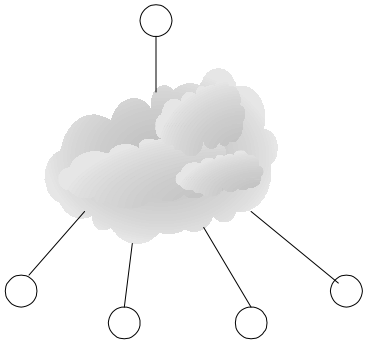
IP Multicast



Shivkumar Kalyanaraman
shivkuma@ecse.rpi.edu
http://www.ecse.rpi.edu/Homepages/shivkuma

- Why IP multicast ? Multicast apps ...
- Concepts: groups, scopes, trees
- Multicast addresses, LAN multicast
- Group management: IGMP
- Multicast routing and forwarding: MBONE, PIM etc
- Reliable Multicast Transport Protocols

Why IP multicast ?

- Need for efficient delivery to multiple destinations across inter/intranets
- Broadcast:
  - Send a copy to every machine on the net
  - Simple, but inefficient
  - All nodes "must" process the packet even if they don't care
  - Wastes *more* CPU cycles of slower machines ("*broadcast radiation*")
  - Network loops lead to "*broadcast storms*"

Why IP multicast ? (contd)

- Replicated Unicast:
  - Sender sends a copy to each receiver in turn
  - Receivers need to register or sender must be preconfigured
  - Sender is focal point of all control traffic
  - Latency = time between the first and last receiver getting a copy {can be large if transmission times are large}

Why IP multicast ?

- Application-layer relays:
  - A “relay” node or set of nodes does the replicated unicast function instead of the source
  - Multiple relays can handle “groups” of receivers and reduce number of packets per multicast => efficiency
  - Manager has to manually configure names of receivers in relays etc => too much administrative burden
- Alternative: build replication/multicast engine at the network layer

Multicast applications

- News/sports/stock/weather updates
- Distance learning
- Routing updates (OSPF, RIP etc)
- Pointcast-type “push” apps
- Videoconferencing, shared whiteboards
- Distributed interactive gaming or simulations
- Email distribution lists
- IPv6 over IPv4
- Voice-over-IP
- Database replication

Multicast apps characteristics

- Number of (simultaneous) senders to the group
- The size of the groups
  - Number of members (receivers)
  - Geographic extent
  - Diameter of the group measured in router hops
- The longevity of the group
- Number of aggregate packets/second
- The peak/average used by source
- Level of human interactivity
  - Lecture mode vs interactive
  - Data-only (eg database replication) vs multimedia

IP multicast concepts

- Message sent to multicast "group" (of receivers)
  - Senders need not be group members
  - Each group has a "group address"
  - Use "group address" instead of destination address in IP packet sent to group
  - Groups can have any size;
  - End-stations (receivers) can join/leave at will

IP multicast concepts (contd)

- Packets are not unnecessarily duplicated or delivered to destinations outside the group
  - Distribution tree for delivery/distribution of packets
  - Packets forwarded "away" from the source
  - No more that one copy of packet appears on any subnet
  - Packets delivered only to "interested" receivers => multicast delivery tree changes dynamically
  - Network has to actively discover paths between senders and receivers

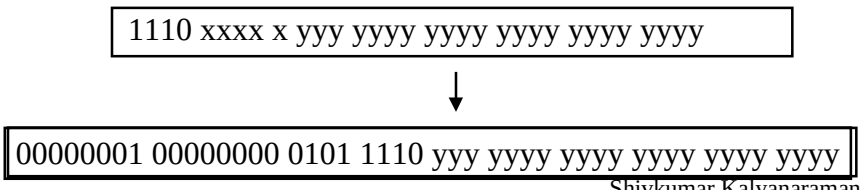
  - Non-member nodes even on a single subnet do not receive packets (unlike subnet-specific broadcast)
- Group membership on a single subnet is achieved through IGMP (and ICMPv6 in IPv6)
- Tree is built by multicast routing protocols. Current multicast tree over the internet is called MBONE.
- Anycast: delivers a packet to *one of a group* (hopefully the closest)

Multicast addresses

- Class D addresses: 224.0.0.0 thru' 239.255.255.255
- Each multicast address represents a *group of arbitrary size, called a "host group"*
- There is no structure within class D address space like subnetting => flat address space
- Addresses 224.0.0.x and 224.0.1.x are reserved. See assigned numbers RFC 1700
  - Eg: 224.0.0.2 = all routers on this subnet
- Addresses 239.0.0.0 thru 239.255.255.255 are reserved for private network (or intranet) use

Multicast IP over IEEE 802 LANs

- MAC address = 6 bytes = OUI + 3-byte address
- Special OUI for IETF: 0x01-00-5E.
  - Of remaining 3 bytes (24 bits), one bit is reserved
  - Remaining 23 bits = lower order 23 bits from IP class D address. Simpler than unicast forwarding ! No ARP etc.
- 32 class D addrs may map to one MAC addr



Multicast over LANs & Scoping

- Multicasts are flooded across MAC-layer bridges along a spanning tree
  - LAN NICs must be specifically programmed to filter multicasts on behalf of the end station
  - But flooding may steal sending opportunity for nonmember stations which want to transmit
- Scope: How far do transmissions propagate?
- Implicit scoping: Reserved Mcast addresses => don't leave subnet. Also called "link-local" addresses

Scope of multicast forwarding

- TTL-based scoping:
  - Each multicast router has a configured TTL threshold
  - It does not forward multicast datagram if TTL <= TTL-threshold
  - Useful at edges of a large intranet as a blanket parameter.
- Administrative scoping:
  - Use a portion of class D address space (239.0.0.0 thru 239.255.255.255)
  - Truly local to admin domain; address reuse possible.
  - In IPv6 scoping is an internal attribute of an an IPv6 multicast address

IGMP

- IGMP: “signaling” protocol to establish, maintain, remove groups on a subnet.
- Router sends *Host Membership Query* to 224.0.0.1 (all multicast hosts on subnet)
- Host responds with *Host Membership report* for each group to which it belongs, *sent to group address*
- Membership report => other hosts in the same group “suppress” reports
- Router periodically broadcasts query to detect if groups have gone away
- Asynchronous reports possible.

IGMPv2

- Distributed with the mrouted source code
- Has a querier election protocol (lowest IP address)
- Hosts may send a “Leave group” message to “all routers” (224.0.0.2) address
  - Querier responds with a Group-specific Query message to see if any group members are available
  - Lower leave latency => responds quickly to membership changes
- Router alert IP option is also enabled
- Bunch of rules for coexistence of IGMPv1 and v2 hosts and routers on a single subnet

Multicast Routing Protocols

- Multicast routing protocols build trees where the "leaves" are the subnets containing at least one group member (detected by IGMP)
- Tree types:
  - ***Source-based trees:*** one tree per (source, group) pair
  - ***Shared trees:*** one tree per group
- Tree building methods:
  - ***Data driven***: calculate the tree only when the first packet is seen
    - ***Broadcast-and-prune:*** Multicast tree = broadcast tree - non-multicast branches

Multicast Routing Protocols

  - Run Dijkstra's algorithm to build tree when first packet is seen (MOSPF)
  - A priori: Build tree before any data is transmitted
- Join-styles:
  - ***Explicit-join:*** The leaves explicitly join the tree
  - ***Implicit-join:*** All subnets are assumed to be receivers unless they say otherwise (eg via tree pruning)
- Modes:
  - ***Dense-mode:*** many (or closely located) subnets have at least one group member
  - ***Sparse-mode***: few (or widely separated/bandwidth-limited) subnets have at least one group member

Reverse Path Multicast (RPM)

- Setup broadcast tree (***reverse path broadcasting, RPB***)
  - Each node maintains "parent" and "child" links
  - If packet from parent (***"reverse-path check")*** send to children; else drop
  - If child is actually downstream (eg in terms of the routing metric), remove the child link
- ***Truncated RPB (TRPB):*** Truncate ***leaf*** if IGMP says that there are no receivers for the group.
- ***Reverse-Path Multicasting (RPM):*** truncate ***branch*** if IGMP says that there are no receivers for the group

DVMRP

- RPM forwarding tree built on demand from a DVMRP group-independent routing table
- Source-based trees, data-driven (broadcast-and-prune), implicit join, dense mode
- TTL and admin scoping available; physical or tunnel interfaces possible
- Limitations:
  - distance-vector => slow to adapt to topology changes;
  - Must store source-specific sate even when not on tree => more scaling problems
  - No hierarchy (flat routing domain)

MBONE

- Internet Multicast Backbone: testbed
- Thousands of regions connected by virtual point-to-point links called “tunnels”.
  - Multicast traffic passes through non-multicast regions using IP-in-IP encapsulation.
  - Intermediate routers see only wrappers (regular IP-unicast packets)
  - Tunnel endpoint recognizes IP-in-IP (protocol type = 4) and decapsulates datagram for processing
- MBONE uses DVMRP (mrouted)
  - Limited to few senders. Many small groups also undesired
- Tools: sdr (session directory), vic, vat, wb

Protocol-Independent Multicast

- PIM has two variants: Dense mode (DM) and sparse mode (SM)
  - DM builds source-based trees in a data-driven (broadcast-and-prune), implicit join manner
  - SM allows both source-based and shared trees. But the trees are built a priori and using explicit join.
- Not dependent upon mechanisms provided by any particular unicast protocol. Can leverage upon RIP, OSPF, BGP-4 etc
- PIM: broadcasts on all non-incoming interfaces until explicit prune messages are received

MOSPF

- Flood the multicast group membership information along with the link states
- The shortest path multicast tree is built upon demand using Dijkstra's algorithm
  - Note that all routers calculate the same source-based shortest-path delivery tree
  - The datagram is not flooded, only the group membership info is flooded
- For each transmission, determine the downstream branch and forward the packet
  - Use caching to avoid tree calculation for each packet
  - The forwarding is not TTL based

Core-Based Trees (CBT)

- Sparse Mode: shared tree set up before forwarding. Good scaling properties for WAN multicast, with scattered receivers
- Each group has a "core router" which is dynamically discovered (bootstrapping)
- A host which wants to join the group sends a JOIN_REQUEST towards the core and gets a JOIN_ACK from the nearest router already on the tree.
- Forwarding cache = group, {outgoing interface list}
  - Packet is forwarded onto all outgoing interfaces except the one in which it arrived
  - Packet transmission can be bidirectional ("upstream" in a CBT refers to the direction towards the core, not the source)
- Non-tree source employs IP-in-IP encapsulation to send packets to the core

Reliable Multicast Transport

- Problems:
  - Retransmission can make reliable multicast as inefficient as replicated unicast
  - Ack-implosion if all destinations ack at once
  - Source does not know # of destinations
  - "Crying baby": one bad link affects entire group
  - Heterogeneity: receivers, links, group sizes
  - Not all multicast applications need reliability of the type provided by TCP. Some can tolerate reordering, delay, etc
- Egs: Scalable Reliable Multicast (SRM), Lightweight Reliable Multicast Protocol (LRMP), Reliable Multicast Transport Protocol (RMTP), Pragmatic General Multicast (PGM)

Scalable Reliable Multicast (SRM)

- All members get all the data that has been sent to the the multicast group (minimalist reliability)
- Repair requests and responses (retransmissions) are multicast.
- Scope of repair requests and responses can be TTL limited or a separate "local recovery group" can be formed
- Techniques to avoid implosion of repair requests, and reduce control traffic
- An example of an "application level framing" paradigm {like RTP}

Summary

- IP multicast issues and applications
- Multicast over LANs and scoping
- IGMP
- Multicast Routing and MBONE
- Reliable multicast transports