**Motivation Example: Internet Radio**

- **www.digitallyimported.com** (techno station)
  - Sends out 128Kb/s MP3 music streams
  - Peak usage ~9000 simultaneous streams
    - Only 5 unique streams (trance, hard trance, hard house, eurodance, classical)
  - Consumes ~1.1Gb/s
    - Bandwidth costs are a large fraction of their expenditures (maybe 50%?)
  - If 1000 people are getting their groove on in Berkeley, 1000 unicast streams are sent from NYC to Berkeley

**Multicast Routing Approaches**

- **Kinds of Trees**
  - Source Specific Trees
  - Shared Tree

- **Tree Computation Methods**
  - Link state
  - Distance vector

**This approach does not scale...**

**Source Specific Trees**

- Each source is the root of its own tree
- One tree per source
- Tree can consist of shortest paths to each receiver
Source Specific Trees

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

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

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

- One tree used by all members in a group

Easier to construct/maintain but hard to pick "good" trees for everyone!

Shared Tree

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

Finding Steiner Tree is NP hard

Alternatively, find a minimum-spanning tree — minimum-weighted tree connecting all nodes in the network

Finding a minimum spanning tree is much easier

Alternatively, 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**

- Alternatively, 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**

- Receivers join a multicast group which is identified by a multicast address (e.g., G).
- Senders(s) send data to address G.
- Network routes data to each of the receivers.
- Note: multicast vs. broadcast
  - Broadcast: packets are delivered to all end-hosts in the network.
  - Multicast: packets are delivered only to end-hosts that are in (have joined) the multicast 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 Service Model (cont’d)**

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

**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
  1. Routing updates identify parent
  2. Since distances are known, each router can easily figure out if it's the parent for a given link
  3. 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

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
  - Downstream routers re-prune
- Note: a soft-state approach

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 an report, then suppress own
- Packet forwarding
  - If not a leaf router or have members
  - Out all links except incoming

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.
  - Shared tree
- 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 One tree per group tree involved
- Reduce routing table state from $O(S \times G)$ to $O(G)$

Example

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

Disadvantages

- Sub-optimal delay
- Single point of failure
  - Core goes out and everything lost until error recovery elects a new core
- Small, local groups with non-local core
  - Need good core selection
  - 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
  - ISP’s reluctant to turn on NLM

Example

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

Disadvantages

- Sub-optimal delay
- Single point of failure
  - Core goes out and everything lost until error recovery elects a new core
- Small, local groups with non-local core
  - Need good core selection
  - 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
  - ISP’s 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 ask 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

**Narada: End System Multicast**

- Set up tree between hosts
- Small group sizes <= hundreds of nodes
- Hosts do the copying of packets
- Only require unicast from infrastructure
- Denial-of-service attacks on known groups
- Anyone can send to a group
- Details of multicast were very hard to get right
- Charging done at edge, but single packet from edge can explode into millions of packets within network

**Barriers to Multicast**

- Hard to change IP
  - Multicast means change 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

**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

**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: chat

**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

**Example**

- Group members: M1, M2, M3

**Single Sender Multicast**

- Many problems with IP multicast disappear if each group is associated with a single source

- Hosts joining multicast group can send join messages to source
  - this sets up delivery tree
  - no worry about “root” being in wrong place

- This solves several problems:
  - better security and charging model
  - simple algorithm

**What’s Wrong with SSM?**

- **Multiple sources?**
  - can set up group per source, or...
  - Source can serve as relay for other senders

- **Algorithm?**
  - Trivial

- So, why isn’t SSM the answer?
  - Multicast no longer serves as “rendezvous”
  - Ok for “broadcast” apps, not good for “meeting” apps

**What Do You Need to Know?**

- DVRMP
- CBT
- SSM
- How they compare