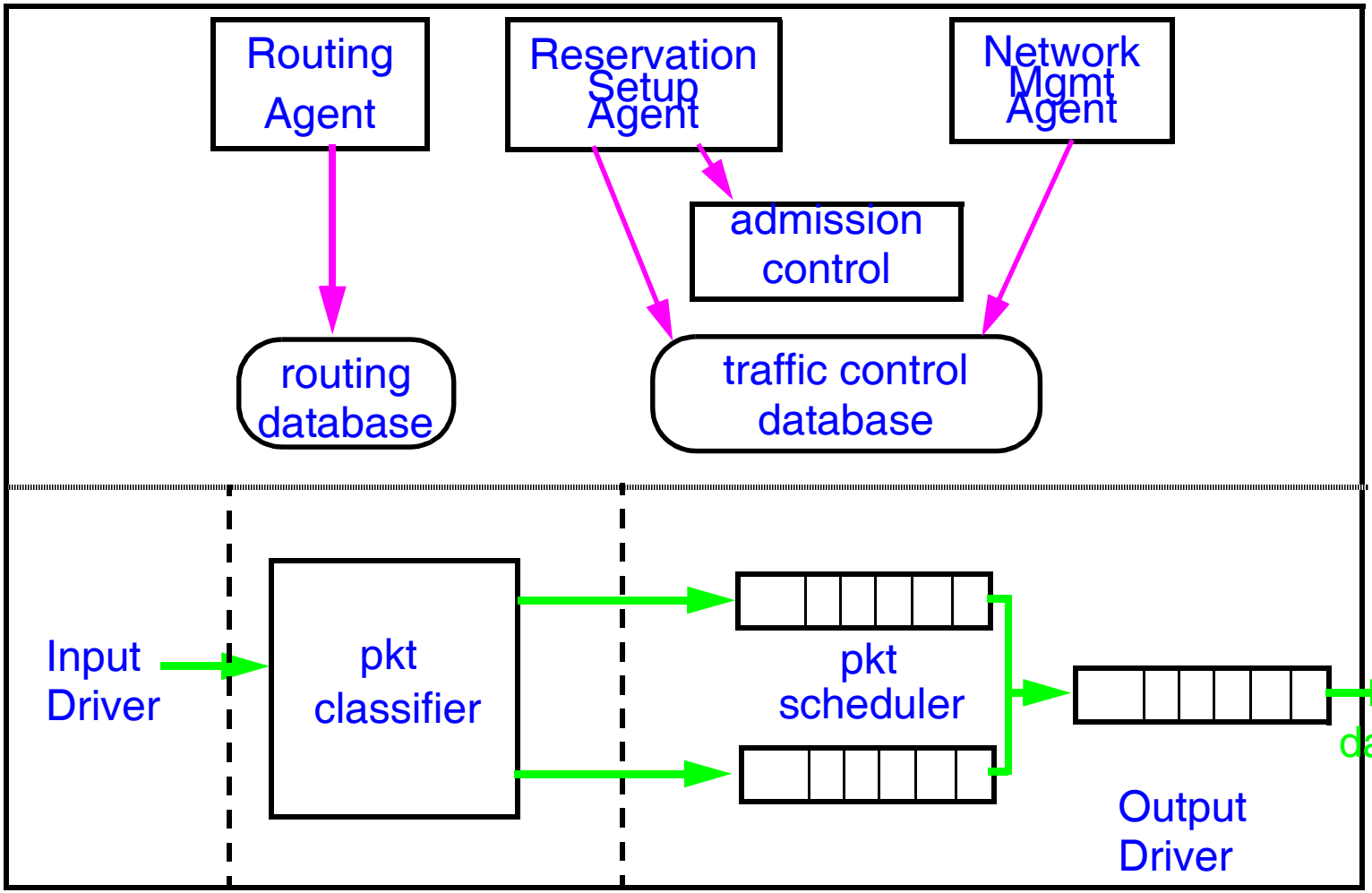
- Multimedia workstations/PCs everywhere
- Multicast backBONE Network exists
- Sophisticated digital audio/video applications appear
- How about the Quality of Service (QoS) provided by Internet:
  - still **best effort delivery unchanged 21 years [CerfKahn74]**
- Argument, Misconception, or Pitfall
  - "Bandwidth is infinite" (not in the near or medium term)
  - "Simple priority is sufficient" (priority is an implementation term, can not provide guarantee when every one sends high priority data)
  - "Application can adapt" (e.g. Jacobson's VAT audio program can adapt to network delay, but only to a degree, human sensors are very sensitive)
- We need **resource reservation** and **admission control** to get QoS.
- What kind of QoS? (For IS model see RFC1633)
  - guarantee? too strict?
  - statistical? too approximately?
  - predictable! Yes
- Need to control link sharing and control over end-to-end packet delay

IS over Internet: Network Assumptions

IS model proposed by RFC 1633:

- Common Infrastructure:
  - Use the same network infrastructure to support real-time and non-real time communication
- Use the existing IP for real-time data transport
  - instead of some real-time protocol such as ST2.
- Unified protocol stack approach
  - provides economy of mechanism
  - handle partial coverage (interoperability among IS-capable and non IS-capable systems) without the complexity of tunneling.
- Resource reservation means some users get privilege service. It requires
  - enforcement of policy
  - administrative control
  - authentication of users who make the reservation request
  - authentication of packets that use the reserved resources

Reference Implementation Model for Routers

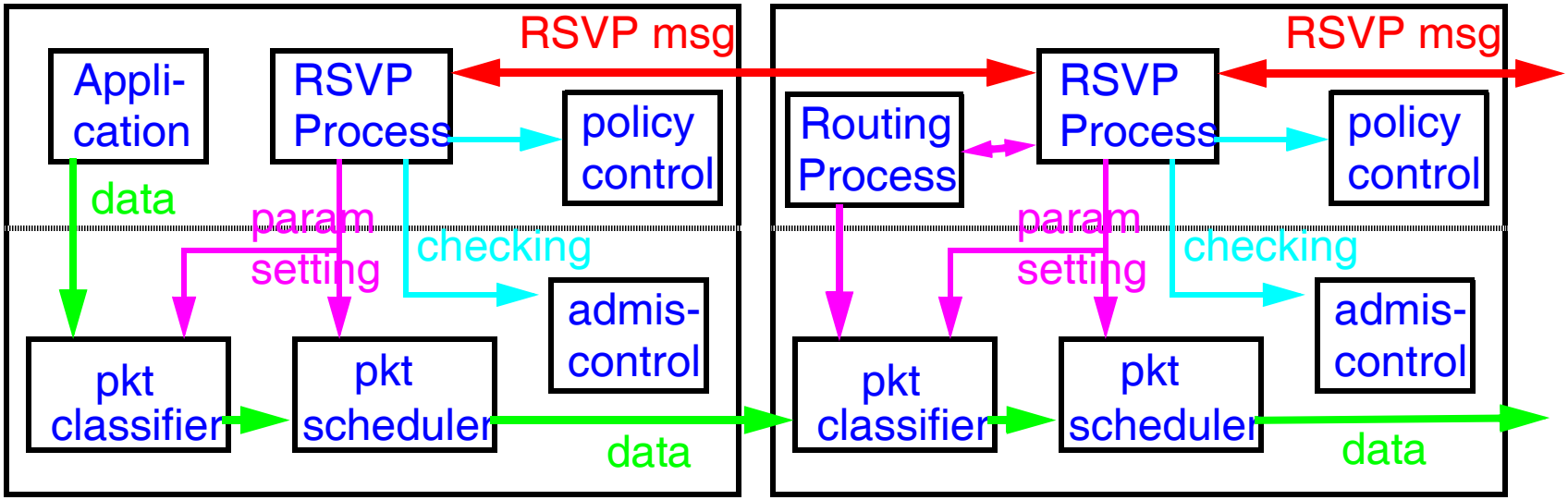


Resource reSerVation Protocol (RSVP)

- A resource reservation setup protocol for an Integrated Services (IS) Internet.
- Used by hosts to request a specific Quality of Service (QoS) for a data stream from a network.
- Used by routers
  - to deliver QoS control request to all nodes along the path of the flow,
  - to establish and maintain state to provide the requested service.
  - The request will generally, although not necessarily, result in resources being reserved in each node along the path (no guarantee?)
- **Flow**: a distinguishable data stream that requires QoS.
- RSVP requests resources for **simplex flow (uni-directional)**.
- RSVP operates on top of IP (IPv4 or IPv6)
- Not a data-transport protocol but an control protocol like ICMP, IGMP, routing
- RSVP operates with the unicast and multicast routing protocols
- Receiver is responsible for requesting QoS control (for scalability reason)
  - Receiver host application passes QoS request to local RSVP process
  - RSVP protocol carries request to all nodes along the path to the source.

RSVP Implementation on Host and Router

- A QoS node passes incoming data packets through a packet classifier.
- The packet classifier determines the route and QoS class for each packet
- On each outgoing interface, a packet scheduler make forward decision
  — to achieve promised QoS in the link-layer medium used by that interface.
- QoS control request passed to admission control and policy control modules
- Admission control determines if the node has sufficient resource.
- Policy control determines if the user has administrative permission
- If both succeed, parameters are set in packet classifier and scheduler.



RSVP "Soft State"

- RSVP router states are dynamically changed (due to membership or routing changes)
- If the state is stale or not adapt to the network status,
  $\rightarrow$RSVP will not be robust.
- The success of TCP/IP suite is attributed to that the state only maintains at end systems and that makes it robust.
- To preserve the robustness, RSVP use "soft state" concept:
  - RSVP sends periodic refresh messages to maintain the state
  - In absence of refresh, the state automatically times out and is deleted.

Reservation Model

- A RSVP reservation request consists of flow descriptor with two specs
  - flow spec: specifies desired QoS; used to set parameters in pkt scheduler
  - filter spec: defines the set of packets to receive the QoS; used to set parameters in pkt classifier
- Flowspec: consists of a service class and
  - Rspec (R for Reserve): defines the desired QoS.
  - Tspec (T for Traffic): describe the data flow.
- RSVP defines the session as a data flow with a particular destination, and transport protocol (e.g., identified by dst IP address and TCP/UDP)
- Filter spec: its format depends on whether IPv4 or IPv6 is in use.
  Most general case, it selects any subsets of the packets of a given session.
  The subset can be defined as
  - senders
  - higher layer protocol
  - any fields in any protocol headers

RSVP Filtering Problem

- For packet classification, routers need to examine protocol header fields, such as the UDP/TCP port numbers.
  $\longrightarrow$ Avoid IP fragmentation.
- Need facility to compute the minimum MTU over a multicast tree.
  and deliver that to senders.
- IPv6 has variable number of variable-length Internet-layer header before transport header
  $\longrightarrow$difficulty and cost in pkt classification
- IP-level security may encrypt the entire transport header, hiding them from intermediate routers.
  $\longrightarrow$extension to RSVP for IP security is being worked on.

One Pass With Advertising (OPWA)

- The basic RSVP reservation model is one pass:
  - A receive sends a request upstream
  - Each node along the path either accepts or rejects the request
  - There is no feedback to the receiver.
- One Pass With Advertising (OPWA) improve that by
  - Sending RSVP control packets (Path msgs) downstream, which follow the data paths, gather information for predicting the end-to-end QoS.
  - The results (advertisements) is delivered by RSVP to the receiver.
  - Allow the receiver to construct or dynamically adjust reservation requests.

Reservation Styles (Options)

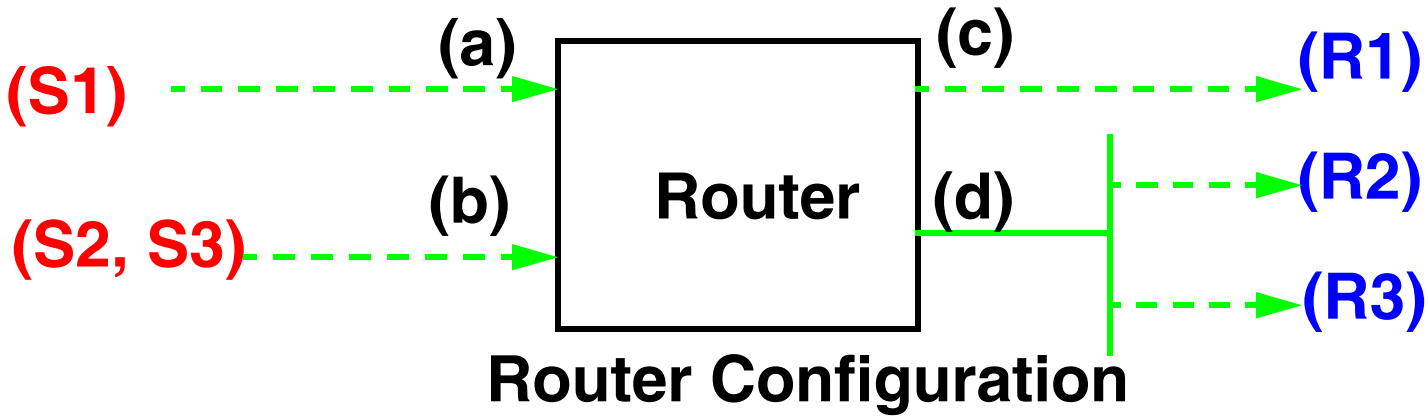
- Concern about the treatment of reservations for different senders within the same session
- Distinct reservation (for each sender) vs. Shared reservation (for selected senders)
- Explicit(list of selected senders) vs. Wildcard (implicitly select all the senders)

Table 2: Reservation Attributes and Styles

Sender Selection	Distinct	Shared
Explicit	Fixed-Filter (FF) style	Shared-Explicit (SE) style
Wildcard	(not defined)	Wildcard-Filter (WF) Style

- WF(* {Q}) where Q represents the flowspec.
- FF(S1{Q1}, S2{Q2},...)
- SE((S1, S2,...){Q})

Examples of Reservation Styles



Router Configuration

(a),(b),(c),(d) are interfaces of the router

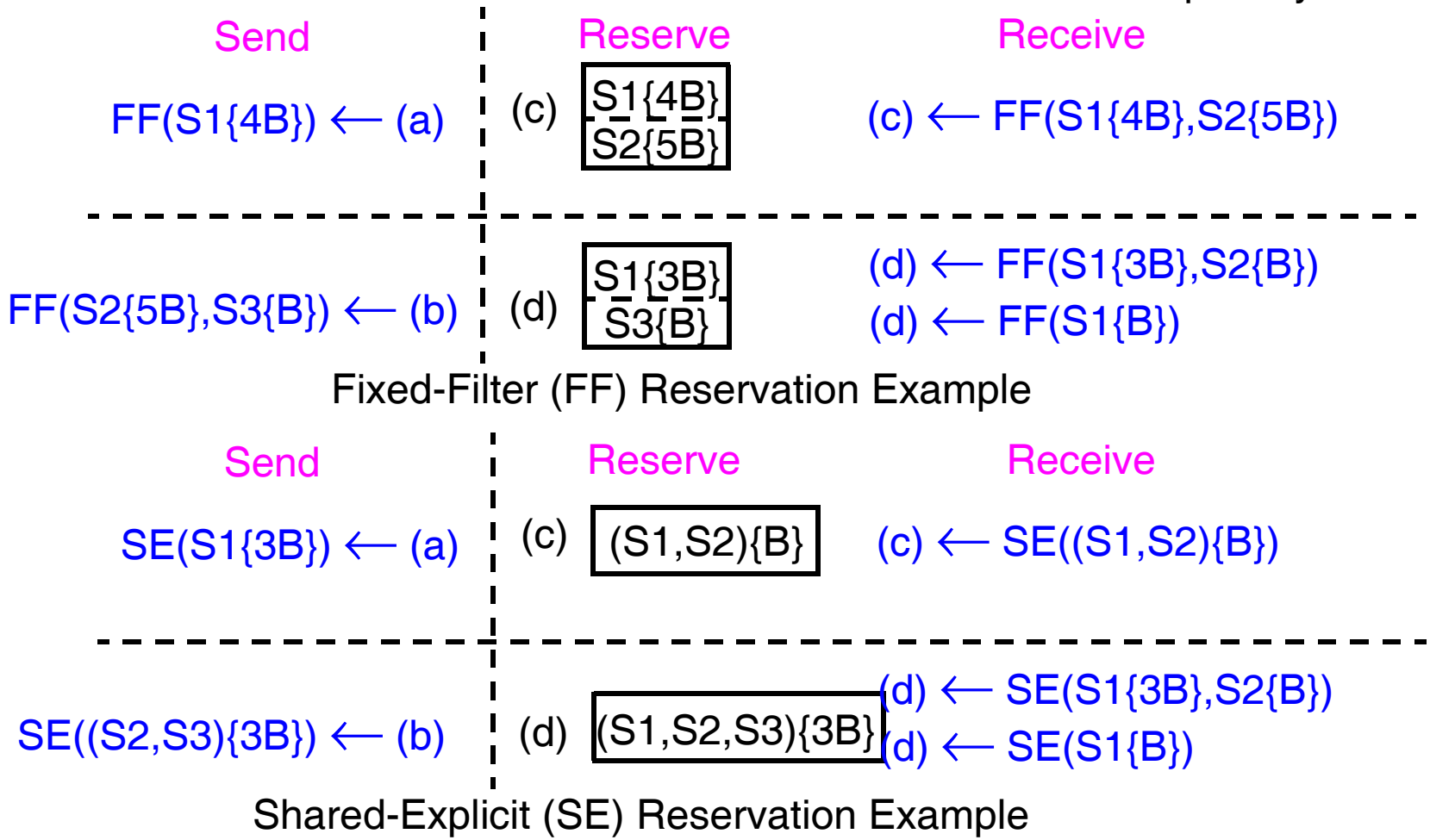
B: some resource quantity

Send	Reserve	Receive
WF(*{4B}) ← (a)	(c) * {4B}	(c) ← WF(*{4B})
WF(*{4B}) ← (b)	(d) * {3B}	(d) ← WF(*{3B}) (d) ← WF(*{2B})

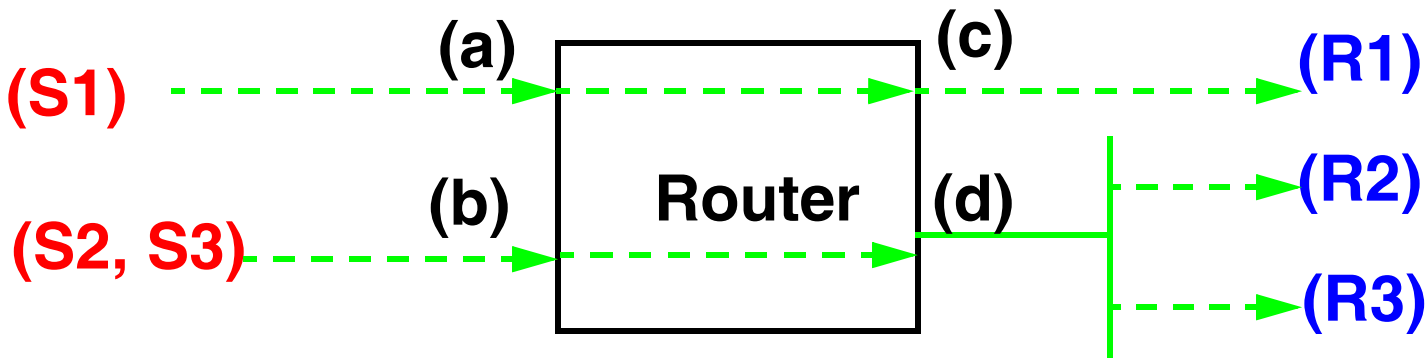
Wildcard-Filter (WF) Reservation Example

FF and SE Reservation Example

B: some resource quantity



Impact of Routing on Reservation Style Values

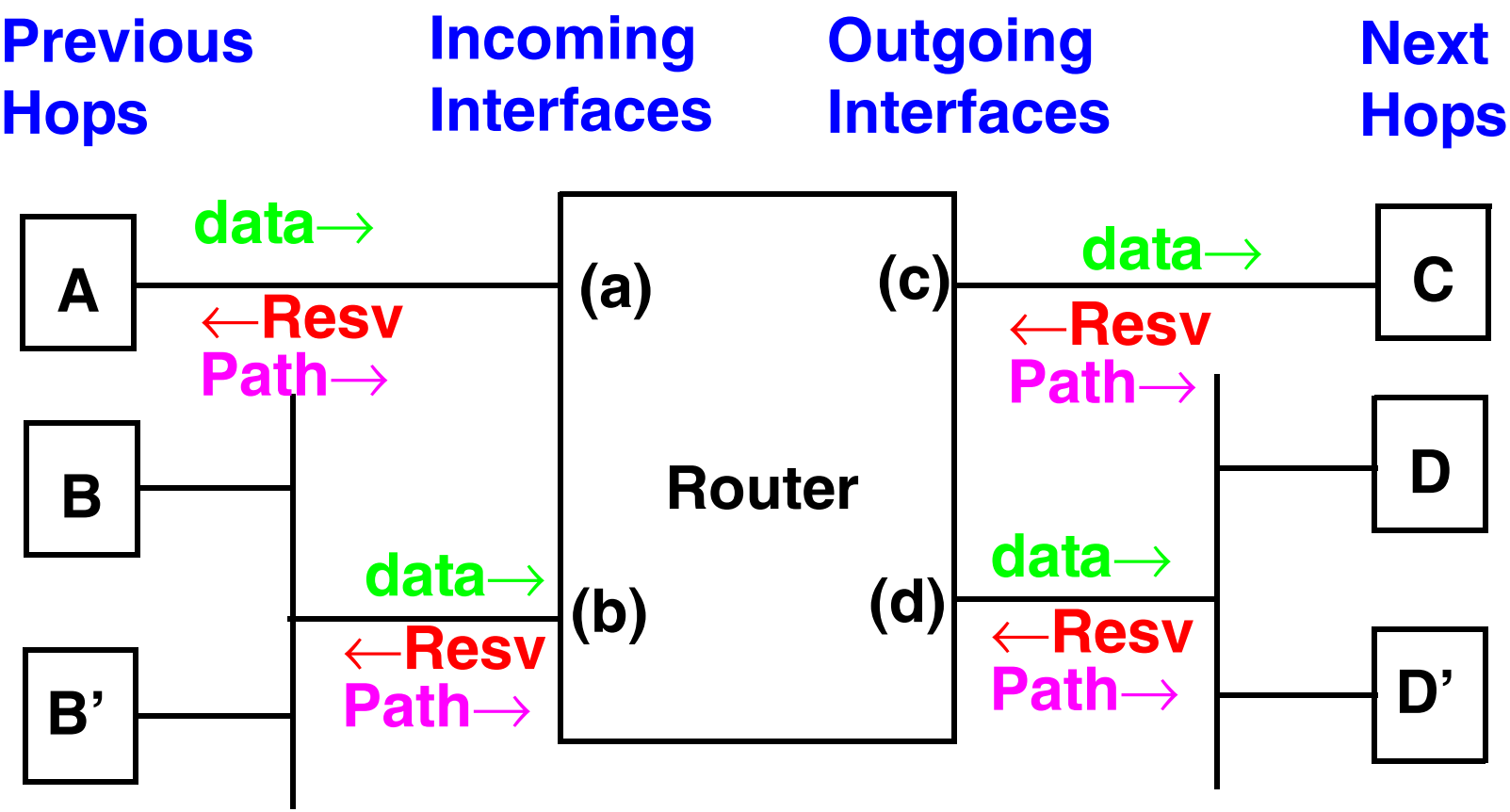


Assume that data from S2 and S3 are not forwarded to interface (c)
There is a better routes from (S2,S3) to R1.

Send	Reserve		Receive
WF(*{4B}) ← (a)	(c)	* {4B}	(c) ← WF(*{4B})
WF(*{3B}) ← (b)	(d)	* {3B}	(d) ← WF(*{3B}) (d) ← WF(*{2B})

Wildcard-Filter (WF) Reservation Example

RSVP Protocol Mechanism



- Resv and Path are the two fundamental message types.
- Resv msgs traverse upstream, create/edit reservation state, reach sources.
- Each sender transmits Path msgs downstream following data paths.

Path Message

Path message contains the following information:

- Sender Template
  - describe the format of data packet
  - use the filter spec format of the Resv msg.
  - include sender IP address, (optionally UDP/TCP sender port)
  - with protocol ID of the session
- Sender Tspec
  - defines the traffic characteristics of the data flow that sender will generate.
  - It is used by traffic control to prevent over-reservation, unnecessary admission control failure.
- Adspec = OPWA advertising information
  - it is passed to the local traffic control, which return an updated adspec
  - the updated adspec is then forwarded in Path msgs downstream
- For detail current RSVP Version 1 Specification, see draft-ietf-rsvp-spec-14.ps in ~cs622/i_draft