Multicasting

Guevara Noubir

Textbook: 1. Computer Networks: A Systems Approach, L. Peterson, B. Davie, Morgan Kaufmann (Chap. 4)
2. Multicasting on the Internet and its applications, Sanjoy Paul, Kluwer Academic Publishers

Lecture Outline

- Introduction to multicast
- Multicast over Ethernet
- Routing protocols for IP multicast (DVRMP, PIM)
- MBone

What is Multicast?

- Multicast is a communication paradigm
  - 1 source, multiple destination
- Applications:
  - **bulk-data distribution to subscribers**
    - **(e.g., newspaper, software, and video tapes distribution),**
  - **connection-time-based charging data distribution**
    - **(e.g., financial data, stock market information, and news tickets broadcasting),**
  - **streaming (e.g., video/audio real-time distribution),**
  - **push applications, web-casting,**
  - **distance learning, conferencing, collaborative work, distributed simulation, and interactive games.**

Why Multicasting?

- Several applications need efficient means to transmit data to multiple destinations with:
  - less bandwidth
  - higher throughput
  - lower delay
  - higher reliability

- Classification
  - Data dissemination
  - Transactions
  - Large Scale Virtual Environments

Ethernet Multicast

- Ethernet is a broadcast medium
  - Every frame can potentially be seen by every host
- Ethernet cards have a unique Ethernet address
- Broadcast address:
  - ff:ff:ff:ff:ff:ff
- Ethernet Multicast address range for IP:
  - 01:00:5e:00:00:00 -to- 01:00:5e:7f:ff:ff

Mapping IP Multicast onto Ethernet Multicast

- IP Multicast (class D IP address):
  - Class D: 224.x.x.x-239.x.x.x (in HEX: Ex.xx.xx.xx): 28 bits
  - No further structure (like Class A, B, or C)
  - Not addresses but identifiers of groups
  - Some of them are assigned by the IANA to *permanent host groups*

- Mapping a class D IP adr. into an Ethernet multicast adr.
  - The least 23 bits of the Class D address are inserted into the 23 bits of ethernet multicast address
  - Many to one mapping: 5 bits are not used
  - More filtering has to be done at IP level

IP Mutlicast: Problems to Solve

- Build on top of the existing Internet and take into account group communication constraints
  - Manage groups
  - Create and maintain multicast routes
  - Efficient end-to-end delay (reliability, flow control, time constraints)

Shortest Path Tree Routing Algorithm

- Apply point-to-point shortest path for all the receivers
- Multiple sources compute different trees
- For dynamic networks: 2 techniques to gather info
  - Distance vector algorithm
    - Each router sends to its neighbors its distance to the sender (called vector distance)
    - After receiving the vector distance from its neighbors, each router computes its own vector distance (minimum(received_vectors)+cost-to-neighbor)
  - Link state algorithm
    - Network connectivity information is broadcast to all routers
    - Every router has a complete knowledge of the network state
    - Every router centrally computes (using Dijkstra’s algorithm) the shortest path to the sender

Minimum Cost Tree Routing Algorithm

- Goal: minimize the overall cost of the multicast tree
- Minimum Spanning Tree:
  – Minimum cost tree which spans all nodes (Prim-Dijkstra's algorithm: add nearest members one by one to the tree)
  – Example:
- Minimum Steiner Tree:
  – Minimum cost tree which spans at least all the group members
  – This problem is NP-complete: we don't have an algorithm that can solve it in polynomial time of the size of the graph (stays NP-complete when link cost = 1, planar graph, bipartite graph)
  – Heuristics exist for approximating the minimum Steiner tree

Constrained Tree Routing Algorithm

- Goal: minimize both the distance between the sender and the receiver (delay) and the overall tree cost (bandwidth)

- Reason: real applications have constraints on delay/cost.

- Heuristics:
  - e.g., [Kompella, Pasquale, Polyzos 93: IEEE/ACM Trans. Net.]

Practical Systems

- MOSPF: shortest path algorithm (link-state Dijkstra's Alg.)
- DVMRP: distributed implementation of Shortest Path (Bellman-Ford Alg.)
- CBT: center-based tree
- PIM (sparse mode): center-based tree + Bellman-Ford

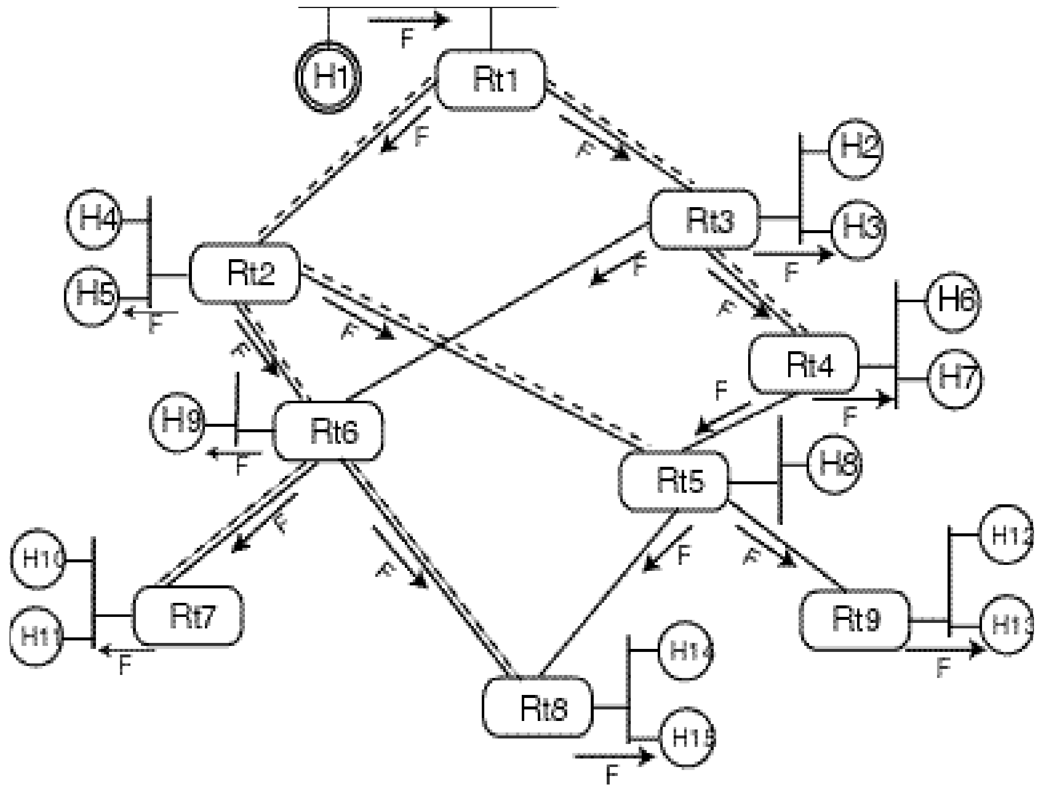
Multicast Routing Protocols: The Evolution

- Reverse Path Forwarding (RPF)
- Internet Group Management Protocol
- Truncated Broadcasting
- Distance Vector Multicast Routing Protocol (DVMRP)
- Multicast extensions to Open Shortest Path First (MOSPF)
- Protocol Independent Multicast (PIM)
- Core Based Tree (CBT)
- Ordered Core Based Tree (OCBT)
- Hierarchical DVMRP (HDVMRP)
- Hierarchical PIM (HPIM)
- Border Gateway Multicast Protocol (BGMP)

Reverse Path Forwarding

- If a router receives a packet on the interface that leads to the multicast sender, he forwards the packet on the other interfaces. Otherwise, he drops the packet
- This protocol achieves broadcasting, but not multicasting
- We need a mechanism to know where are the members of the group

Illustration of RPF



Internet Group Management Protocol [RFC1112]

- IGMP router periodically broadcasts a *Host-Membership Query* on its subnet
- If there is a host subscribing to the group, the host schedules a random timer to send an *IGMP Host-Membership Report*
- When the timer expires the *IGMP H-M Report* is multicasted. The purpose of this report is:
  - The other members of the group in the same subnet cancel their timer
  - The router knows that there is a member on its subnet listening to a given group

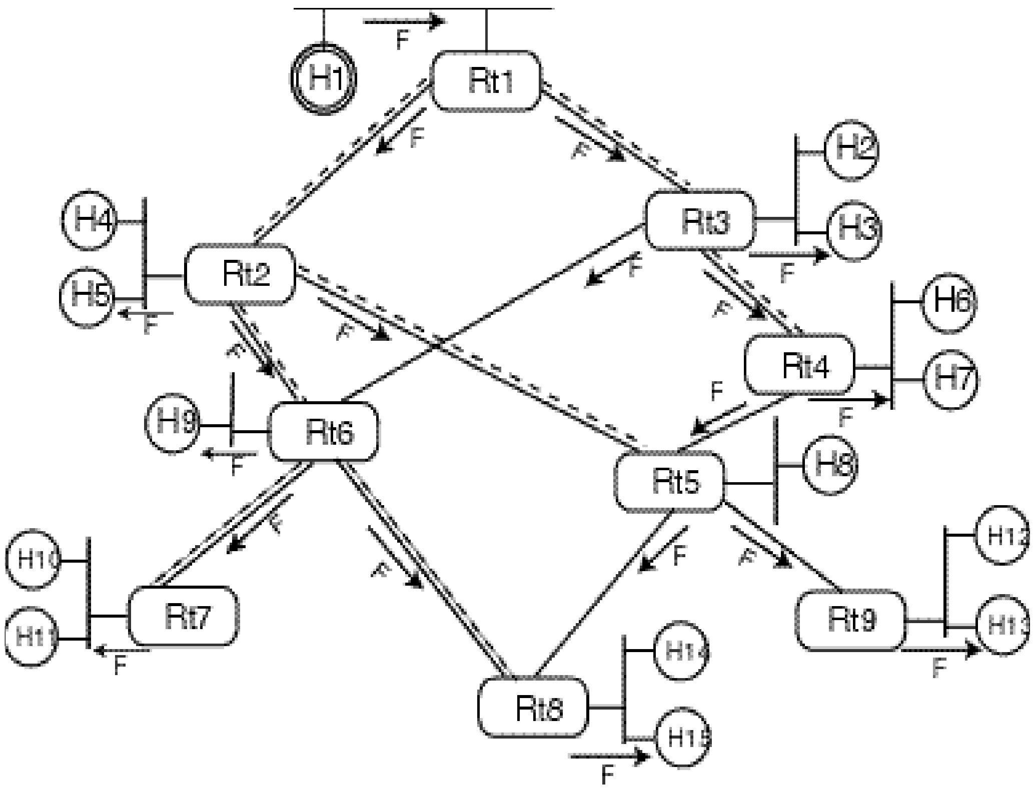
Truncated Broadcasting

- Uses the group membership information to decide if the packets will be broadcast on the leaf subnet
- Reduces the traffic in the leaf subnet
- Does not reduce the traffic in the core network

Distance Vector Multicast Routing Protocol (DVMRP): RFC1075(1988-97)

- Distance vector routing
  - Similar to RIP and extended to multicast routing
  - Extends truncated broadcast by using *pruning* and *grafting*
  - *Soft-state* protocol: pruning and flooding is periodically repeated
- Pruning:
  - On reception of a flooded packet by a leaf-router:
    - if the leaf- router is not interested (no members) it sends a *prune* message to all its neighbors
    - otherwise it sends the *prune* message only on the interfaces different from the *reverse shortest path*
  - If a router receives a prune on all its interfaces except the reverse shortest path, it propagates the prune through the reverse shortest path
- Grafting: If a host wants to join before the next flooding:
  - a graft is forwarded upstream (RPF) to the closest router in the tree

Illustrating DVMRP



Summary of some of the problems

- Flooding/pruning:
  - good for small dense networks
  - bad in poorly populated networks
- Sender specific trees:
  - low delay
  - complex routing tables
- Shared trees:
  - small routing tables
  - traffic concentration, non-optimal delay
- Steiner trees:
  - optimal overall cost
  - too complex to compute on the fly

Protocol Independent Multicast (PIM: 1996)

- Goals:
  - does not depend on any unicast protocol
  - optimize traffic depending on the density of receivers in the region
  - low-latency data distribution (source-based trees instead of shared-trees)
- Modes:
  - Dense mode: flooding
  - Sparse mode: use Rendezvous Points (RPs)
- Sparse mode regions:
  - number of networks/domains with members is significantly smaller than the total number of networks/domains in the region
  - group members are widely distributed
  - overhead of flooding + pruning is high

Components of PIM

- Rendezvous Point (RP):
  - each multicast group uses one RP:
    - (SM) receivers explicitly join the group by sending a *JOIN* to the RP
    - senders unicast to the RP, which sends the packets on the shared tree
- Designated Router (DR):
  - each sender/receiver communicates with a directly connected router (PIM-Reg: Join/Prune)
  - the DR may be the IGMP querier
- Last Hop Router (LHR):
  - router directly connected to the receiver: forwards the multicast packets
  - generally: LHR = DR
- Boot Strap Router: elected router within a domain
  - constructs the set of RP and distribute it to the routers in the domain

Key Steps of PIM

- Creating the PIM framework:
  - some routers are configured as candidate RPs (C-RPs)
  - C-RPs periodically send C-RP-Advs to the BSR
  - BSR distributes the RP-set to all the routers (Bootstrap Messages: BSM)
  - any router: RP-set + Group Address -> RP for the group
- Multicast shared tree:
  - Receiver join:
    - IGMP-report message from receiver to DR
    - DR creates an entry (*, G), DR sends a PIM Join/Prune message to RP
  - Source Join:
    - IGMP-report message from sender to DR
    - Data packets are unicast to the RP by the DR: PIM-register
    - Packets are forwarded through the shared tree (if there is no (S, G) entry: no shortest path tree)

Key Steps of PIM (*Cont'd*)

- Switching from shared tree to shortest path tree:
  - PIM starts with a shared tree (RP-tree)
  - when the traffic > TH, the receiver DR/LHR initiates the switch:
    - creates a source specific entry (S1, G)
    - sends a PIM <u>Join</u>/Prune to the sender through the next best hop router for S1
    - intermediate routers send a PIM <u>Join</u>/Prune to the sender on the shortest path
    - intermediate routers send a PIM Join/<u>Prune</u> to the RP if the path to the RP is different from the shortest path
- Steady state maintenance:
  - soft state protocol: periodic join/prune messages
- Data forwarding:
  - first check for a (S, G) entry: SPT, otherwise for (*, G): shared tree
- Multi-access network: resolution of multipath, ...

Multicast in IPv6

- Multicast address format (128 bits): FF.FlagScope.G-ID
  - Flag (4bits):
    - 0: permanently assigned group (NTP, …)
    - 1: transient group
    - others: undefined
  - Scope (4bits):
    - limits of transmission (node, link, site, organization)
  - Group-ID (112bits):
    - unique group ID
    - reserved values: 0 (never used), 1: all nodes, 2: all routers
- Group-ID is assigned using random number generators
- IGMP is incorporated inside ICMP

Multicast Backbone (Mbone)

- Multicast chicken-and-egg problem:
  - multicast cannot be deployed (and fully tested) without the support of router vendors
  - router vendors would not support IP multicast before it is mature and robust
- Mbone solution:
  - connect multicast capable routers using IP tunnels
  - First IP tunnel 1988: BBN (Boston) and Stanford University
  - IEEE INFOCOM, IEEE GLOBECOM, ACM SIGCOMM over MBone
- Tunneling:
  - IP multicast packets are encapsulated into unicast packets and sent to next-hop MBone router
  - Next MBone router strip off the outer packet header:
    - multicast to its subnet (if there is any members)
    - re-encapsulate the packet and send it to the next-hop using IP tunnel

Mbone (*Cont'd*)

- Traffic level in the MBone
  - Upper limit per tunnel: 500 KBps
  - Typical conference sessions: 100-300 KBps
  - TTL (0-255) to limit the scope of sessions

- MBone tools
  - session directory (sd, sdr)
  - audio conferencing tool (vat, nevot, rat)
  - video conferencing tool (nv, ivs, vic, nevit)
  - shared whiteboard tool (wb)
  - Network text editor (nte)