15-744: Computer Networking

L-16 Multicast Routing

Multicast Routing

- IP Multicast
- IGMP
- Multicast routing
- Assigned reading
  - [DC90] Multicast Routing in Datagram Internetworks and Extended LANs
  - [CRSZ01] Enabling Conferencing Applications on the Internet using an Overlay Multicast Architecture

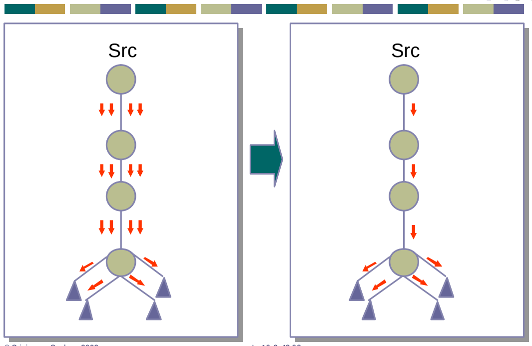
Multicast Routing

- 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
- Host/Router Interaction
- Multicast Routing Basics
- DVMRP
- Overlay Multicast

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

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

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

Scope

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

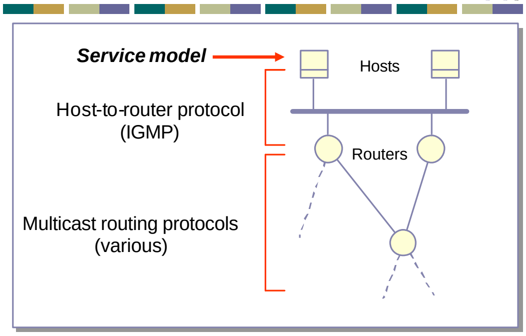
Example Applications

- 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

Overview

- What/Why Multicast
- IP Multicast Service Basics
- Host/Router Interaction
- Multicast Routing Basics
- DVMRP
- Overlay Multicast

IP Multicast Architecture



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.

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

IP Multicast Service — Sending

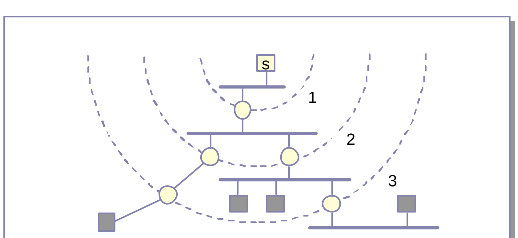
- Uses normal IP-Send operation, with an IP multicast address specified as the destination
- Must provide sending application a way to:
  - Specify outgoing network interface, if >1 available
  - Specify IP time-to-live (TTL) on outgoing packet
  - Enable/disable loop-back if the sending host is a member of the destination group on the outgoing interface

IP Multicast Service — 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

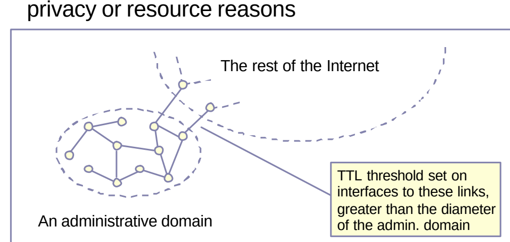
Multicast Scope Control – Small TTLs

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



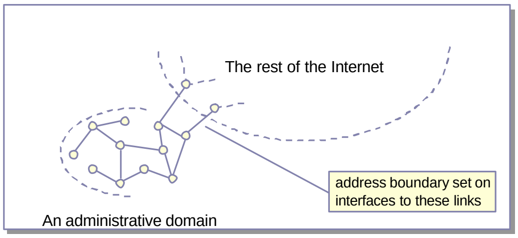
Multicast Scope Control – Large TTLs

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



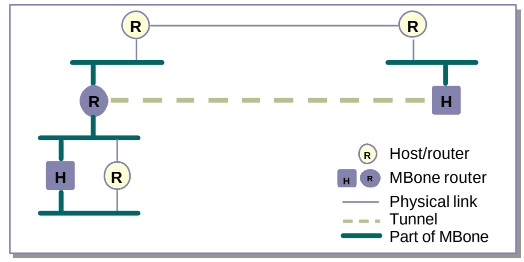
Multicast Scope Control

- Administratively-Scoped Addresses (RFC 1112)
  - Uses address range 239.0.0.0 — 239.255.255.255
  - Supports overlapping (not just nested) domains



Multicast Backbone (MBone)

- An overlay network of IP multicast-capable routers



MBone Tunnels

- A method for sending multicast packets through multicast-ignorant routers
- IP multicast packet is encapsulated in a unicast packet addressed to far end of tunnel:

IP header, dest = unicast	IP header, dest = multicast	Transport header and data…

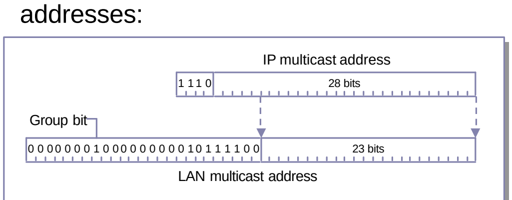
- Tunnel acts like a virtual point-to-point link
- Each end of tunnel is manually configured with unicast address of the other end

Link-Layer Transmission/Reception

- Transmission
  - IP multicast packet is transmitted as a link-layer multicast, on those links that support multicast
  - Link-layer destination address is determined by an algorithm specific to the type of link
- Reception
  - Necessary steps are taken to receive desired multicasts on a particular link, such as modifying address reception filters on LAN interfaces
  - Multicast routers must be able to receive all IP multicasts on a link, without knowing in advance which groups will be used

Using Link-Layer Multicast Addresses

- Ethernet and other LANs using 802 addresses:

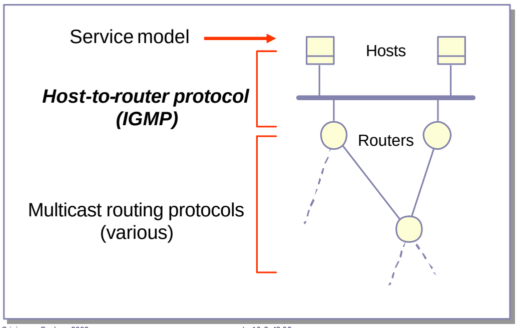


- No mapping needed for point-to-point links

Overview

- What/Why Multicast
- IP Multicast Service Basics
- Host/Router Interaction
- Multicast Routing Basics
- DVMRP
- Overlay Multicast

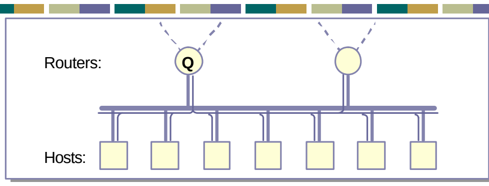
IP Multicast Architecture



Internet Group Management Protocol

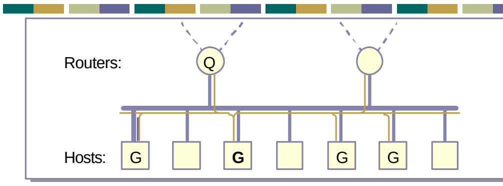
- End system to router protocol is IGMP
- Each host keeps track of which mcast groups are subscribed to
  - Socket API informs IGMP process of all joins
- Objective is to keep router up-to-date with group membership of entire LAN
  - Routers need not know who all the members are, only that members exist

How IGMP Works



- On each link, one router is elected the "querier"
- Querier periodically sends a Membership Query message to the all-systems group (224.0.0.1), with TTL = 1
- On receipt, hosts start random timers (between 0 and 10 seconds) for each multicast group to which they belong

How IGMP Works (cont.)



- When a host's timer for group G expires, it sends a Membership Report to group G, with TTL = 1
- Other members of G hear the report and stop their timers
- Routers hear all reports, and time out non-responding groups

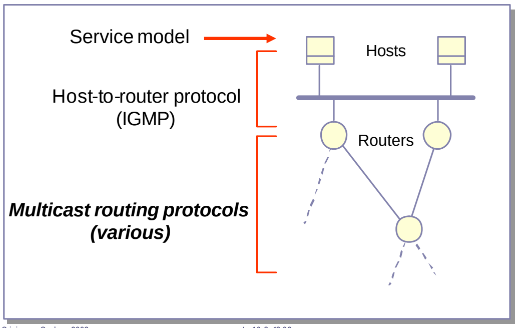
How IGMP Works (cont.)

- Note that, in normal case, only one report message per group present is sent in response to a query
- Query interval is typically 60-90 seconds
- When a host first joins a group, it sends one or two immediate reports, instead of waiting for a query

Overview

- What/Why Multicast
- IP Multicast Service Basics
- Host/router Interaction
- Multicast Routing Basics
- DVMRP
- Overlay Multicast

IP Multicast Architecture



Multicast Routing

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

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*

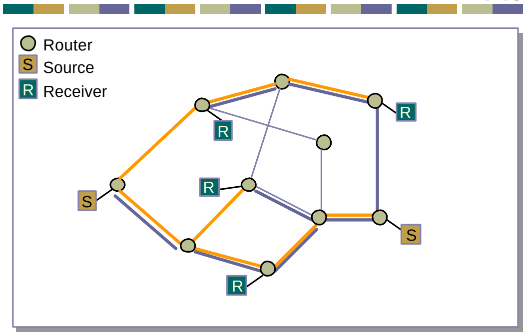
Routing Techniques

- 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

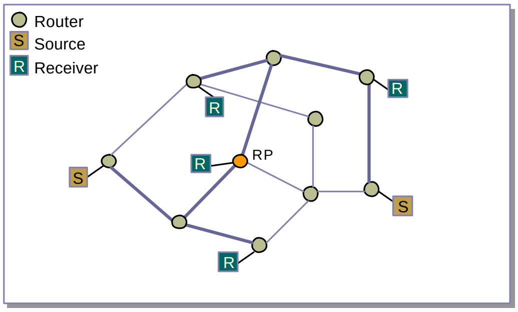
Shared vs. Source-based Trees

- 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

Source-based Trees



A Shared Tree



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

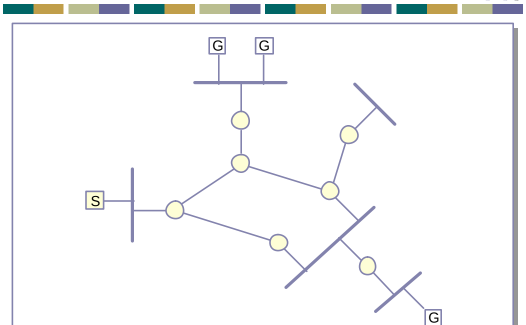
Overview

- What/Why Multicast
- IP Multicast Service Basics
- Host/Router Interaction
- Multicast Routing Basics
- DVMRP
- Overlay Multicast

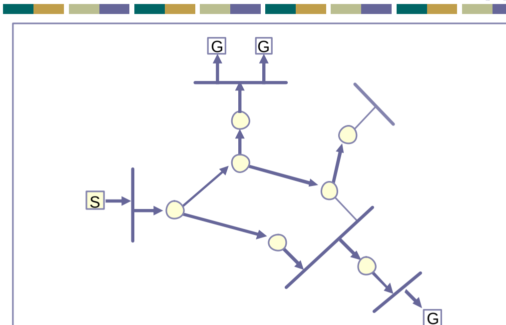
Distance-Vector Multicast Routing

- 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

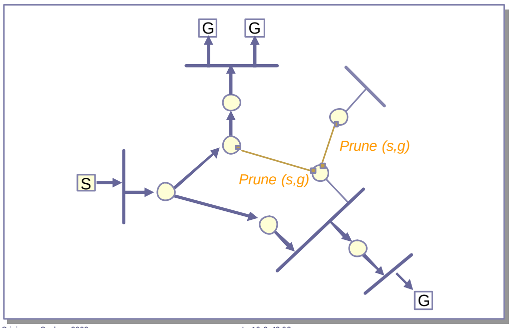
Example Topology



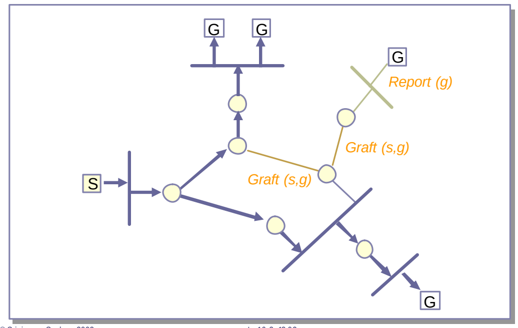
Broadcast with Truncation



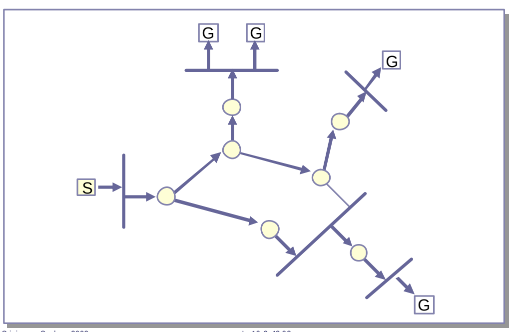
Prune



Graft



Steady State



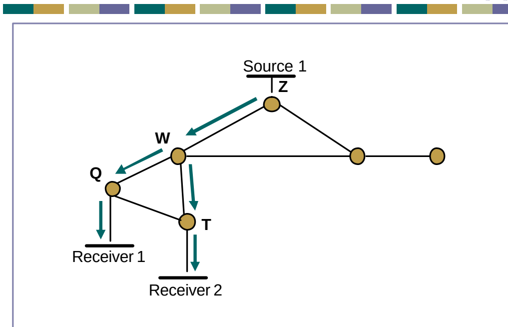
Overview

- MOSPF

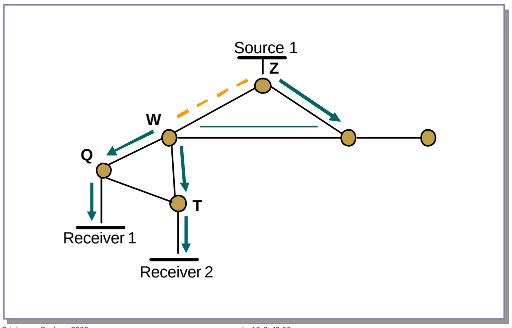
Multicast OSPF (MOSPF)

- Add-on to OSPF (Open Shortest-Path First, a link-state, intra-domain routing protocol)
- Multicast-capable routers flag link state routing advertisements
- Link-state packets include multicast group addresses to which local members have joined
- Routing algorithm augmented to compute shortest-path distribution tree from a source to any set of destinations

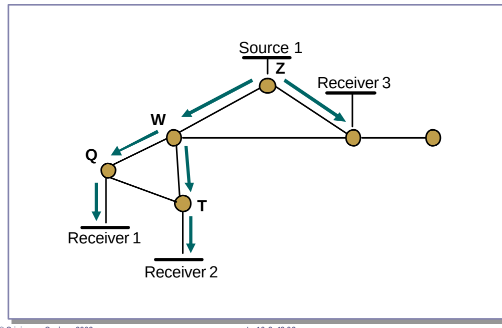
Example



Link Failure/Topology Change



Membership Change



Impact on Route Computation

- Can't pre-compute all source multicast trees
- Compute on demand when first packet from a source S to a group G arrives
- New link-state advertisement
  - May lead to addition or deletion of outgoing interfaces if it contains different group addresses
  - May lead to re-computation of entire tree if links are changed

Overview

- PIM

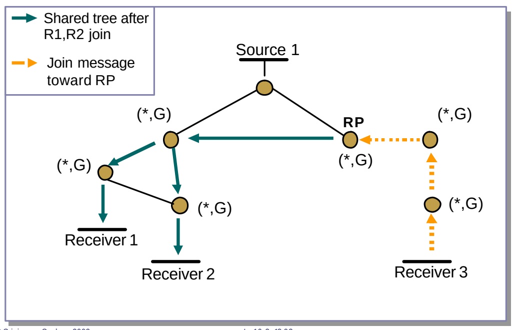
Protocol Independent Multicast (PIM)

- Support for both shared and per-source trees
- Dense mode (per-source tree)
  - Similar to DVMRP
- Sparse mode (shared tree)
  - Core = rendezvous point (RP)
- Independent of unicast routing protocol
  - Just uses unicast forwarding table

PIM Protocol Overview

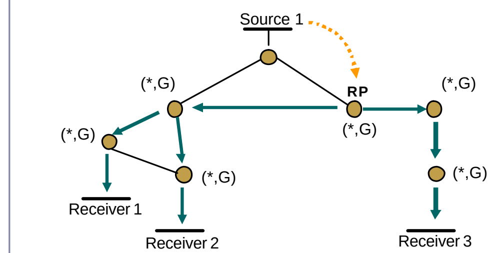
- Basic protocol steps
  - Routers with local members Join toward Rendezvous Point (RP) to join shared tree
  - Routers with local sources encapsulate data in Register messages to RP
  - Routers with local members may initiate data-driven switch to source-specific shortest path trees
- PIM v.2 Specification (RFC2362)

PIM Example: Build Shared Tree



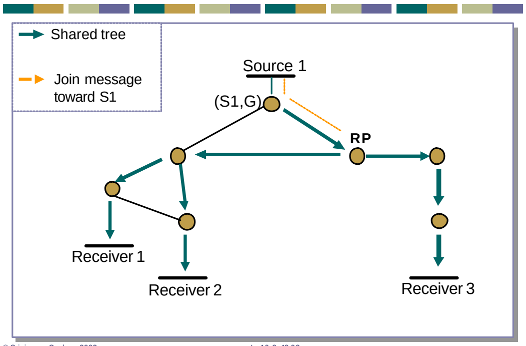
Data Encapsulated in Register

Unicast encapsulated data packet to RP in **Register**

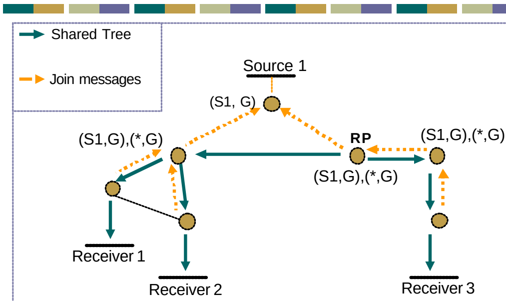


RP decapsulates, forwards down shared tree

RP Send Join to High Rate Source

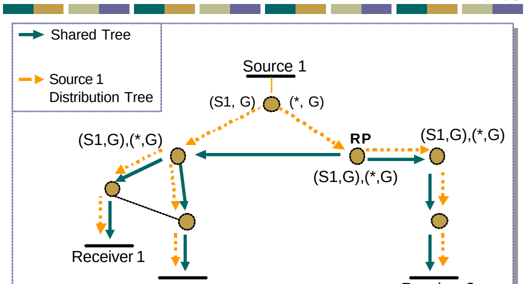


Build Source-Specific Distribution Tree



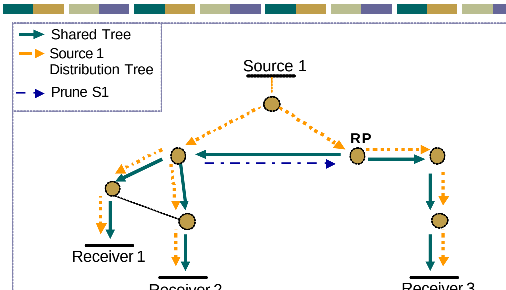
Build source-specific tree for high data rate source

Forward On "Longest-match" Entry



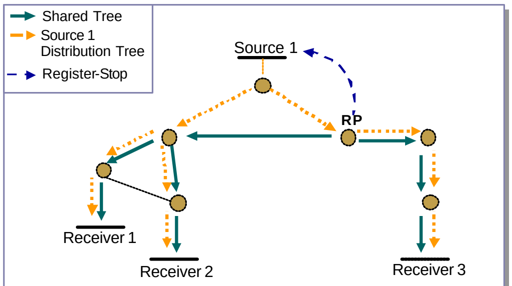
Source-specific entry is "longer match" for source S1 than is Shared tree entry that can be used by any source

Prune S1 off Shared Tree



Prune S1 off shared tree where iif of S1 and RP entries differ

Register-Stop

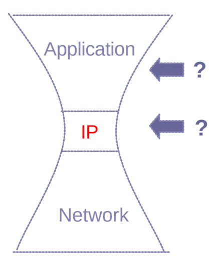


RP unicasts Register-Stop to S1 when packets received natively

Overview

- What/Why Multicast
- IP Multicast Service Basics
- Host/Router Interaction
- Multicast Routing Basics
- DVMRP
- Overlay Multicast

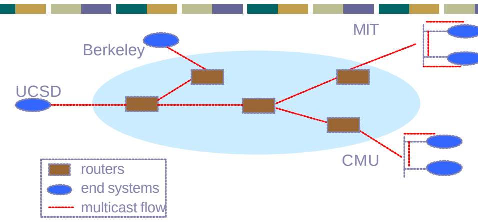
Supporting Multicast on the Internet



At which layer should multicast be implemented?

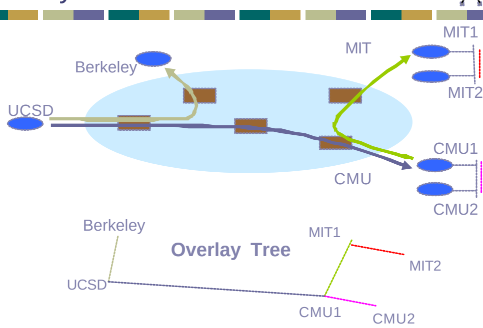
Internet architecture

IP Multicast



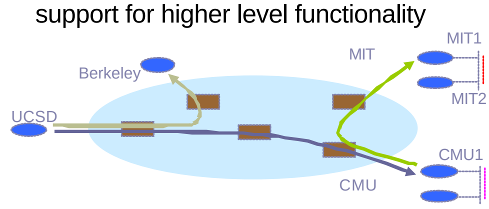
- Highly efficient
- Good delay

End System Multicast



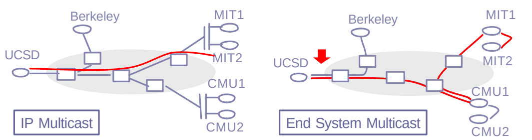
Potential Benefits Over IP Multicast

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



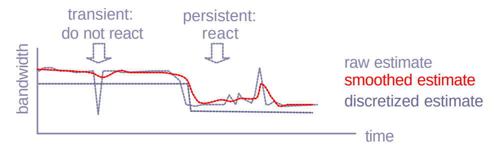
Concerns with End System Multicast

- 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)



Adapt to Dynamic Metrics

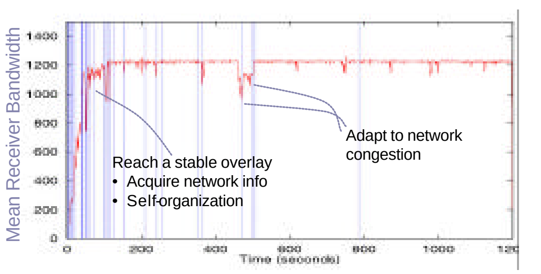
- Adapt overlay trees to changes in network condition
  - Monitor bandwidth and latency of overlay links
- Link measurements can be noisy
  - Aggressive adaptation may cause overlay instability



- Capture the long term performance of a link
  – Exponential smoothing, Metric discretization

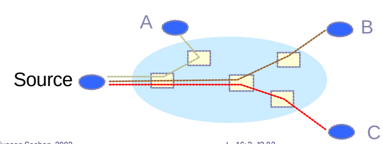
Example of Protocol Behavior

- All members join at time 0
- Single sender, CBR traffic

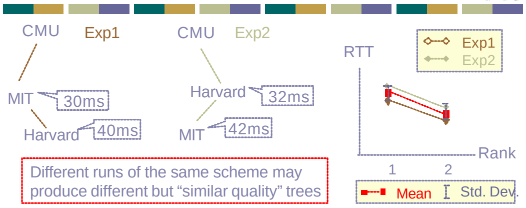


Benchmark Scheme

- IP Multicast not deployed
- Sequential Unicast: an approximation
  - Bandwidth and latency of unicast path from source to each receiver
  - Performance similar to IP Multicast with ubiquitous deployment



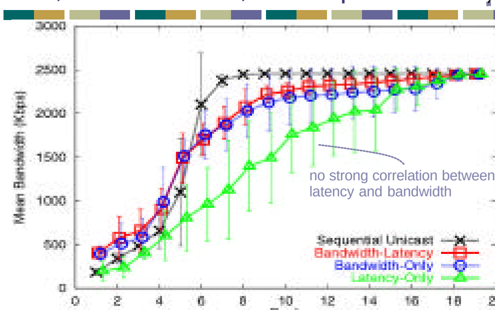
Performance of Overlay Scheme



"Quality" of overlay tree produced by a scheme

- Sort ("rank") receivers based on performance
- Take mean and std. dev. on performance of same rank across multiple experiments
- Std. dev. shows variability of tree quality

BW, Extended Set, 2.4 Mbps



Optimizing only for latency has poor bandwidth performance

Resource Usage (RU)

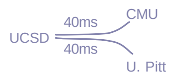
Captures consumption of network resource of overlay tree

- Overlay link RU = propagation delay
- Tree RU = sum of link RU

Scenario: Primary Set, 1.2 Mbps (normalized to IP Multicast RU)

IP Multicast	1.0
Bandwidth-Latency	1.49
Random	2.24
Naïve Unicast	2.62

Efficient (RU = 42ms)



Inefficient (RU = 80ms)

Next Lecture: Multicast Routing

- Reliable multicast
- Multicast congestion control
- Multicast routing
- Assigned reading
  - [F+97] A Reliable Multicast Framework for Light-Weight Sessions and Application Level Framing
  - [MJV96] Receiver-driven Layered Multicast