EECS 122:
Introduction to Computer Networks
Multicast and Overlay Networks

Ion Stoica (and Brighten Godfrey)
TAs: Lucian Popa, David Zats and Ganesh Ananthanarayanan
http://inst.eecs.berkeley.edu/~ee122/
Materials with thanks to Vern Paxson, Jennifer Rexford, and colleagues at Princeton and UC Berkeley

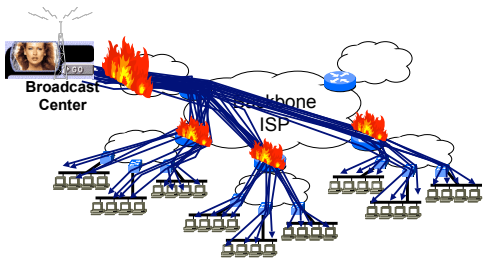
Announcements

- No class Wednesday. Happy Thanksgiving!
- Project 2a code posted… use at will for part 2b
  - but, better and probably easier to use your own code

Motivational Example: Streaming Media

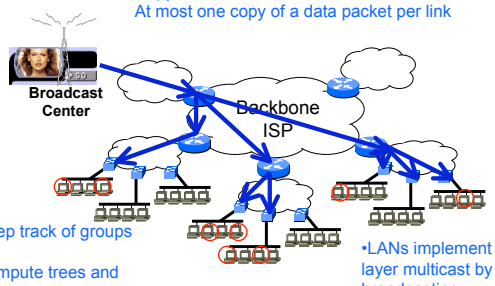
- Live 8 concert
  - Send ~300 Kb/s video streams
  - Peak usage > 100,000 simultaneous users
  - Consumes > 30 Gb/s
  - If 1000 people are in Berkeley, and if the concert were broadcast from a single location, 1000 unicast streams are sent from that location to Berkeley

This approach does not scale…



Instead build trees

Copy data at routers
At most one copy of a data packet per link



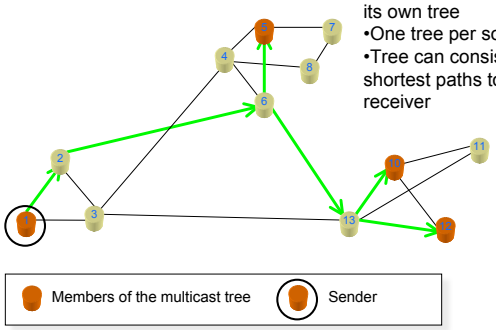
•Routers keep track of groups in real-time
•Routers compute trees and forward packets along them

•LANs implement link layer multicast by broadcasting

Multicast Routing Approaches

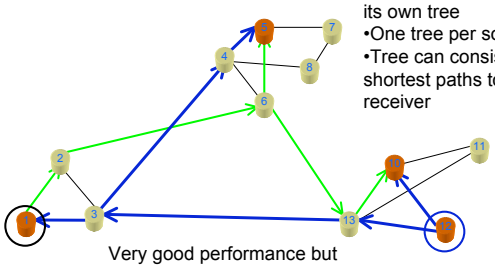
- Kinds of Trees
  - Source Specific Trees
  - Shared Tree
- Tree Computation Methods
  - Link state
  - Distance vector

Source Specific Trees



- Each source is the root of its own tree
- One tree per source
- Tree can consists of shortest paths to each receiver

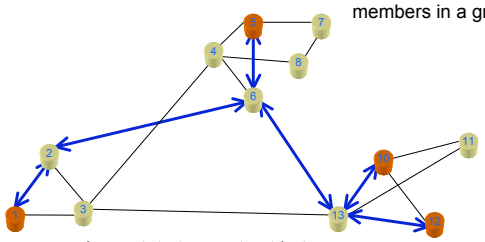
Source Specific Trees



- Each source is the root of its own tree
- One tree per source
- Tree can consists of shortest paths to each receiver

Very good performance but expensive to construct/maintain; routers need to manage a tree per source

Shared Tree

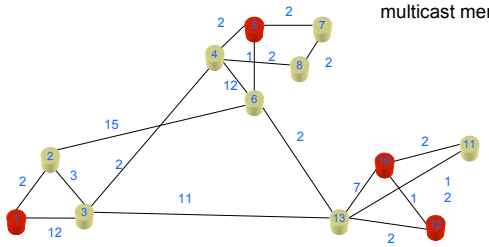


One tree used by all members in a group

Less state to construct but hard to pick "good" trees for everyone!

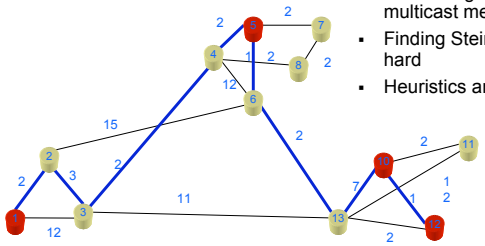
Shared Tree

- Ideally, find a Steiner tree - the minimum-weighted tree connecting only the multicast members



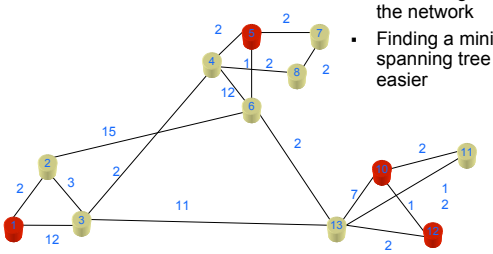
Shared Tree

- Ideally, find a Steiner tree – minimum-weighted tree connecting only the multicast members
- Finding Steiner Tree is NP hard
- Heuristics are known



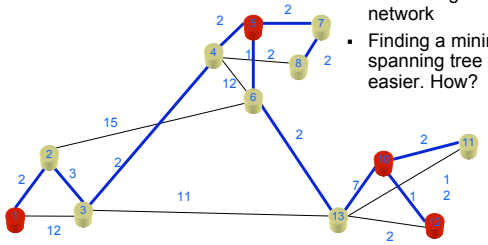
Shared Tree

- Example heuristic: find a minimum-spanning tree – minimum-weighted tree connecting all nodes in the network
- Finding a minimum spanning tree is much easier



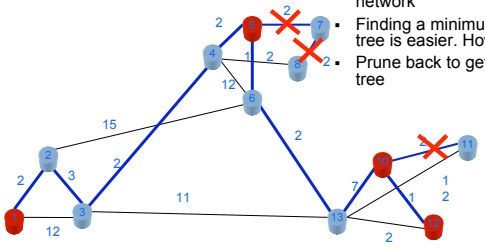
Shared Tree

- Example heuristic: find a minimum-spanning tree – minimum-weighted tree connecting all nodes in the network
- Finding a minimum spanning tree is much easier. How?



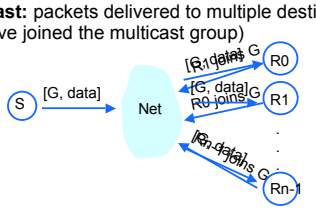
Shared Tree

- Example heuristic: find a minimum-spanning tree – minimum-weighted tree connecting all nodes in the network
- Finding a minimum spanning tree is easier. How?
- Prune back to get multicast tree



Multicast Service Model

- **Unicast:** packets delivered to one destination
- **Broadcast:** packets delivered to all end-hosts
- **Multicast:** packets delivered to multiple destinations (those that have joined the multicast group)



- Receivers join a multicast group which is identified by a multicast address (e.g. G)
- Sender(s) send data to address G
- Network routes data to each of the receivers

Multicast Service Model (cont'd)

- Membership access control
  - Open group: anyone can join
  - Closed group: restrictions on joining
- Sender access control
  - Anyone can send to group
  - Anyone in group can send to group
  - Restrictions on which host can send to group

Multicast and Layering

- Multicast can be implemented at different layers
  - data link layer
    - e.g. Ethernet multicast
  - network layer
    - e.g. IP multicast
  - application layer
    - e.g. End system multicast
- Which layer is best?

Multicast Implementation Issues

- How are multicast packets addressed?
- How is join implemented?
- How is send implemented?
- How much state is kept and who keeps it?

Data Link Layer Multicast

- Recall: end-hosts in the same local area network (LAN) can hear from each other at the data link layer (e.g., Ethernet)
- Reserve some data link layer addresses for multicast
- Join group at multicast address G
  - Network interface card (NIC) normally only listens for packets sent to unicast address A and broadcast address B
  - To join group G, NIC also listens for packets sent to multicast address G (NIC limits number of groups joined)
  - Implemented in hardware, thus efficient
- Send to group G
  - Packet is flooded on all LAN segments, like broadcast
  - Can waste bandwidth, but LANs should not be very large
- Only host NICs keep state about who has joined → scalable to large number of receivers, groups

Problems with Data Link Layer Multicast

- Single data link technology
- Single LAN
  - Limited to small number of hosts
  - Limited to low diameter latency
  - Essentially all the limitations of LANs compared to internetworks

Network Layer (IP) Multicast

- Overcomes limitations of data link layer multicast
- Performs inter-network multicast routing
  - Relies on data link layer multicast for intra-network routing
- Portion of IP address space defined as multicast addresses
  - 2^{28} addresses for entire Internet
- Open group membership
- Anyone can send to group
  - Flexible, but leads to problems

IP Multicast Routing

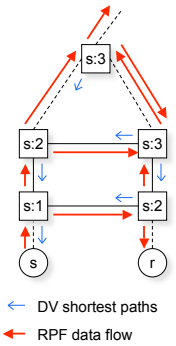
- Intra-domain
  - Distance-vector multicast
  - Link-state multicast
- Inter-domain
  - Protocol Independent Multicast
  - Single Source Multicast

Distance Vector Multicast Routing Protocol (DVRMP)

- An elegant extension to DV routing
- Use shortest path DV routes to determine if link is on the source-rooted spanning tree
- Three steps in developing DVRMP
  - Reverse Path Flooding
  - Reverse Path Broadcasting
  - Truncated Reverse Path Broadcasting

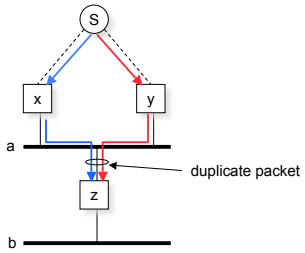
Reverse Path Flooding (RPF)

- Extension to DV unicast routing
- Packet forwarding
  - If incoming link is shortest path to source
  - Send on all links except incoming
  - Packets always take shortest path
    - assuming delay is symmetric
- Issues
  - Some links (LANs) may receive multiple copies
  - Every link receives each multicast packet, even if no interested hosts



Example

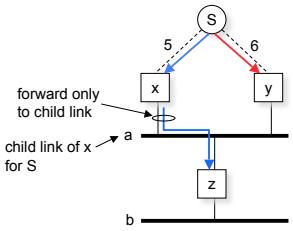
- Flooding can cause a given packet to be sent multiple times over the same link



- Solution: Reverse Path Broadcasting

Reverse Path Broadcasting (RPB)

- Chose parent of each link along reverse shortest path to source
- Only parent forward to a link (child link)
- Identify Child Links
  - Routing updates identify parent
  - Since distances are known, each router can easily figure out if it's the parent for a given link
  - In case of tie, lower address wins

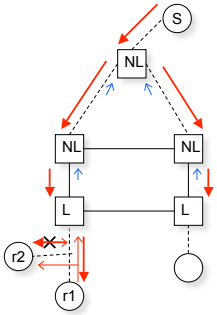


Don't Really Want to Flood!

- This is still a broadcast algorithm – the traffic goes everywhere
- Need to "Prune" the tree when there are subtrees with no group members
- Solution: Truncated Reverse Path Broadcasting

Truncated Reverse Path Broadcasting (TRPB)

- Extend DV/RPB to eliminate unneeded forwarding
- Identify leaves
  - Routers announce that a link is their next link to source S
  - Parent router can determine that it is not a leaf
- Explicit group joining on LAN
  - Members periodically (with random offset) multicast report locally
  - Hear and report, then suppress own
- Packet forwarding
  - If not a leaf router or have members
  - Out all links except incoming



L – leaf node
NL – Non-leaf node

Pruning Details

- Prune (Source,Group) at leaf if no members
  - Send Non-Membership Report (NMR) up tree
- If all children of router R send NRM, prune (S,G)
  - Propagate prune for (S,G) to parent R
- On timeout:
  - Prune dropped
  - Flow is reinstated
  - Down stream routers re-prune
- Note: a soft-state approach

Pruning Details

- How to pick prune timers?
  - Too long → large join time
  - Too short → high control overhead
- What do you do when a member of a group (re)joins?
  - Issue prune-cancellation message (graft)

Distance Vector Multicast Scaling

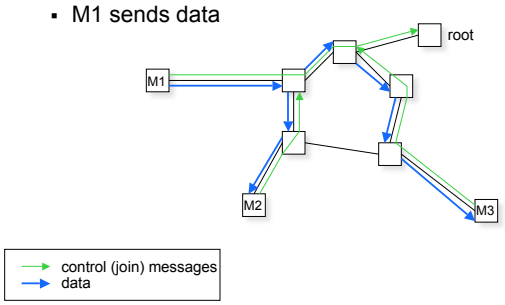
- State requirements:
  - O(Sources × Groups) active state
- How to get better scaling?
  - Hierarchical Multicast
  - Core-based Trees

Core Based Trees (CBT)

- Pick a "rendezvous point" for the group called the core.
  - One tree per group (same tree for all senders in group)
- Unicast packet to core and bounce it back to multicast group
- Tree construction is receiver-based
  - Joins can be tunneled if required
  - Only nodes on tree involved
- Reduce routing table state from $O(S \times G)$ to $O(G)$

Example

- Group members: M1, M2, M3
- M1 sends data



CBT vs. DV Multicast

- **DV Multicast:** One tree for each source
  - More state in routers
  - Better optimized trees
- **CBT:** Trees shared among all sources in a group
  - Less state in routers
  - Shared tree may not be the best tree for any particular source
  - Need to pick a good core node
    - Sub-optimal setup delay
    - Optimal choice (computing topological center) is NP hard

Problems with Network Layer Multicast (NLM)

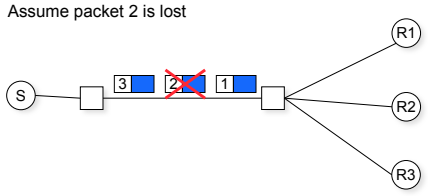
- Scales poorly with number of groups
  - A router must maintain state for every group that traverses it
  - Many groups traverse core routers
- Supporting higher level functionality is difficult
  - NLM: best-effort multi-point delivery service
  - Reliability and congestion control for NLM complicated
- Deployment is difficult and slow
  - ISPs reluctant to turn on NLM

NLM Reliability

- Assume reliability through retransmission
- Sender can not keep state about each receiver
  - E.g., what receivers have received
  - Number of receivers unknown and possibly very large
- Sender can not retransmit every lost packet
  - Even if only one receiver misses packet, sender must retransmit, lowering throughput
- N(ACK) implosion
  - Described next

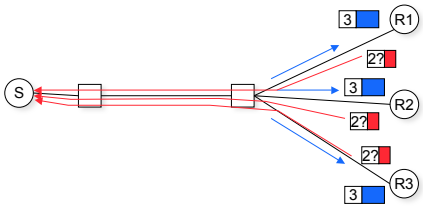
(N)ACK Implosion

- (Positive) acknowledgements
  - Ack every n received packets
  - What happens for multicast?
- Negative acknowledgements
  - Only ack when data is lost
  - Assume packet 2 is lost



NACK Implosion

- When a packet is lost all receivers in the sub-tree originated at the link where the packet is lost send NACKs



Barriers to Multicast

- Hard to change IP
  - Multicast means changes to IP
  - Details of multicast were very hard to get right
- Not always consistent with ISP economic model
  - Charging done at edge, but single packet from edge can explode into millions of packets within network
- Troublesome security model
  - Anyone can send to a group
  - Denial-of-service attacks on known groups

Overlay Networks

Overlay Networks: Motivations

- Protocol changes in the network happen very slowly
- Why?
  - Internet network is a shared infrastructure; need to achieve consensus (IETF)
  - Many of proposals require to change a large number of routers (e.g., IP Multicast, QoS); otherwise end-users won't benefit
- Proposed changes that haven't happened yet on large scale:
  - More Addresses (IPv6 '91)
  - Security (IPSEC '93); Multicast (IP multicast '90)

Motivations (cont'd)

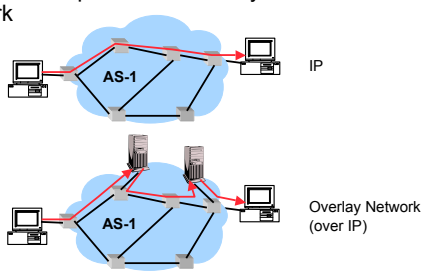
- One size does not fit all
- Applications need different levels of
  - Reliability
  - Performance (latency)
  - Security
  - Access control (e.g., who is allowed to join a multicast group)
  - …

Goals

- Make it easy to deploy new functionalities in the network → accelerate the pace of innovation
- Allow users to customize their service

Solution

- Deploy processing in the network
- Have packets processed as they traverse the network



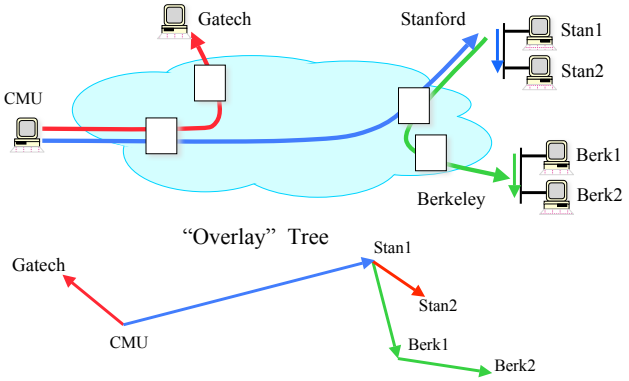
Overlay network overview

- Application Layer (Overlay) multicast
- Resilient Overlay Network (RON)
- Next lecture: Peer-to-peer systems

Application Layer Multicast (ALM)

- Let the hosts do all the "special" work
  - Only require unicast from infrastructure
- Basic idea:
  - Hosts do the copying of packets
  - Set up tree between hosts
- Example: Narada [Yang-hua et al, 2000]
  - Small group sizes <= hundreds of nodes
  - Typical application: streaming video
- (What do you use that's a lot like overlay multicast?)

Narada: End System Multicast



Algorithmic Challenge

- Choosing replication/forwarding points among hosts
  - how do the hosts know about each other
  - and know which hosts should forward to other hosts

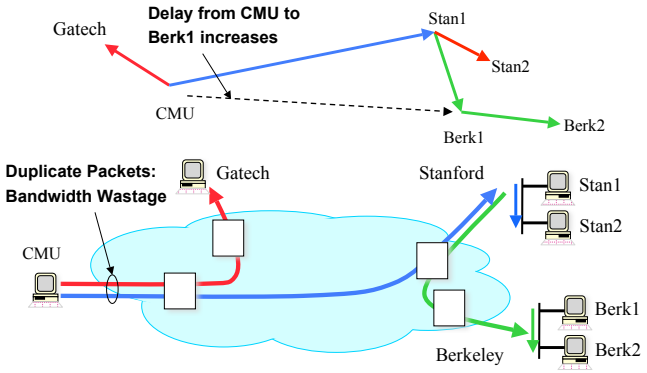
Advantages of ALM

- No need for changes to IP or routers
- No need for ISP cooperation
- End hosts can prevent other hosts from sending
- Easy to implement reliability
  - use hop-by-hop retransmissions

Performance Concerns

- Stretch
  - ratio of latency in the overlay to latency in the underlying network
- Stress
  - number of duplicate packets sent over the same physical link
- Network Layer Multicast can get perfect or near-perfect stretch; ALM can't in general
  - but it can come pretty close

Performance Concerns



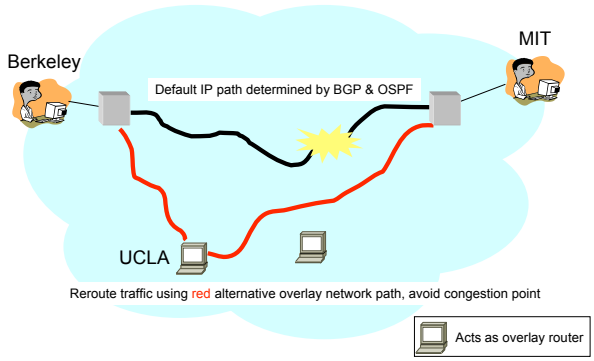
Overlay network overview

- Overlay Multicast (Narada)
- ➢ Resilient Overlay Network (RON)
- Next lecture: Peer-to-peer systems

Resilient Overlay Network (RON)

- Premise: by building application overlay network, can increase performance and reliability of routing
- Install *N* computers at different Internet locations
- Each computer acts as an overlay network router
  - Between each overlay router is an IP tunnel (logical link)
  - Logical overlay topology is all-to-all (N^2 total links)
- Run a link-state routing algorithm in that overlay topology
  - Computers actively measure each logical link in real time for packet loss rate, latency, throughput, etc
  - These define link costs
  - Route overlay network traffic based on measured characteristics

Example



Summary: What You Need to Know

- Multicast protocols
  - DVRMP
  - CBT
  - How they compare
- Overlay networks
  - Benefits and drawbacks
  - More to come: P2P