11: IP Multicast

Last Modified:
4/9/2003 1:15:00 PM

Based on slides by Gordon Chaffee
Berkeley Multimedia Research Center
URL: http://bmrc.berkeley.edu/people/chaffee

Outline

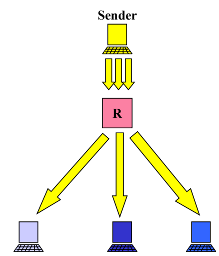
- IP Multicast
- Multicast routing
  - Design choices
  - Distance Vector Multicast Routing Protocol (DVMRP)
  - Core Based Trees (CBT)
  - Protocol Independent Multicast (PIM)
  - Border Gateway Multicast Protocol (BGMP)
- Issues in IP Multicast Deplyment

What is multicast?

- 1 to N communication
- Nandwidth-conserving technology that reduces traffic by simultaneously delivering a single stream of information to multiple recipients
- Examples of Multicast
  - Network hardware efficiently supports multicast transport
    - Example: Ethernet allows one packet to be received by many hosts
  - Many different protocols and service models
    - Examples: IETF IP Multicast, ATM Multipoint

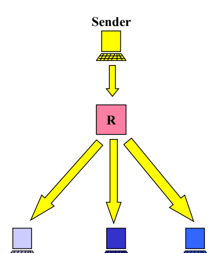
Unicast

- Problem
  - Sending same data to many receivers via unicast is inefficient
- Example
  - Popular WWW sites become serious bottlenecks



Multicast

- Efficient one to many data distribution



IP Multicast Introduction

- Efficient one to many data distribution
  - Tree style data distribution
  - Packets traverse network links only once
- Location independent addressing
  - IP address per multicast group
- Receiver oriented service model
  - Applications can join and leave multicast groups
  - Senders do not know who is listening
  - Similar to television model
  - Contrasts with telephone network, ATM

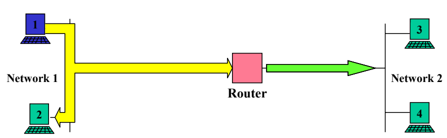
IP Multicast

- Service
  - All senders send at the same time to the same group
  - Receivers subscribe to any group
  - Routers find receivers
- Unreliable delivery
- Reserved IP addresses
  - 224.0.0.0 to 239.255.255.255 reserved for multicast
  - Static addresses for popular services (e.g. Session Announcement Protocol)

Internet Group Management Protocol (IGMP)

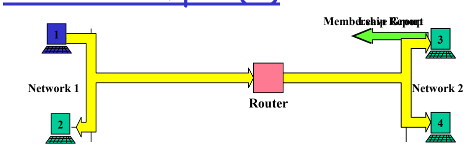
- Protocol for managing group membership
  - IP hosts report multicast group memberships to neighboring routers
  - Messages in IGMPv2 (RFC 2236)
    - Membership Query (from routers)
    - Membership Report (from hosts)
    - Leave Group (from hosts)
- Announce-Listen protocol with Suppression
  - Hosts respond only if no other hosts has responded
- Soft State protocol

IGMP Example (1)



- Host 1 begins sending packets
  - No IGMP messages sent
  - Packets remain on Network 1
- Router periodically sends IGMP Membership Query

IGMP Example (2)



- Host 3 joins conference
  - Sends IGMP Membership Report message
- Router begins forwarding packets onto Network 2
- Host 3 leaves conference
  - Sends IGMP Leave Group message
  - Only sent if it was the last host to send an IGMP Membership Report message

Source Specific Filtering: IGMPv3

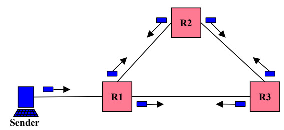
- Adds Source Filtering to group selection
  - Receive packets **only** from specific source addresses
  - Receive packets from **all but** specific source addresses
- Benefits
  - Helps prevent denial of service attacks
  - Better use of bandwidth
- Status: Internet Draft?

Multicast Routing Discussion

- What is the problem?
  - Need to find all receivers in a multicast group
  - Need to create spanning tree of receivers
- Design goals
  - Minimize unwanted traffic
  - Minimize router state
  - Scalability
  - Reliability

Data Flooding

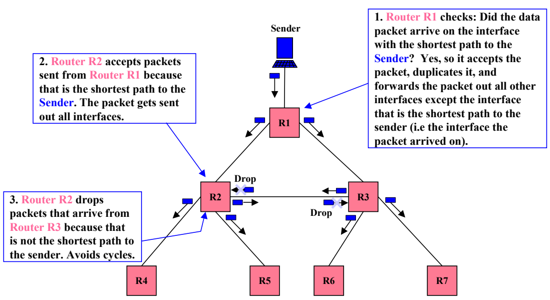
- Send data to all nodes in network
- Problem
  - Need to prevent cycles
  - Need to send only once to all nodes in network
  - Could keep track of every packet and check if it had previously visited node, but means too much state



Reverse Path Forwarding (RPF)

- Simple technique for building trees
- Send out all interfaces except the one with the shortest path to the sender
- In unicast routing, routers send to the destination via the shortest path
- In multicast routing, routers send away from the shortest path to the sender

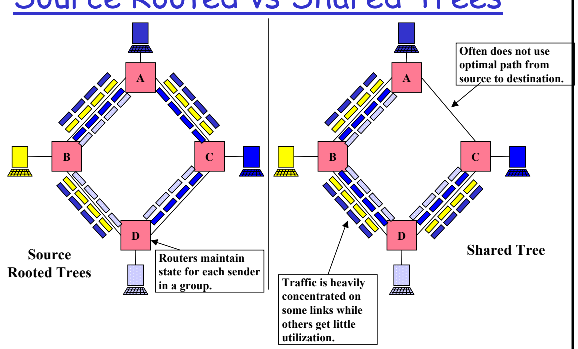
Reverse Path Forwarding Example



Data Distribution Choices

- Source rooted trees
  - State in routers for each sender
  - Forms shortest path tree from each sender to receivers
  - Minimal delays from sources to destinations
- Shared trees
  - All senders use the same distribution tree
  - State in routers only for wanted groups
  - No per sender state (until IGMPv3)
  - Greater latency for data distribution

Source Rooted vs Shared Trees



Distance Vector Multicast Routing (DVMRP)

- Steve Deering, 1988
- Source rooted spanning trees
  - Shortest path tree
  - Minimal hops (latency) from source to receivers
- Extends basic distance vector routing
- Flood and prune algorithm
  - Initial data sent to all nodes in network(!) using Reverse Path Forwarding
  - Prunes remove unwanted branches
  - State in routers for all unwanted groups
  - Periodic flooding since prune state times out (soft state)

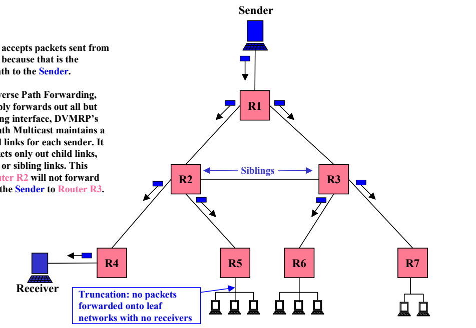
DVMRP Algorithm

- Truncated Reverse Path Multicast
  - Optimized version of Reverse Path Forwarding
  - Truncating
    - No packets sent onto leaf networks with no receivers
  - Still how "truncated" is this?
- Pruning
  - Prune messages sent if no downstream receivers
  - State maintained for each unwanted group
- Grafting
  - On join or graft, remove prune state and propagate graft message

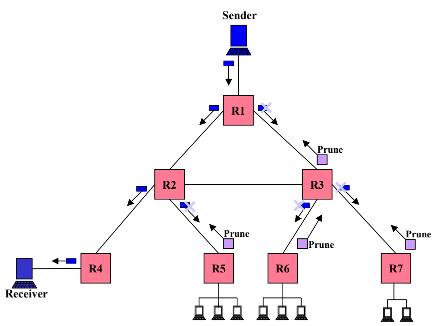
Truncated Reverse Path Multicast Example

Router R2 accepts packets sent from Router R1 because that is the shortest path to the Sender.

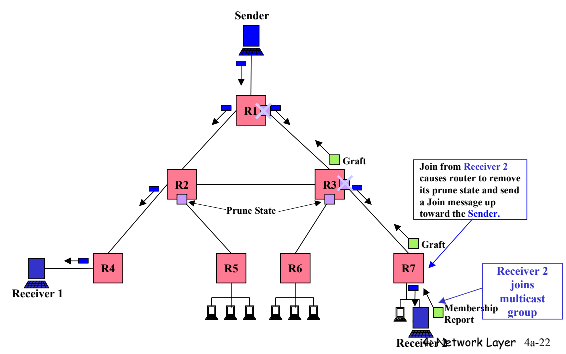
Unlike Reverse Path Forwarding, which simply forwards out all but the incoming interface, DVMRP's Reverse Path Multicast maintains a list of child links for each sender. It sends packets only out child links, not parent or sibling links. This means Router R2 will not forward data from the Sender to Router R3.



DVMRP Pruning Example



DVMRP Grafting Example



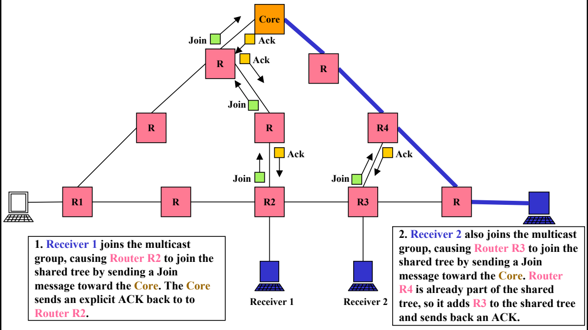
DVMRP Problems

- State maintained for unwanted groups
- Bandwidth intensive
  - Periodic data flooding per group
    - No explicit joins, and prune state times out
  - Not suitable for heterogeneous networks
- Poorly handles large number of senders
  - Scaling = O(Senders, Groups)
- Problems of distance vector routing
  - slow convergence
  - cycles due to lack of global knowledge

Core Based Trees (CBT)

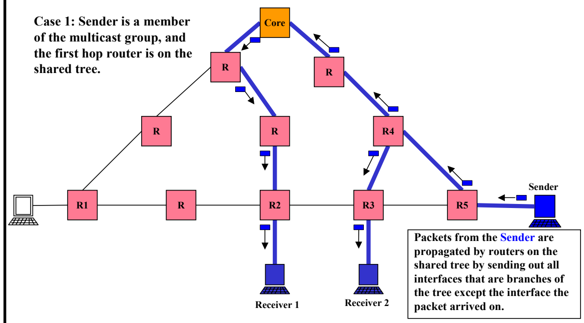
- Attributes
  - Single shared tree per group => sparse trees
  - Large number of senders
  - Routing tables scale well, size = O(Groups)
  - Bi-directional tree

Group Management in CBT



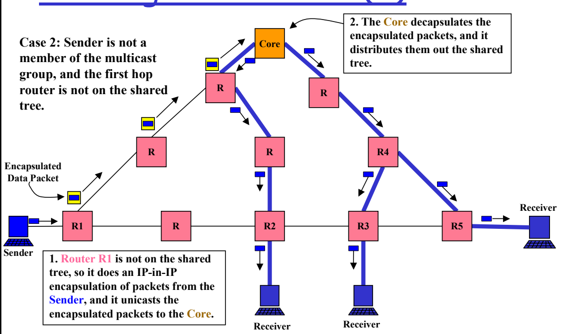
Sending Data in CBT (1)

Case 1: Sender is a member of the multicast group, and the first hop router is on the shared tree.



Sending Data in CBT (2)

Case 2: Sender is not a member of the multicast group, and the first hop router is not on the shared tree.



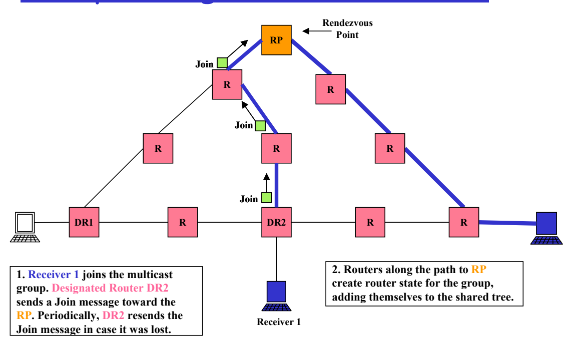
Protocol Independent Multicast (PIM)

- Uses unicast routing table for topology
- Dense mode (PIM-DM)
  - For groups with many receivers in local/global region
  - Like DVMRP, a flood and prune algorithm
- Sparse mode (PIM-SM)
  - For groups with few widely distributed receivers
  - Builds shared tree per group, but may construct source rooted tree for efficiency
  - Explicit join

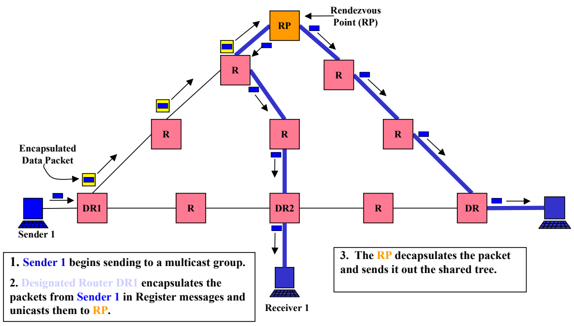
PIM Sparse Mode

- Hybrid protocol that combines features of DVMRP and CBT
- Suited to widely distributed, heterogeneous networks
- Shared tree centered at Rendezvous Point (RP)
- Shared tree introduces sources to receivers
- Source specific trees for heavy traffic flows
- Unidirectional distribution tree

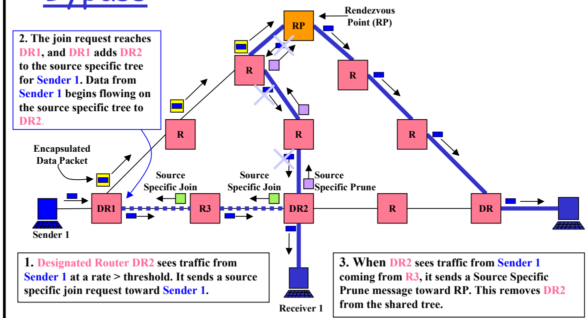
Group Management in PIM-SM



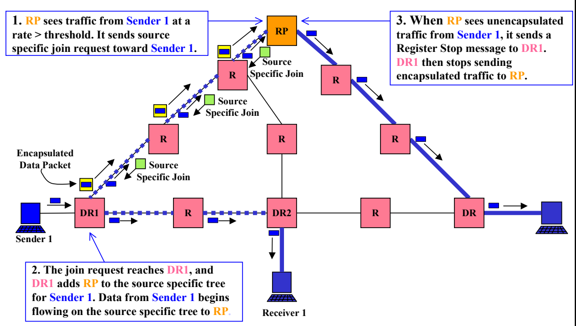
Sending Data in PIM-SM



PIM-SM Source Specific Bypass



RP Joins Source Specific Tree



Problems with PIM

- Global broadcasts of all Rendezvous Points
- Sensitive to location of RP
- No administrative control over multicast traffic; policy controls lacking
- Conceived as inter-domain, but now considered intra-domain

Classification of Tree Building Choices

- Flood network topology to all routers
  - Link state protocol
  - Multicast Extensions to OSPF (MOSPF)
- Flood and prune
  - Distance Vector Multicast Routing Protocol (DVMRP)
  - Protocol Independent Multicast Dense Mode (PIM-DM)
- Explicit join
  - Core Based Trees (CBT)
  - Protocol Independent Multicast Sparse Mode (PIM-SM)

Border Gateway Multicast Protocol (BGMP)

- Administrative control of multicast traffic
- Hierarchical multicast address allocation
- Uses BGP for routing tables
- No global broadcasts of anything
- Bi-directional shared multicast routing tree

IP Multicast in the Real World

Commercial Motivation

- Problem
  - Traffic on Internet is growing about 100% per year
  - Router technology is getting better at 70% per year
  - Routers that are fast enough are very expensive
- ISPs need to find ways to reduce traffic
- Multicast could be used to…
  - WWW: Distribute data from popular sites to caches throughout Internet
  - Send video/audio streams multicast
  - Software distribution

ISP Concerns

- Multicast causes high network utilization
  - One source can produce high total network load
  - Experimental multicast applications are relatively high bandwidth: audio and video
  - Flow control non-existent in many multicast apps
- Multicast breaks telco/ISP pricing model
  - Currently, both sender and receiver pay for bandwidth
  - Multicast allows sender to buy less bandwidth while reaching same number of receivers
  - Load on ISP network not proportional to source data rate

Economics of Multicast

- One packet sent to multiple receivers
- Sender
  - \+ Benefits by reducing network load compared to unicast
  - \+ Lower cost of network connectivity
- Network service provider
  - \- One packet sent can cause load greater than unicast packet load
  - \+ Reduces overall traffic that flows over network
- Receiver
  - = Same number of packets received as unicast

Multicast Problems

- Multicast is immature
  - Immature protocols and applications
  - Tools are poor, difficult to use, debugging is difficult
  - Routing protocols leave many issues unresolved
    - Interoperability of flood and prune/explicit join
    - Routing instability
- Multicast development has focused on academic problems, not business concerns
  - Multicast breaks telco/ISP traffic charging and management models
  - Routing did not address policy
    - PIM, DVMRP, CBT do not address ISP policy concerns
    - BGMP addresses some ISP concerns, but it is still under development

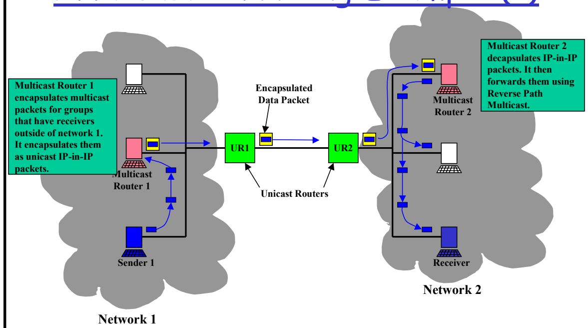
Current ISP Multicast Solution

- Restrict senders of multicast data
- Charge senders to distribute multicast traffic
  - Static agreements
- Do not forward multicast traffic
  - Some ISP's offer multicast service to customers (e.g. UUNET UUCast)
  - ISP beginning to discuss peer agreements

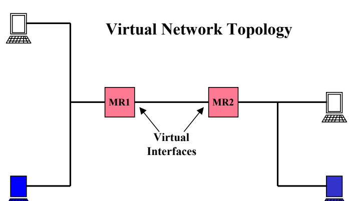
Multicast Tunneling

- Problem
  - Not all routers are multicast capable
  - Want to connect domains with non-multicast routers between them
- Solution
  - Encapsulate multicast packets in unicast packet
  - Tunnel multicast traffic across non-multicast routers
  - We will see more examples of tunneling later

Multicast Tunneling Example (1)



Multicast Tunneling Example (2)



MBone

- MBONE
  - Multicast capable virtual network, subset of Internet
  - Native multicast regions connection with tunnels
- In 1992, the MBone was created to further the development of IP multicast
  - Experimental, global multicast network
  - Served as a testbed for multicast applications development
    - vat -- audio tool
    - vic -- video tool
    - wb -- shared whiteboard

MBone Usage

- Dramatic increase in use...
  - Research: telecollaboration, protocol development
  - Learning: conferences, seminars, and classes
  - Entertainment: Rolling Stones concert
- Leads to much higher bandwidth demand
  - Groups range from < 10 to 1000's, will grow to millions
  - Number of programs/groups -- thousands of channels

Future?

Outtakes

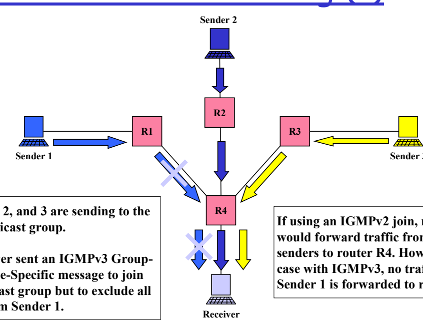
Multicast

- History
  - Long history of usage on shared medium networks
  - Data distribution
  - Resource discovery: DHCP , Bootp, ARP
- Ethernet
  - Broadcast (software filtered)
  - Multicast (hardware filtered)
- Multiple LAN multicast protocols
  - DECnet, AppleTalk, IP

Source Specific Filtering: IGMPv3

- Adds Source Filtering to group selection
  - Receive packets **only** from specific source addresses
  - Receive packets from **all but** specific source addresses
- Benefits
  - Helps prevent denial of service attacks
  - Better use of bandwidth
- Status: Internet Draft?

IGMPv3 Source Filtering (1)

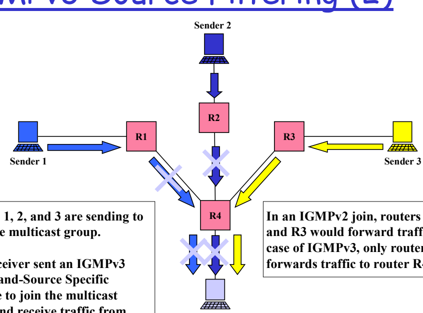


Senders 1, 2, and 3 are sending to the same multicast group.

The receiver sent an IGMPv3 Group-and-Source-Specific message to join the multicast group but to exclude all traffic from Sender 1.

If using an IGMPv2 join, router R1 would forward traffic from all senders to router R4. However, in this case with IGMPv3, no traffic from Sender 1 is forwarded to router R4.

IGMPv3 Source Filtering (2)



Senders 1, 2, and 3 are sending to the same multicast group.

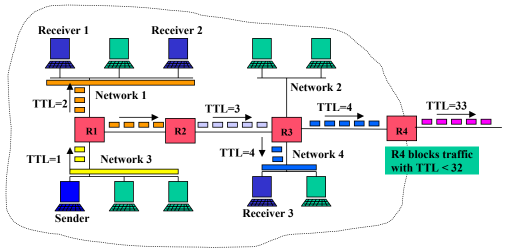
The Receiver sent an IGMPv3 Group-and-Source Specific message to join the multicast group and receive traffic from only Sender 3.

In an IGMPv2 join, routers R1, R2, and R3 would forward traffic. In the case of IGMPv3, only router R3 forwards traffic to router R4.

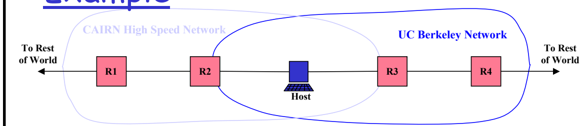
Scoping Multicast Traffic

- TTL based
  - Based on Time to Live (TTL) field in IP header
  - Only packets with a TTL > threshold cross boundary
- Administrative scoping
  - Set of addresses is not forwarded past domain
  - More flexible than TTL based.
- Scoped addresses
  - 224.0.0.* never leaves subnet

TTL Scoping Example



Administrative Scoping Example



- Administrative scoping allows traffic to be limited to a region based on its multicast group address, resulting in more flexible network configurations.
- The Host can send traffic that is limited to only the CAIRN High Speed Network, to only the UC Berkeley Network, to both, or to the rest of the world.
- 239.2.0.0 - 239.2.255.255: Traffic scoped to only the CAIRN High Speed Network
- 239.3.0.0 - 239.3.255.255: Traffic scoped to only the UC Berkeley Network
- 239.4.0.0 - 239.4.255.255: Traffic scoped to both the CAIRN and UC Berkeley Networks
- 224.0.1.0 - 238.255.255.255: Traffic scoped to the rest of the world

Reliable Multicast

- Some applications need the same data to be delivered reliably to many receivers
  - Distributed collaboration tools (e.g. shared whiteboard)
  - Stock history
  - Software distribution
- Status
  - Many different proposals
  - Proposals solve some problems but have not considered commercial limitations of multicast
  - Still exploring applications for reliable multicast

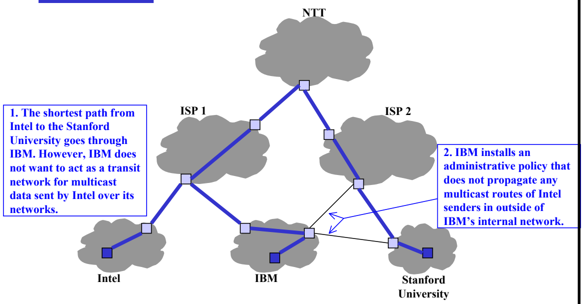
PIM Rendezvous Point (RP)

- Requirement
  - Different groups map to different RPs
- Bootstrap Router (BSR)
  - Dynamically elected
  - Constructs a set of RP IP addresses based on received Candidate-RP messages
- How do routers know RP for a group?
  - Bootstrap Router broadcasts Bootstrap message with RP set to PIM
  - Hash function on group address maps to an RP

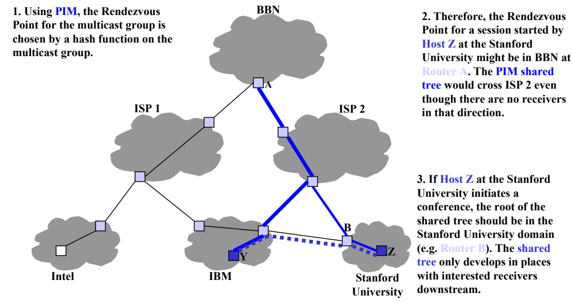
Border Gateway Multicast Protocol (BGMP)

- Motivation
  - Hierarchy for multicast routing
  - Combine design of multicast address allocation and multicast routing
  - Inter-domain routing protocols need administrative control of multicast traffic
- Scalability issues
  - Need to minimize router state
  - Need to minimize control messages
  - Only send data where it is needed

Administrative Control of Traffic



Choosing a Shared Tree Root



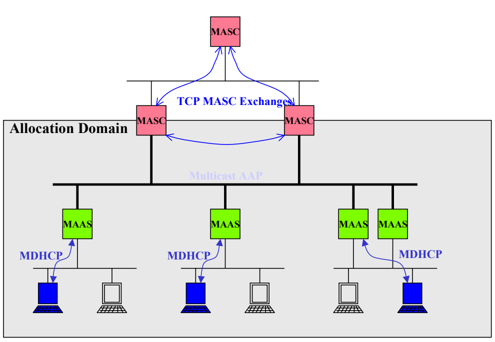
Multicast Address Allocation

- Problem
  - Multicast addresses are a limited resource
  - Current multicast address allocation scheme does not scale and makes multicast routing more difficult
- Solution
  - Use dynamically allocated addresses
  - Address allocation location determines root of shared tree
  - Hierarchical address allocation scales better and helps multicast routing

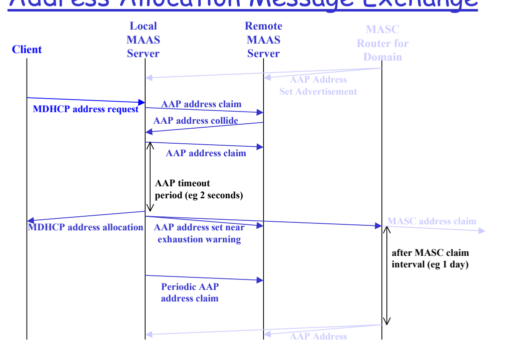
Multicast Address Allocation Architecture

- Multicast Address Set Claim (MASC)
  - Protocol to allocate multicast address sets to domains
  - Algorithm: Listen and claim with collision detection
  - Makes hierarchy available to routing infrastructure
- Address Allocation Protocol (AAP)
  - Protocol for allocating multicast addresses within domains
  - Used by Multicast Address Allocation Servers (MAAS)
- MDCHP (Multicast DHCP)
  - Protocol for end hosts to request multicast address
  - Extension to DHCP (Dynamic Host Configuration Protocol)

Multicast Address Allocation Example



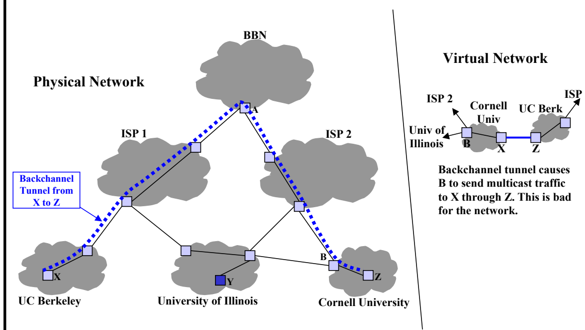
Address Allocation Message Exchange



Operational Problems

- Debugging is difficult
- Misconfigured routers inject unicast routing tables into multicast routing tables
- Black holes
  - Cisco to Cisco tunneling using DVMRP doesn't work
    - Routes exchanged, but no data flows
  - RPF checks on different routers think multicast traffic should be coming from the other router
- Backchannel tunnels
  - Improper tunnels cause non-optimal routing behavior

Backchannel Tunneling



Debugging Multicast Problems

- Local LAN debugging
  - tcpdump
    - `tcpdump ip multicast`
    - `tcpdump igmp`
- Routing debugging
  - mrinfo
  - mstat
  - mtrace