CS 356: Computer Network Architectures

Lecture 17: IP Multicast
[PD] Chapter 4.2

Xiaowei Yang
xwy@cs.duke.edu

Overview

- Multicast routing protocols
- Challenges

https://www.cisco.com/c/en/us/about/press/internet-protocol-journal/back-issues/table-contents-3/ipj-archive/article09186a00800c851e.html

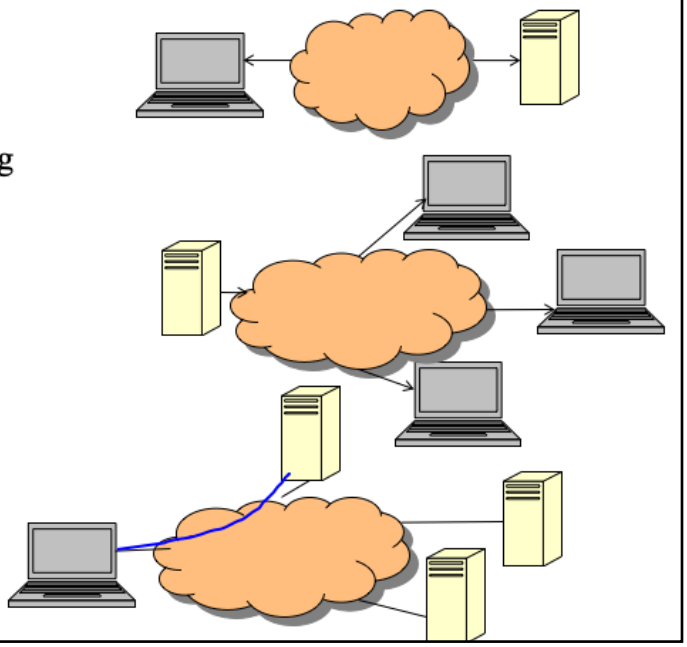
IP Multicast

What is Multicast

- Many-to-many communications
- Applications
  - Internet radio
  - Video conferencing
  - News dissemination

Communication models

- Unicast
  - One-to-one
  - Unicast routing
- Multicast
- Anycast
- Broadcast



1. Internet radio, tv broadcast, pub/sub, news, stock quotes..

2. Packet duplication, app needs to track all recipients

Design questions

- How does a sender know who is interested in the packet?
- How to send a packet to each receiver?

Multicast Architecture

- Nodes interested in many-to-many communications form a multicast group
- Each group is assigned a multicast address
- Routers establish forwarding state to multicast addresses
- Members of a multicast group receive packets sent to the group's multicast address

Group Management

- Routers maintain which outgoing links connect to multicast group members
- A host signals to its local router its desire to join or leave a group
  - Internet Group Management protocol (IPv4)
  - Multicast Listener Discovery (IPv6)

Multicast Addresses

- IPv4: 224.0.0.0/4 (28 bits)
- IPv6: 1111 1111 / 8

- Mapping an IP multicast address to an Ethernet multicast address
  - 01-00-5E-00-00-00 to 01-00-5E-7F-FF-FF
  - Internet Multicast [RFC1112]
  - Map the lower-order 23-bit IP address to Ethernet multicast address
  - No ARP for multicast

- IPv6 has a similar mapping scheme

http://www.iana.org/assignments/ethernet-numbers

In the normal Internet dotted decimal notation this is 0.0.94 since the bytes are transmitted higher order first and bits within bytes are transmitted lower order first.

33-33-00-00-00-00 to 33-33-FF-FF-FF-FF are used for IPv6 multicast

Router forwarding algorithm

- if IP-destination is on the same local network
  or IP-destination is a host group, send
  datagram locally to IP-destination
- else
  - send datagram locally to NextHop (IP-destination
)

Receiving a Multicast Packet

- Host configures the network adaptor to listen to the multicast group
- Examine the IP multicast address, and discard packets from non-interested groups

Types of multicast

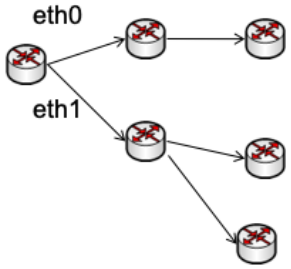
- Any source multicast
  - Many-to-many
  - A receiver does not specify a sender
- Source specific multicast
  - A receiver specifies both the group and the sender
  - TV, radio channels

Design questions

- How does a sender know who is interested in the packet?
  - Receivers join the group
  - Sends to a multicast group
  - Routers maintain the group membership
- How to send a packet to each receiver?
  - Unicast?
  - Flooding?

Multicast routing

224.16.0.10	eth0
	eth1



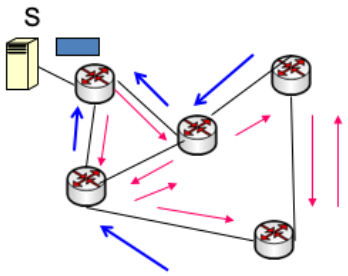
- Multicast distribution trees: multiple outgoing interfaces for a multicast destination address

Distance Vector Multicast Routing Protocol

- Using existing distance vector routing protocol

- Establish multicast forwarding state
  - Flood to all destinations (reverse path flooding)
    - Key design challenge: loop-avoidance
    - Q: how many broadcast loop-avoidance mechanisms have we learned?
  - Prune those not in the group

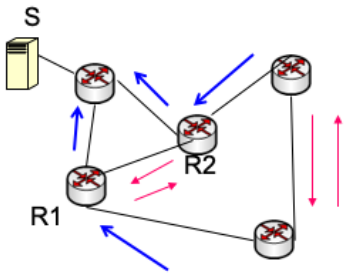
Reverse path flooding



- Reverse shortest-path flooding
  - If packet comes from link L, and next hop to S is L, broadcast to all outgoing links except the incoming one
- Packets do not loop back
  - Why?

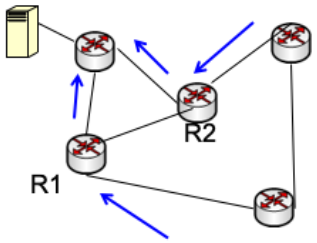
Thus, whenever it receives a multicast packet from source S, the router forwards the packet on all outgoing links (except the one on which the packet arrived) if and only if the packet arrived over the link that is on the shortest path to S (i.e., the packet came from the NextHop associated with S in the routing table).

Problems with RPF



- Problems
  - multiple routers on a LAN → receiving multiple copies of packets
  - Not all hosts are in the multicast group. Broadcast is a waste

Designated router election

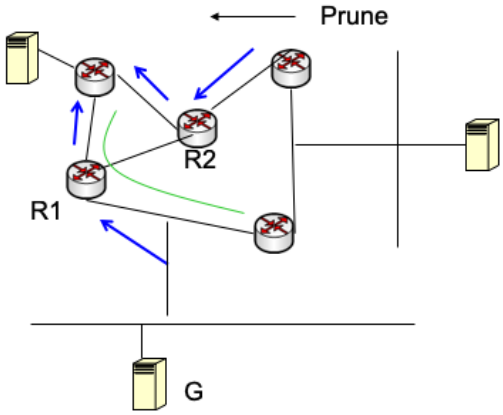


- Address the duplicate broadcast packet problem
- Routers on the same LAN elect a parent that has the shortest distance to S
  - Parent is one with the shortest path
    - Routers can learn this from DV routing messages
  - If tie, elect one with the smallest router ID

Truncated reverse path flooding

- Start with a full broadcast tree to all links (RPB)
- Prune unnecessary links
  - Hosts interested in G periodically announce membership
  - If a leaf network does not have any member, sends a prune message to parent
    - Augment distance vector to propagate groups interested to other routers
    - Only do so when S starts to multicast
  - This prune message can be propagated from router to router to prune non-interested branches

A pruning example



Protocol Independent Multicast (PIM)

- Problem with DVMRP
  - Broadcast is inefficient if few nodes are interested
  - Most routers must explicitly send prune messages
  - Dependent on routing protocols
- Solution
  - Dense mode: flood & prune similar to DVMRP
  - Sparse mode: send join messages to rendezvous point (RP)
  - Not dependent on any unicast routing protocol, unlike DVMRP

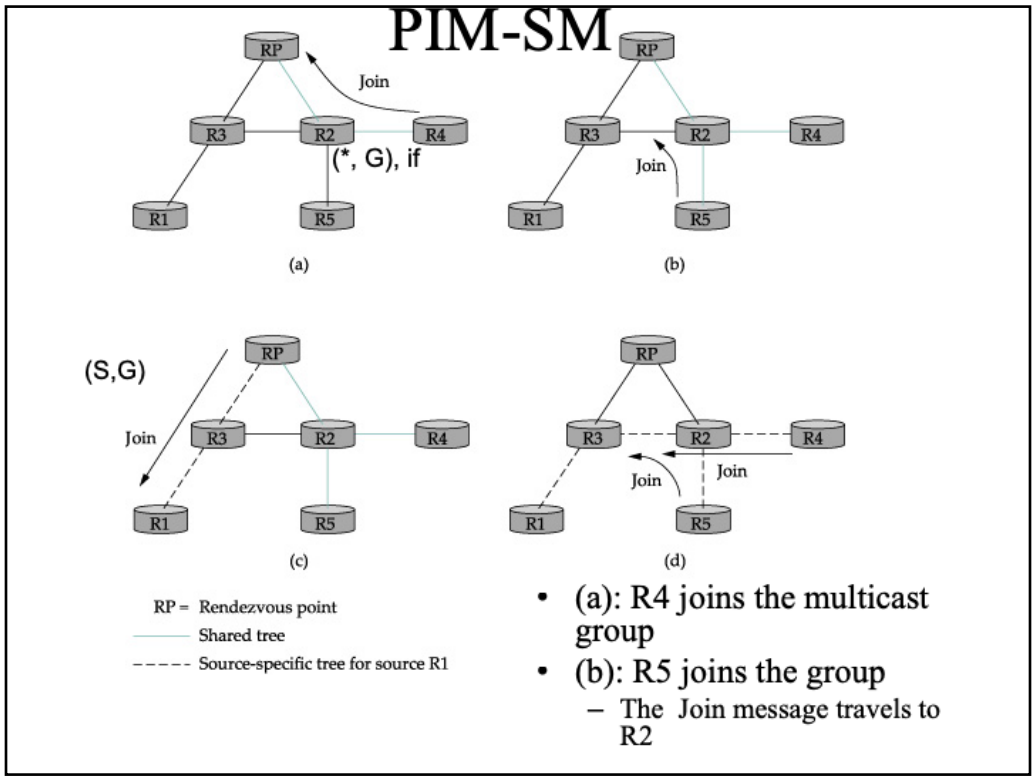
PIM-SM

1. Routers explicitly join a shared distribution tree
   - Unlike DVMRP which starts from a broadcast tree

2. Source-specific trees are created later for more efficient distribution if there is sufficient traffic

Join

- PIM-SM assigns each group a special router known as the rendezvous point (RP)
- A router that has hosts interested in G sends a Join message to RP
- A router looks at the join message and create a multicast routing entry (*,G) pointing to the incoming interface. This is called an all sender forwarding entry
- It propagates join to previous hop closer to RP



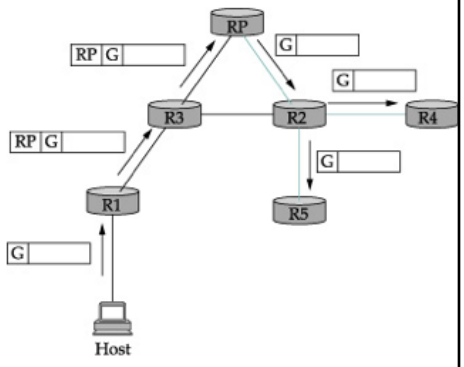
After B, a host attached to R1 can send to the tree by tunneling the packet to RP first. But it's inefficient. So RP will send a source specific Join to R1. RP sends a source specific joint so R1 can sends with native multicast. R4, R5, can send (S,G) to build a source specific tree.

To do so, it constructs a packet with the appropriate multicast group address as its destination and sends it to a router on its local network known as the designated router (DR). Suppose the DR is R1 in Figure 4.14. There is no state for this multicast group between R1 and the RP at this point, so instead of simply forwarding the multicast packet, R1 tunnels it to the RP. That is, R1 encapsulates the multicast packet inside a PIM Register message that it sends to the unicast IP address of the RP. Just like a tun- nel endpoint of the sort described in Section 3.2.9, the RP receives the packet addressed to it, looks at the payload of the Register message, and finds inside an IP packet addressed to the multicast address of this group. The RP, of course, does know what to do with such a packet—it sends it out onto the shared tree of which the RP is the root. In the example of Figure 4.14, this means that the RP sends the packet on to R2, which is able to forward it on to R4 and R5. The complete delivery of a packet from R1 to R4 and R5 is shown in Figure 4.15. We see the tunneled packet travel from R1 to the RP with an extra IP header containing the unicast address of RP, and then the multicast packet addressed to G

making its way along the shared tree to R4 and R5.

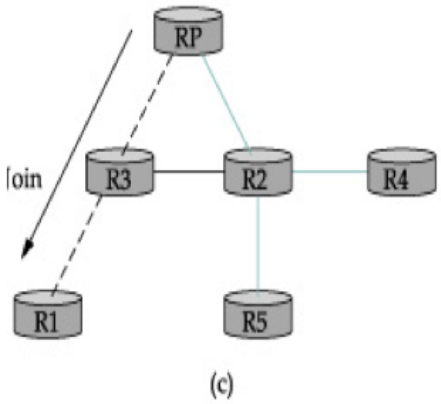
Forwarding along a shared tree

- If a source S wishes to send to the group
  - S sends a packet to its designated router (R1) with the multicast group as the destination address
  - R1 encapsulates the packet into a PIM register message, unicast it to RP
- PR decapsulates it and forwards to the shared trees



Source specific tree

- Problems
  - Encapsulation is inefficient
- Solution:
  - RP sends Join message to source S
  - R3 now knows the group (S,G)



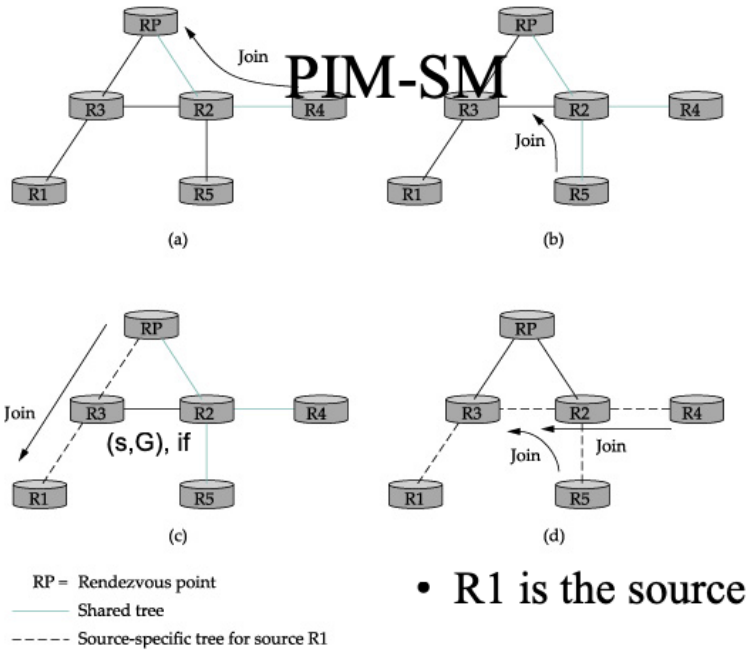
A router state (S,G) is created. Native multicast packets can be sent to RP.

Source specific tree

- Problem: shared trees are inefficient as paths could be longer than shortest path
- Solution
  - If s sends at high rates, routers send source-specific Join messages
  - Trees may no longer involve RP

Later, R4 and R5 sent join messages so each router creates a (s,G) forwarding state. RP Is not involved anymore.

PIM-SM



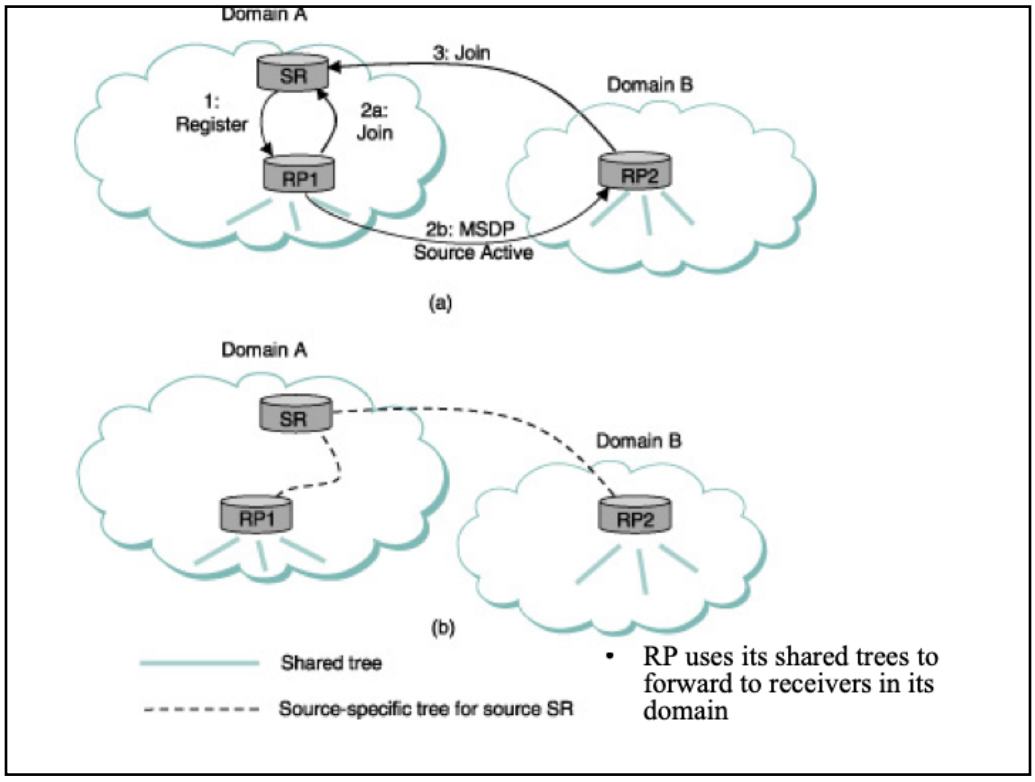
- R1 is the source

PIM: final remarks

- Unicast independent
  - Assuming a unicast routing protocol has established correct forwarding state
- Scalability challenges
  - Per (S,G) forwarding state!

Inter-domain multicast

- Problem: how can the entire Internet agree on a single RP for a group G?
- Multicast Source Discovery Protocol
  - Hierarchical
  - Intra-domain: PIM-SM
  - Inter-domain: a distribution tree among all domain's RPs



A hierarchical design similar to intra- and inter-domain routing protocols

Each domain runs PIM-SM internally

RPs of each domains form an overlay mesh using TCP connection (similar to BGP sessions)

An RP periodically broadcasts active sources to peer RPs: (S,G)

Reverse path forwarding

A peer RP that has active receivers sends join to S on behalf of the receivers

If an MSDP peer RP that receives one of these broadcasts has active receivers for that multicast group, it sends a source-specific Join, on that RP's own behalf, to the source host, as shown in Figure 4.16(a). The Join message builds a branch of the source-specific tree to this RP, as shown in Figure 4.16(b). The result is that every RP that is part of the MSDP network and has active receivers for a particular multicast group is added to the source-specific tree of the new source. When an RP receives a multicast from the source, the RP uses its shared tree to forward the multicast to the receivers in its domain.

Source-specific multicast (PIM-SSM)

- One-to-many
  - Considered more common than many-to-many

- Channel: (S,G)
- Hosts join a channel
- Join messages are propagated to S to create a source specific tree
- Only S can use the tree

- Advantages
  - More efficient distribution than shared tree
  - More multicast groups
  - More secure: only S can send
  - No need for MSDR

Remarks on IP multicast

- Many design choices
- Facing many challenges: used to be a very active resource topic
  - Economic model's not clear: who pays for the service?
  - Security
  - Reliability
  - Scalability
  - Heterogeneity

https://tools.ietf.org/html/rfc3170

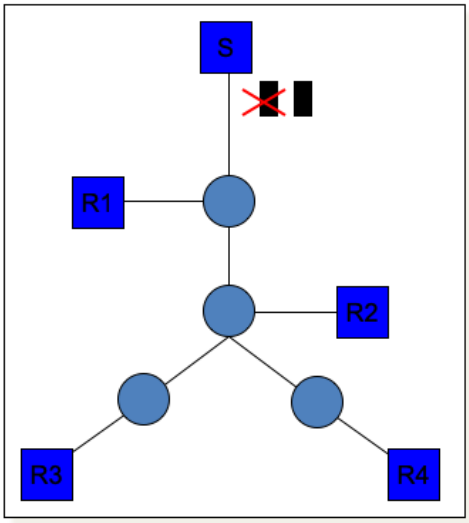
https://en.wikipedia.org/wiki/IP_multicast#Development

Reliable multicast

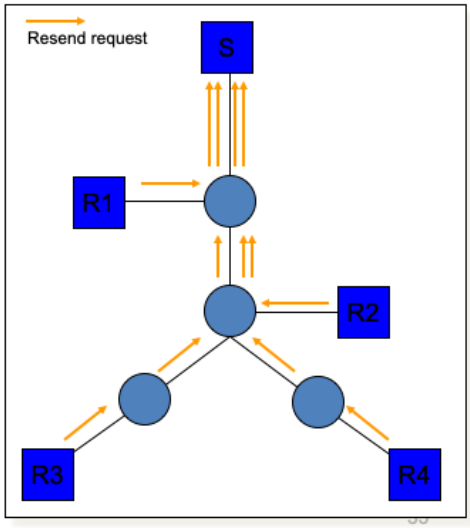
- Problems
  - Acknowledgment implosion
  - Retransmission exposure

Implosion

Packet 1 is lost



All 4 receivers request a resend



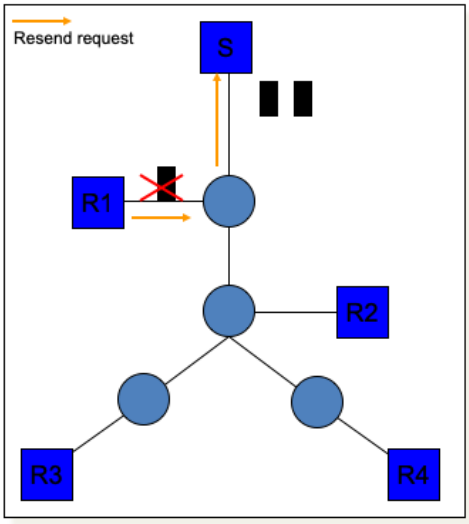
Retransmission

- Re-transmitter
  - Options: sender, other receivers
- How to retransmit
  - Unicast, multicast, scoped multicast, retransmission group, …
- Problem: Exposure

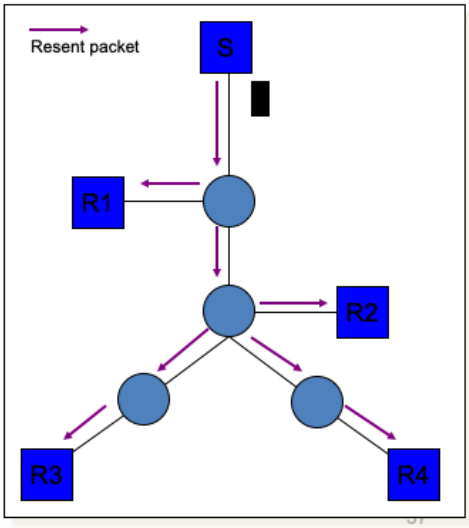
36

Exposure

Packet 1 does not reach R1;
Receiver 1 requests a resend

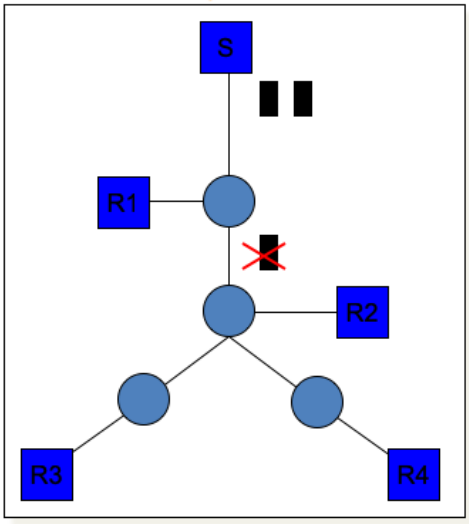


Packet 1 resent to all 4 receivers

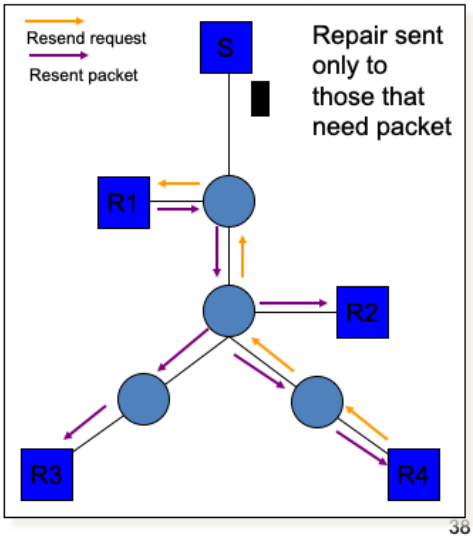


Ideal Recovery Model

Packet 1 reaches R1 but is lost before reaching other Receivers



Only one receiver sends NACK to the nearest S or R with packet

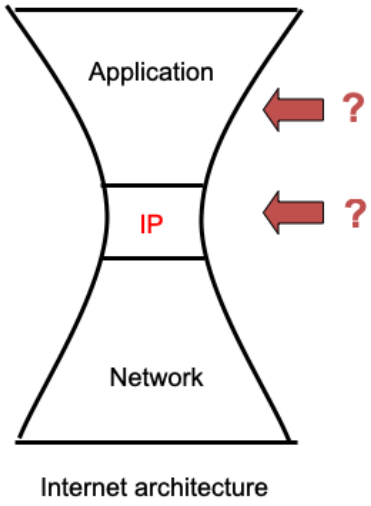


38

Multicast Challenges

- Reliability
- Scalability
- Heterogeneity

Supporting Multicast on the Internet



Internet architecture

- At which layer should multicast be implemented?
- Can routers afford (s,G) state?
- Who pays to create a multicast group?
  - Botnets

40

In the hourglass Internet architecture,

-IP is the compatibility layer in the Internet architecture.

-All hosts must implement IP

-Two choices

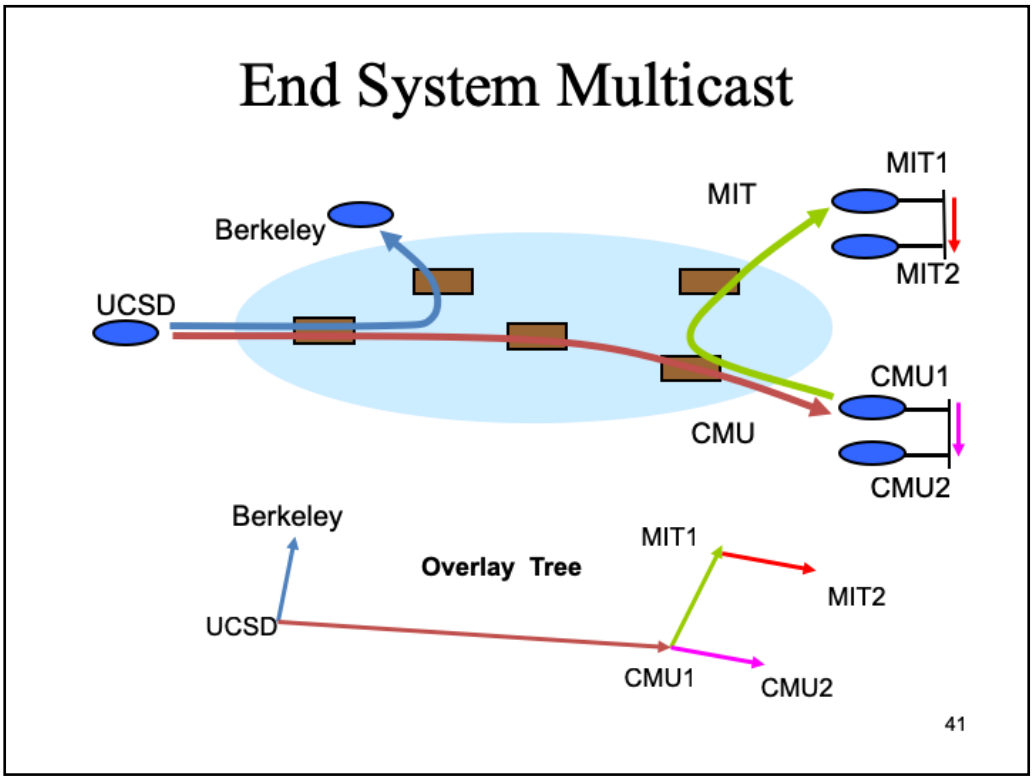
-multicast at IP

-or application: only a subset, customizability

One important architecture question is, at which layer should multicast be implemented.

The convention wisdom has been to support multicast in the IP layer for efficiency and performance reasons. However, more than 10 years since this is proposed, it still has not been widely deployed.

This paper revisits this question with emphasis on Internet evaluation. In particular, we show that multicast at the application layer can be efficient compared to IP Multicast.



Recently, we and others have advocated for an alternative architecture, where all multicast functionality, including pkt replication and group management are pushed to end systems.

- We call this architecture End System Multicast

- In this architecture, end system organize themselves into an overlay tree root at the source

- data is sent along the overlay tree.

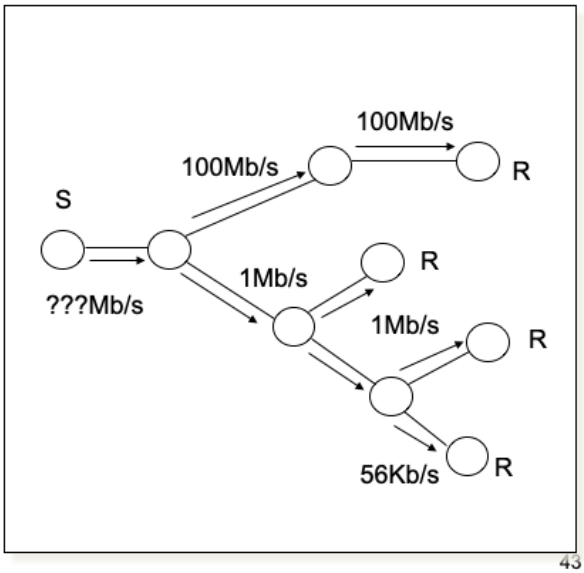
- It is an overlay in the sense that each link in the overlay tree corresponds to a physical path in the underlying network

Multicast Challenges

- Reliability
- Scalability
- Heterogeneity

Multicast sending rates

- What if receivers have very different bandwidths?
- Send at max?
- Send at min?
- Send at avg?



43

Video Adaptation: RLM

- Receiver-driven Layered Multicast
- Layered video encoding
- Each layer uses its own multicast group
- On spare capacity, receivers add a layer
- On congestion, receivers drop a layer
- Join experiments used for shared learning

44

Layered Media Streams