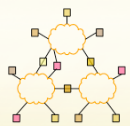
15-446 Distributed Systems Spring 2009



L-20 Multicast

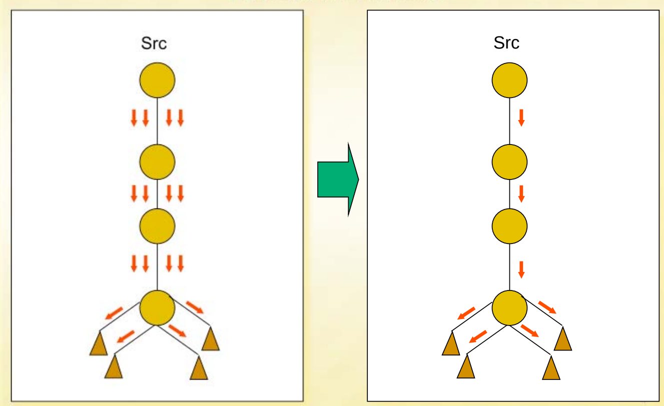
Multicast Routing

- Unicast: one source to one destination
- Multicast: one source to many destinations
- Two main functions:
  - Efficient data distribution
  - Logical naming of a group

Overview

- What/Why Multicast
- IP Multicast Service Basics
- Multicast Routing Basics
- DVMRP
- Reliability
- Congestion Control
- Overlay Multicast
- Publish-Subscribe

Multicast – Efficient Data Distribution



Multicast Router Responsibilities

- Learn of the existence of multicast groups (through advertisement)
- Identify links with group members
- Establish state to route packets
  - Replicate packets on appropriate interfaces
  - Routing entry:

Src, incoming interface	List of outgoing interfaces

5

Logical Naming

- Single name/address maps to logically related set of destinations
  - Destination set = multicast group
- How to scale?
  - Single name/address independent of group growth or changes

6

Multicast Groups

- Members are the intended receivers
- Senders may or may not be members
- Hosts may belong to many groups
- Hosts may send to many groups
- Support dynamic creation of groups, dynamic membership, dynamic sources

7

Scope

- Groups can have different scope
  - LAN (local scope)
  - Campus/admin scoping
  - TTL scoping
- Concept of scope important to multipoint protocols and applications

8

Example Applications

- Broadcast audio/video
- Push-based systems
- Software distribution
- Web-cache updates
- Teleconferencing (audio, video, shared whiteboard, text editor)
- Multi-player games
- Server/service location
- Other distributed applications

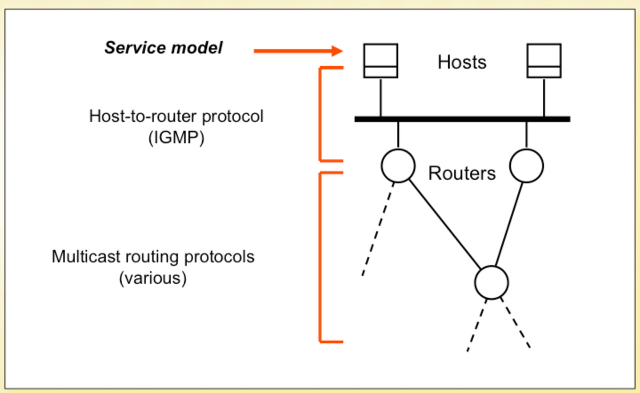
9

Overview

- What/Why Multicast
- IP Multicast Service Basics
- Multicast Routing Basics
- DVMRP
- Reliability
- Congestion Control
- Overlay Multicast
- Publish-Subscribe

10

IP Multicast Architecture



11

IP Multicast Service Model (rfc1112)

- Each group identified by a single IP address
- Groups may be of any size
- Members of groups may be located anywhere in the Internet
- Members of groups can join and leave at will
- Senders need not be members
- Group membership not known explicitly
- Analogy:
  - Each multicast address is like a radio frequency, on which anyone can transmit, and to which anyone can tune-in.

12

IP Multicast Addresses

- Class D IP addresses
  - 224.0.0.0 – 239.255.255.255

1 1 1 0	Group ID

- How to allocated these addresses?
  - Well-known multicast addresses, assigned by IANA
  - Transient multicast addresses, assigned and reclaimed dynamically, e.g., by "sdr" program

13

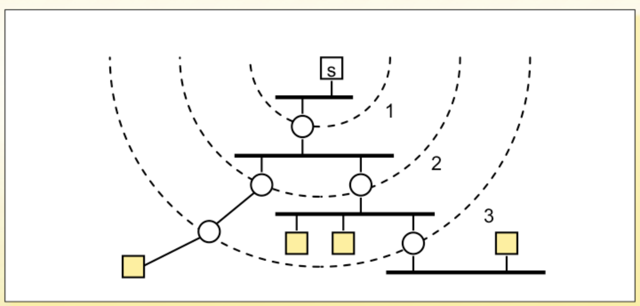
IP Multicast Service

- Sending – same as before
- Receiving – two new operations
  - Join-IP-Multicast-Group(group-address, interface)
  - Leave-IP-Multicast-Group(group-address, interface)
  - Receive multicast packets for joined groups via normal IP-Receive operation

14

Multicast Scope Control – Small TTLs

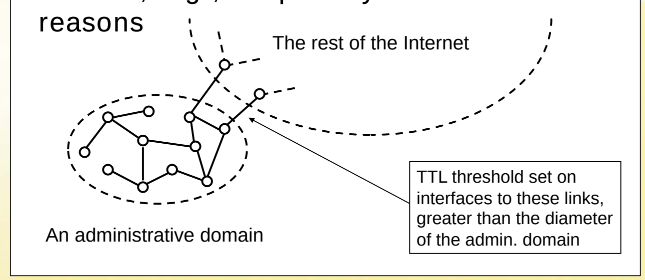
- TTL expanding-ring search to reach or find a nearby subset of a group



15

Multicast Scope Control – Large TTLs

- Administrative TTL Boundaries to keep multicast traffic within an administrative domain, e.g., for privacy or resource reasons



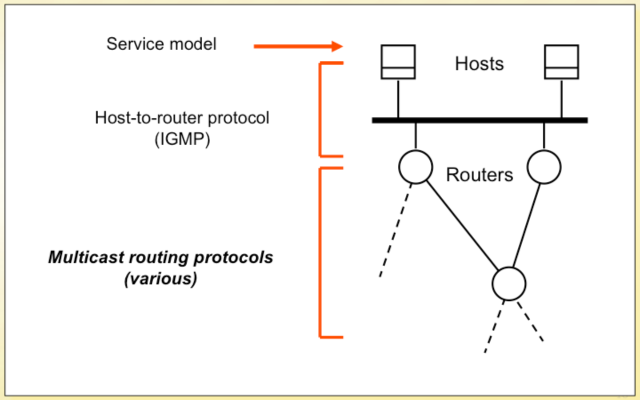
16

Overview

- What/Why Multicast
- IP Multicast Service Basics
- Multicast Routing Basics
- DVMRP
- Reliability
- Congestion Control
- Overlay Multicast
- Publish-Subscribe

17

IP Multicast Architecture



Multicast Routing

- Basic objective – build distribution tree for multicast packets
- Multicast service model makes it hard
  - Anonymity
  - Dynamic join/leave

19

Routing Techniques

- Flood and prune
  - Begin by flooding traffic to entire network
  - Prune branches with no receivers
  - Examples: DVMRP, PIM-DM
  - *Unwanted state where there are no receivers*
- Link-state multicast protocols
  - Routers advertise groups for which they have receivers to entire network
  - Compute trees on demand
  - Example: MOSPF
  - *Unwanted state where there are no senders*

20

Routing Techniques

- Core based protocols
  - Specify "meeting place" aka core
  - Sources send initial packets to core
  - Receivers join group at core
  - Requires mapping between multicast group address and "meeting place"
  - Examples: CBT, PIM-SM

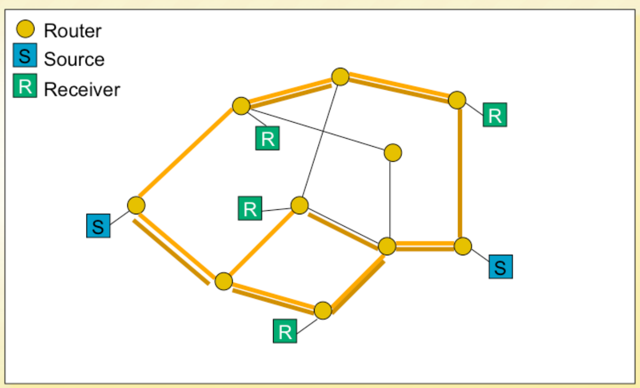
21

Shared vs. Source-based Trees

- Source-based trees
  - Separate shortest path tree for each sender
  - DVMRP, MOSPF, PIM-DM, PIM-SM
- Shared trees
  - Single tree shared by all members
  - Data flows on same tree regardless of sender
  - CBT, PIM-SM

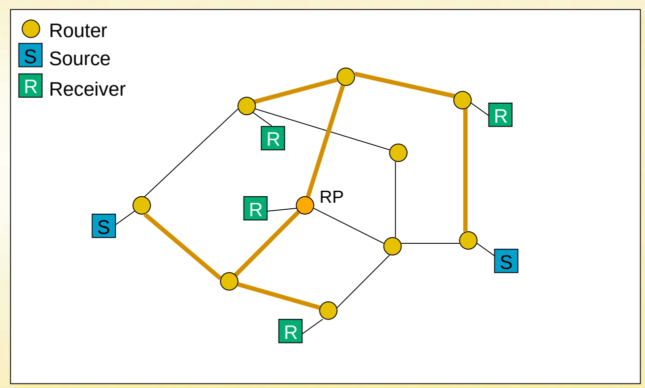
22

Source-based Trees



23

A Shared Tree



24

Shared vs. Source-Based Trees

- Source-based trees
  - Shortest path trees – low delay, better load distribution
  - More state at routers (per-source state)
  - Efficient for in dense-area multicast
- Shared trees
  - Higher delay (bounded by factor of 2), traffic concentration
  - Choice of core affects efficiency
  - Per-group state at routers
  - Efficient for sparse-area multicast
- Which is better? → extra state in routers is bad!

25

Overview

- What/Why Multicast
- IP Multicast Service Basics
- Multicast Routing Basics
- DVMRP
- Reliability
- Congestion Control
- Overlay Multicast
- Publish-Subscribe

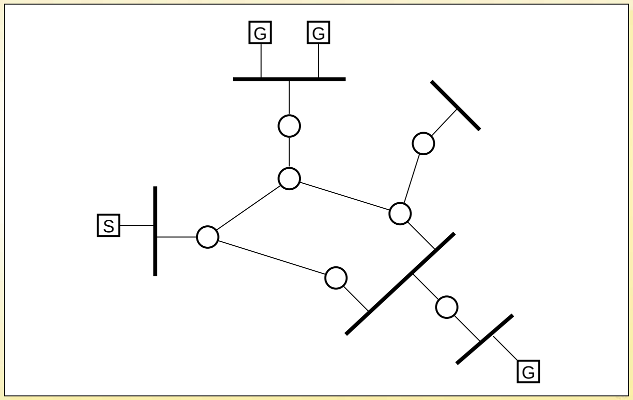
26

Distance-Vector Multicast Routing

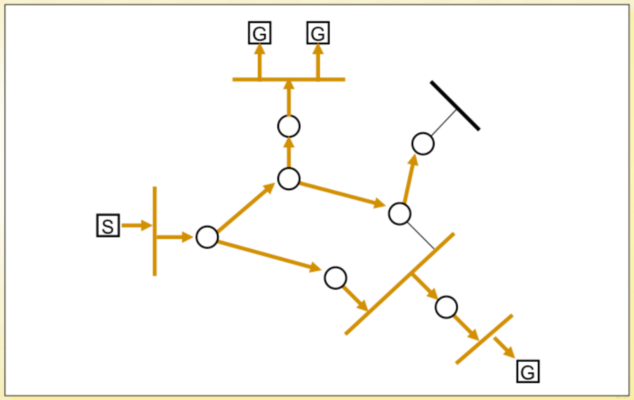
- DVMRP consists of two major components:
  - A conventional distance-vector routing protocol (like RIP)
  - A protocol for determining how to forward multicast packets, based on the routing table
- DVMRP router forwards a packet if
  - The packet arrived from the link used to reach the source of the packet (reverse path forwarding check – RPF)
  - If downstream links have not pruned the tree

27

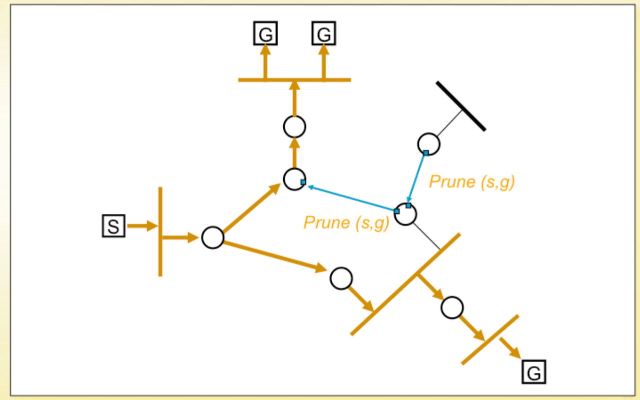
Example Topology



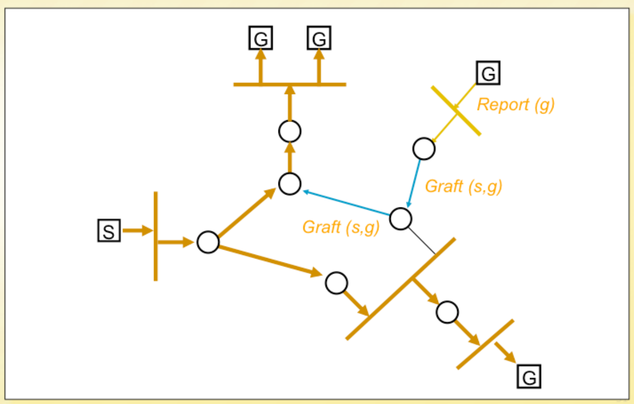
Broadcast with Truncation



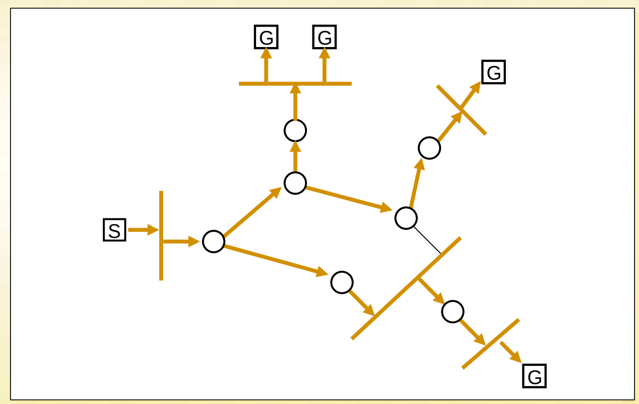
Prune



Graft



Steady State



Overview

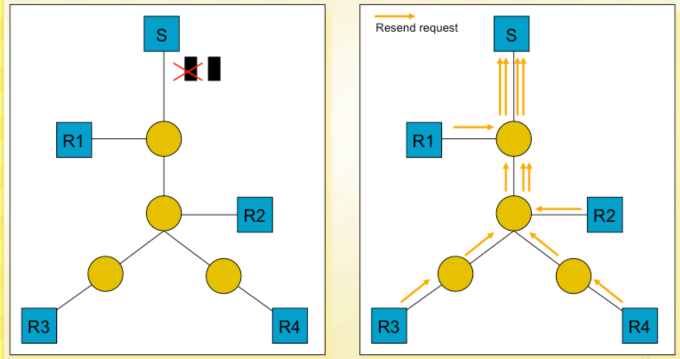
- What/Why Multicast
- IP Multicast Service Basics
- Multicast Routing Basics
- DVMRP
- Reliability
- Congestion Control
- Overlay Multicast
- Publish-Subscribe

33

Implosion

Packet 1 is lost

All 4 receivers request a resend



Retransmission

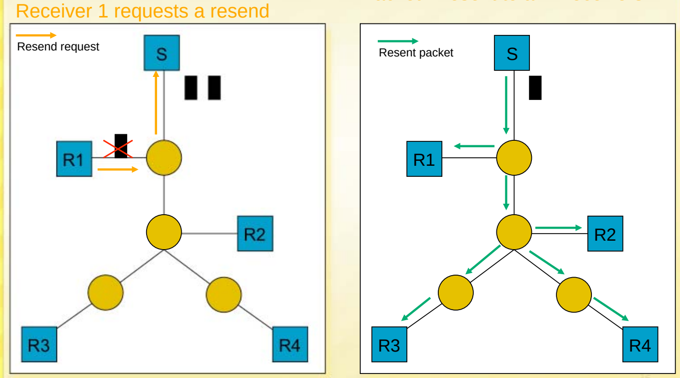
- Re-transmitter
  - Options: sender, other receivers
- How to retransmit
  - Unicast, multicast, scoped multicast, retransmission group, …
- Problem: Exposure

35

Exposure

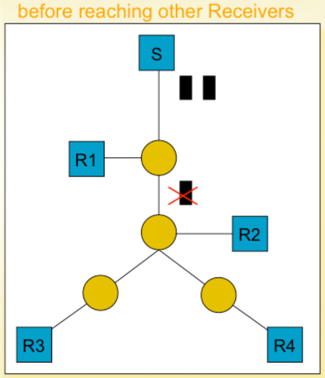
Packet 1 does not reach R1;
Receiver 1 requests a resend

Packet 1 resent to all 4 receivers

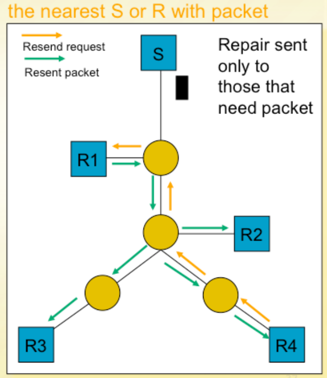


Ideal Recovery Model

Packet 1 reaches R1 but is lost before reaching other Receivers



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

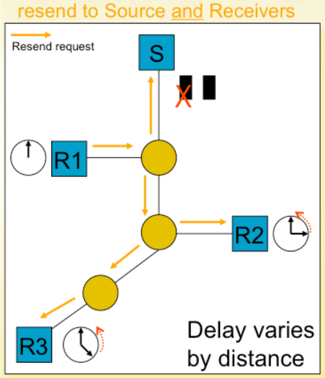


SRM

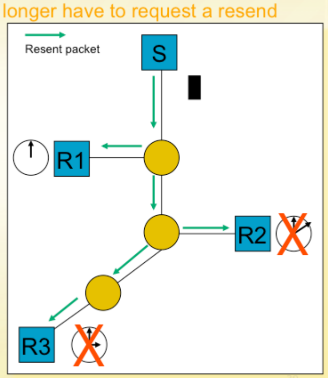
- Originally designed for *wb*
- Receiver-reliable
  - NACK-based
- Every member may multicast NACK or retransmission

SRM Request Suppression

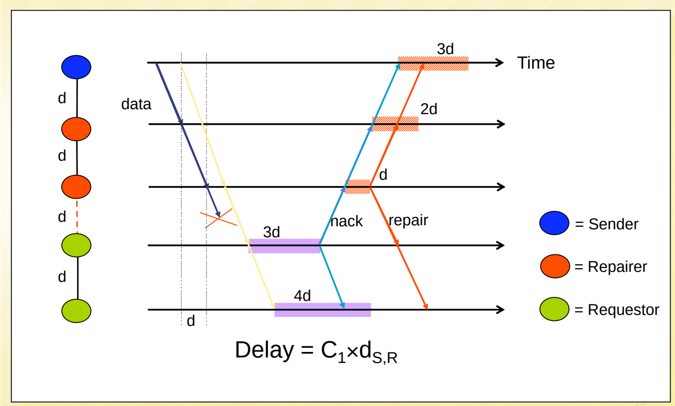
Packet 1 is lost; R1 requests resend to Source and Receivers



Packet 1 is resent; R2 and R3 no longer have to request a resend

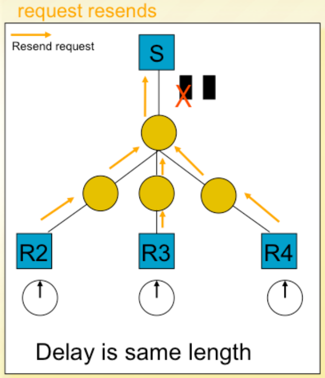


Deterministic Suppression

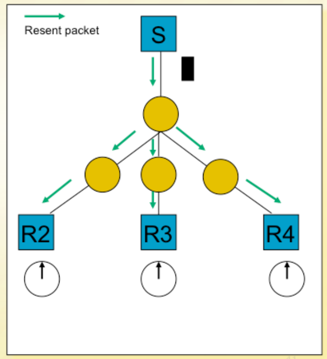


SRM Star Topology

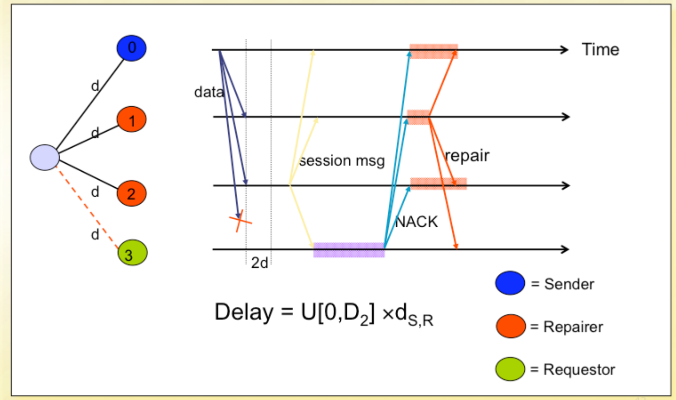
Packet 1 is lost; All Receivers request resends



Packet 1 is resent to all Receivers



SRM: Stochastic Suppression

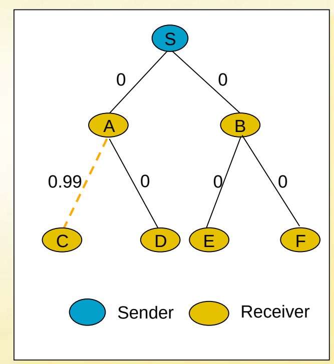


SRM (Summary)

- NACK/Retransmission suppression
  - Delay before sending
  - Delay based on RTT estimation
  - Deterministic + Stochastic components
- Periodic session messages
  - Full reliability
  - Estimation of distance matrix among members

What's Missing?

- Losses at link (A,C) causes retransmission to the whole group
- Only retransmit to those members who lost the packet
- [Only request from the nearest responder]



Local Recovery

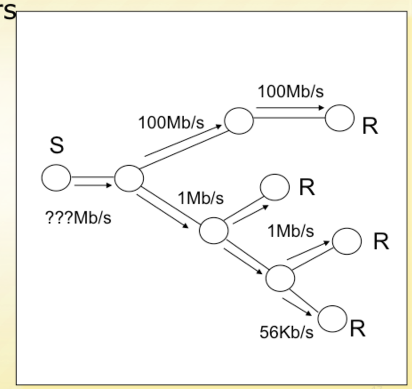
- Different techniques in various systems
- Application-level hierarchy
  - Fixed v.s. dynamic
- TTL scoped multicast
- Router supported

Overview

- What/Why Multicast
- IP Multicast Service Basics
- Multicast Routing Basics
- DVMRP
- Reliability
- Congestion Control
- Overlay Multicast
- Publish-Subscribe

Multicast Congestion Control

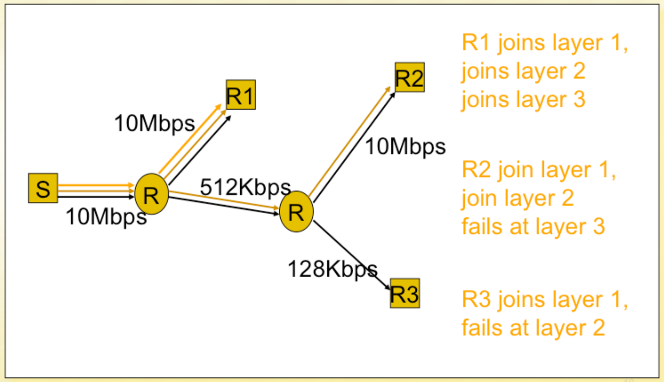
- What if receivers have very different bandwidths?
- Send at max?
- Send at min?
- Send at avg?



Video Adaptation: RLM

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

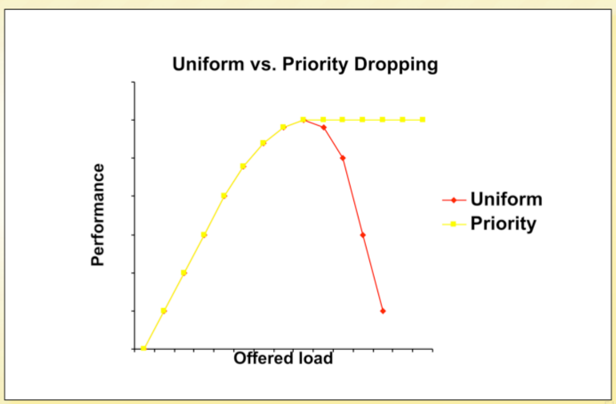
Layered Media Streams



Drop Policies for Layered Multicast

- Priority
  - Packets for low bandwidth layers are kept, drop queued packets for higher layers
  - Requires router support
- Uniform (e.g., drop tail, RED)
  - Packets arriving at congested router are dropped regardless of their layer
- Which is better?
  - Intuition vs. reality!

RLM Intuition



RLM Intuition

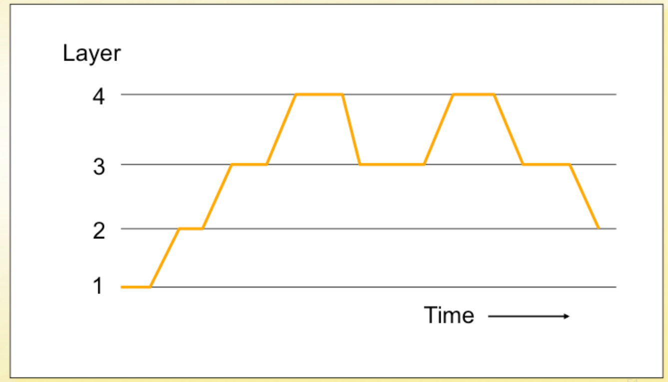
- Uniform
  - Better incentives to well-behaved users
  - If oversend, performance rapidly degrades
  - Clearer congestion signal
  - Allows shared learning
- Priority
  - Can waste upstream resources
  - Hard to deploy
- RLM approaches optimal operating point
  - Uniform is already deployed
  - Assume no special router support

RLM Join Experiment

- Receivers periodically try subscribing to higher layer
- If enough capacity, no congestion, no drops → Keep layer (& try next layer)
- If not enough capacity, congestion, drops → Drop layer (& increase time to next retry)
- What about impact on other receivers?

53

Join Experiments



RLM Scalability?

- What happens with more receivers?
- Increased frequency of experiments?
  - More likely to conflict (false signals)
  - Network spends more time congested
- Reduce # of experiments per host?
  - Takes longer to converge
- Receivers coordinate to improve behavior

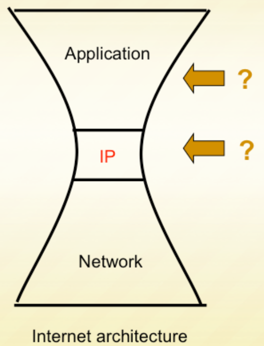
55

Overview

- What/Why Multicast
- IP Multicast Service Basics
- Multicast Routing Basics
- DVMRP
- Reliability
- Congestion Control
- Overlay Multicast
- Publish-Subscribe

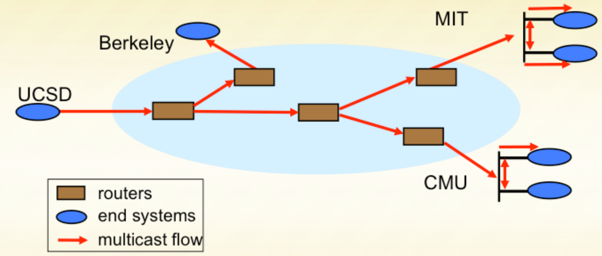
56

Supporting Multicast on the Internet



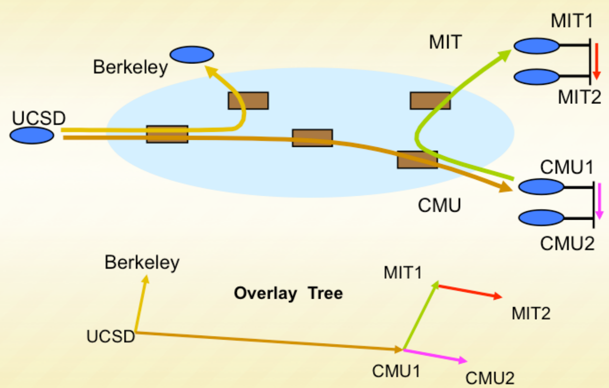
At which layer should multicast be implemented?

IP Multicast



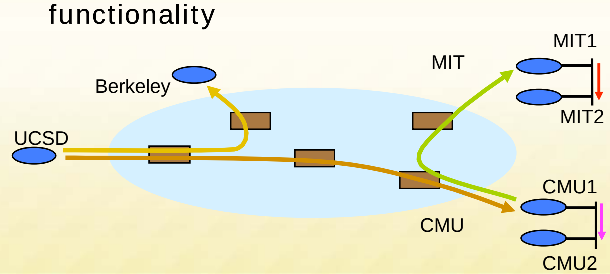
- Highly efficient
- Good delay

End System Multicast



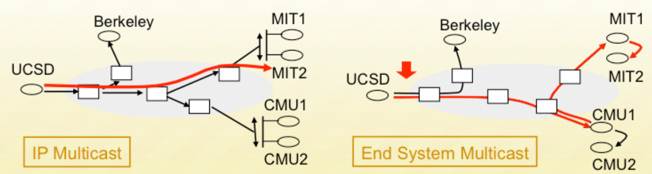
Potential Benefits Over IP Multicast

- Quick deployment
- All multicast state in end systems
- Computation at forwarding points simplifies support for higher level functionality



Concerns with End System Multicast

- Self-organize recipients into multicast delivery overlay tree
  - Must be closely matched to real network topology to be efficient
- Performance concerns compared to IP Multicast
  - Increase in delay
  - Bandwidth waste (packet duplication)

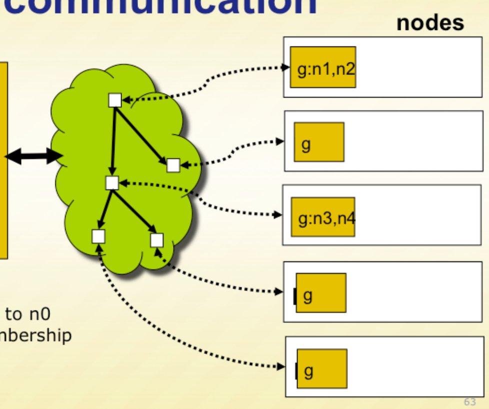


Coordination: Cooperative group communication

- Scribe: Tree-based group management
- Multicast, anycast primitives
- Scalable: large numbers of groups, members, wide range of members/group, dynamic membership
- [IEEE JSAC '02]

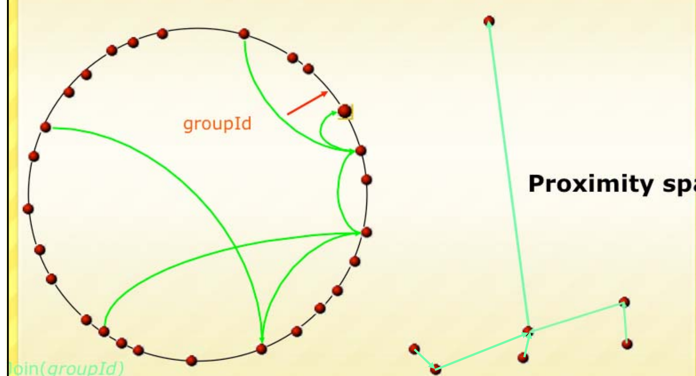
Cooperative group communication

Operations:
create(g)
join(g)
leave(g)
multicast(g,m)
anycast(g,m)



- groupId *g* mapped to n0
- decentralized membership
- robust, scalable

Scribe



Respecting forwarding capacity

- The tree structure described may not respect maximum capacities
- Scribe's push-down fails to resolve the problem because a leaf node in one tree may have children in another tree

Parent location algorithm

- Node adopts prospective child
- If too many children, choose one to reject:
  - First, look for one in stripe without shared prefix
  - Otherwise, select node with shortest prefix match
- Orphan locates new parent in up to two steps:
  - Tries former siblings with stripe prefix match
    - Adopts or rejects using same criteria; continue push-down
  - Use the spare capacity group

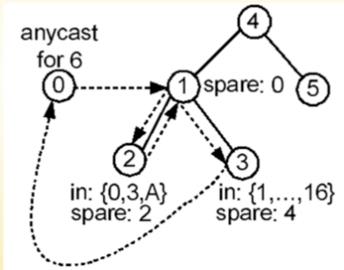
The spare capacity group

- If orphan hasn't found parent yet, anycasts to spare capacity group
- Group contains all nodes with fewer children than their forwarding capacity
- Anycast returns nearby node, which starts a DFS of the spare capacity group tree, sending first to a child…

Spare capacity group (cont.)

- At each node in the search:
  - If node has no children left to search, check whether it receives a stripe the orphan seeks
  - If so, verifies that the orphan is not an ancestor (which would create a cycle)
- If both tests succeed, the node adopts the orphan
  - May leave spare capacity group
- If either test fails, back up to parent (more DFS…)

A spare capacity example



69

Problems

- Imposing bandwidth constraints on Scribe can
- result in:
  - High tree depth
  - non-DHT links
- Observed Cause: mismatch between id space and node bandwidth constraints

70

Overview

- What/Why Multicast
- IP Multicast Service Basics
- Multicast Routing Basics
- DVMRP
- Reliability
- Congestion Control
- Overlay Multicast
- Publish-Subscribe

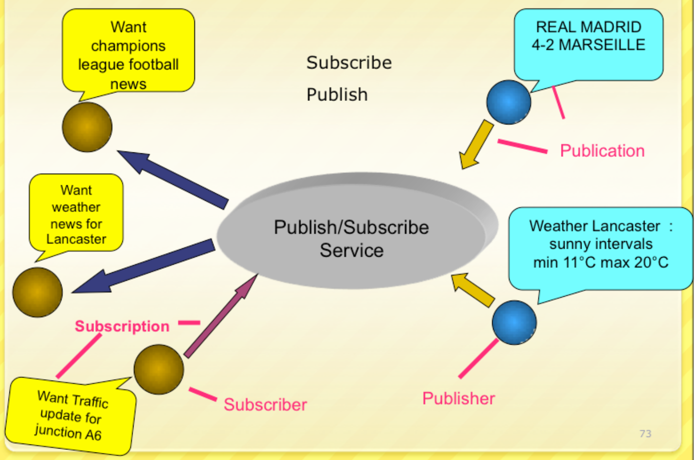
71

Publish-Subscribe

- P/S service is also known as event service
- Publishers role : Publishers generate event data and publishes them
- Subscribers role : Subscribers submit their subscriptions and process the events received
- P/S service: It's the mediator/broker that routes events from publishers to interested subscribers

72

Publish/Subscribe



73

Key attributes of P/S communication model

- The publishing entities and subscribing entities are <u>anonymous</u>
- The publishing entities and subscribing entities are highly <u>de-coupled</u>
- <u>Asynchronous</u> communication model
- The number of publishing and subscribing entities can dynamically change without affecting the entire system

74

Key functions implemented by P/S service

- Event filtering (event selection)- The process which selects the set of subscribers that have shown interest in a given event

- Event routing (event delivery) – The process of routing the published events from the publisher to all interested subscribers

75

Subject based vs. Content based

- Subject based:
  - Generally also known as topic based, group based or channel based event filtering.
  - Here each event is published to one of these channels by its publisher
  - A subscriber subscribes to a particular channel and will receive all events published to the subscribed channel.
  - Simple process for matching an event to subscriptions

76

Subject based vs. Content based

- Content based:
  - More flexibility and power to subscribers, by allowing to express as an arbitrary query over the contents of the event.
  - E.g. Notify me of all stock quotes of IBM from New York stock exchange if the price is greater than 150
  - Added complexity in matching an event to subscriptions

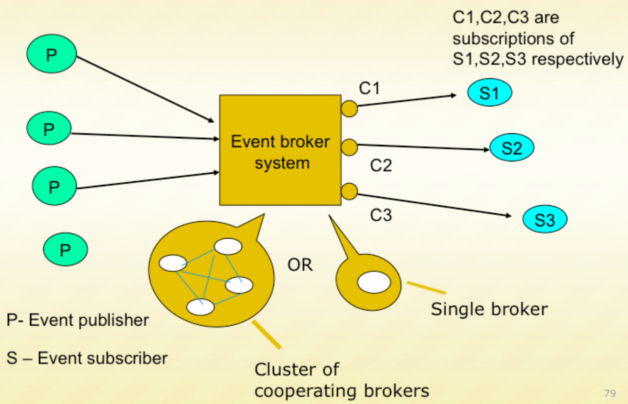
77

Event routing

- The basic P/S system consists of many event publishers, an event broker (or mediator) and many subscribers.
- An event publisher generates an event in response to some change it monitors
- The events are published to an event broker which matches events against all subscriptions forwarded by subscribers in the system.
- Event broker system could have either a single event broker or multiple distributed event brokers coordinating among themselves

78

Event routing



79

Basic elements of P/S model

- Event data model
  - Structure
  - Types
- Subscription model
  - Filter language
  - Scope (subject, content, context)
- General challenge
  - Expressiveness vs. Scalability

80