Z24: IP Multicast

Mark Handley

TV on the net

- Suppose you want to broadcast an event live to a million people worldwide, but your audience is too distributed for mainstream TV to carry. How do you do it?
- Classic example:
  - British and Indian communities in the US want to watch cricket.
  - Americans worldwide want to watch baseball.
- The Internet should be the ultimate minority media channel.
- And eventually the majority media distribution network.

Not just TV

- Conferencing (audio, video, whiteboard)
- Remote teaching
- Multi-user games
- Semi-synchronous streaming for video-on-demand.
- Distributed simulations
- Software distribution
- News distribution

Requirements

- Distribution should scale to large groups
  - Data must traverse each link only once.
- Robustness to link/router failures.
  - End-systems can't know the detailed distribution topology.

IP Multicast

- Dynamically constructs efficient shortest-path distribution trees from senders to receivers.
- Service Model:
  - Receivers announce their interest.
  - Senders just send
  - Routers conspire to deliver data from senders to receivers.

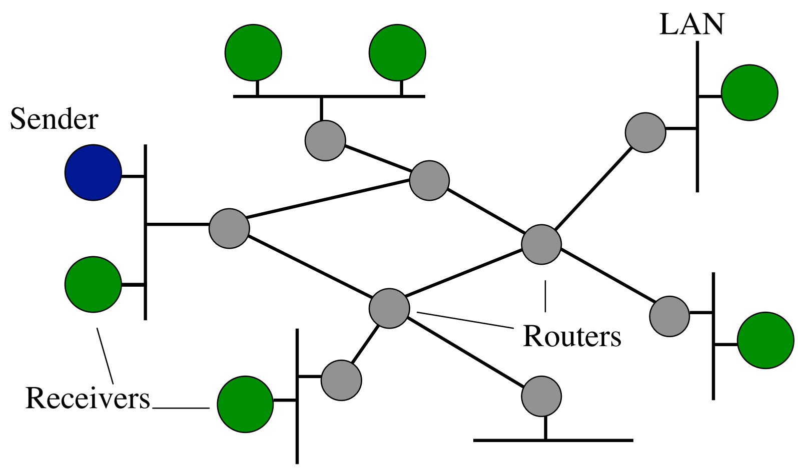
Class D addresses

- Traditionally, IP Addresses were divided into *classes*:
  - Class A for networks with millions of hosts
  - Class B for networks with thousands of hosts
  - Class C for networks with tens to hundreds of hosts
- *Class D* is multicast group addresses
  - A multicast sender just sends to a class D multicast address
  - Multicast receivers express an interest in a class D multicast address
  - If they choose the same address, the network delivers traffic from the senders to the receivers.
  - Class D addresses are in the range 224.0.0.0 to 239.255.255.255

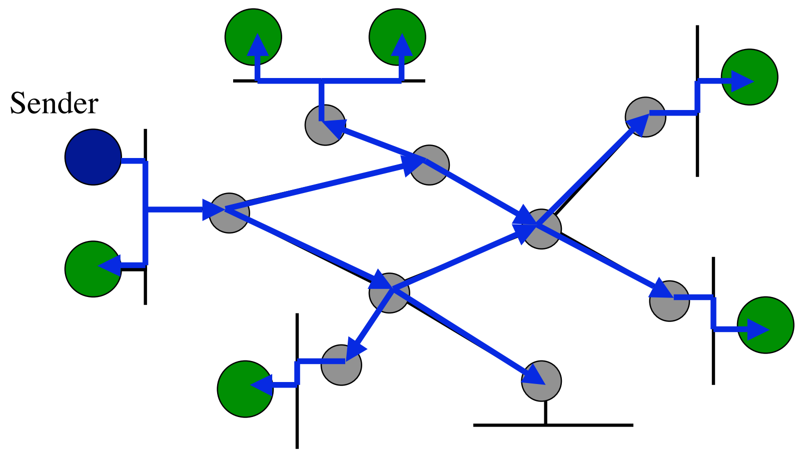
IP Multicast Routing Protocols

- **DVMRP**
  - Distance Vector Multicast Routing
  - The original multicast routing protocol.
- **MOSPF**
  - Multicast OSPF
- **CBT**
  - Core-Based Trees
  - Better scaling properties but not shortest path trees.
- **PIM**
  - Protocol Independent Multicast
  - Two modes: sparse and dense.
  - PIM Sparse-mode (PIM-SM) is widely deployed.

DVMRP and PIM Dense-Mode

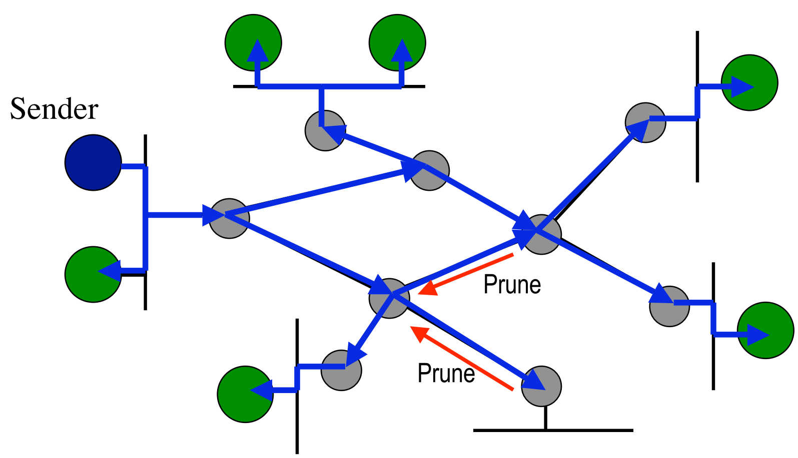


Broadcast and Prune



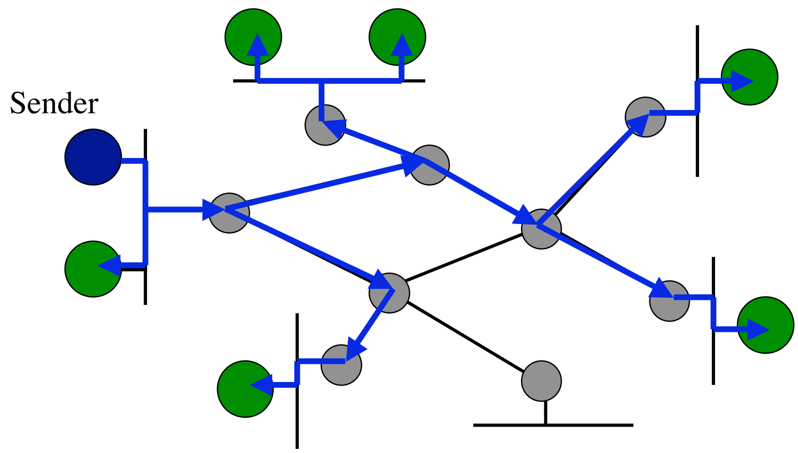
Basic model is “broadcast and prune”

Broadcast and Prune



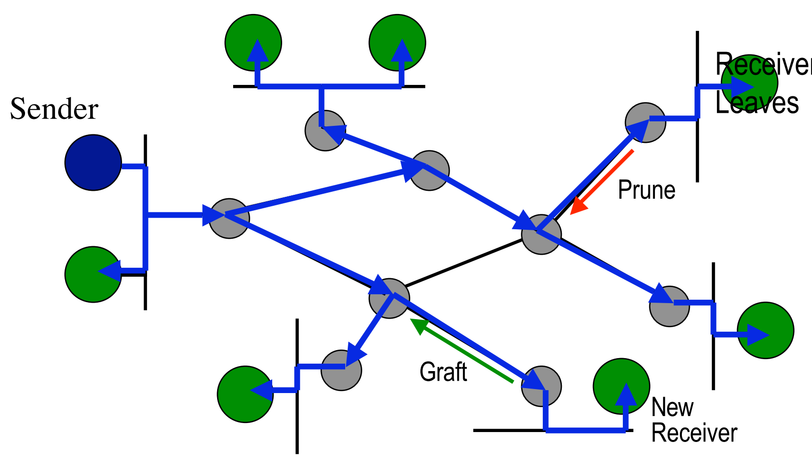
Branches not on the shortest path tree to a receiver are pruned off

Shortest Path Tree



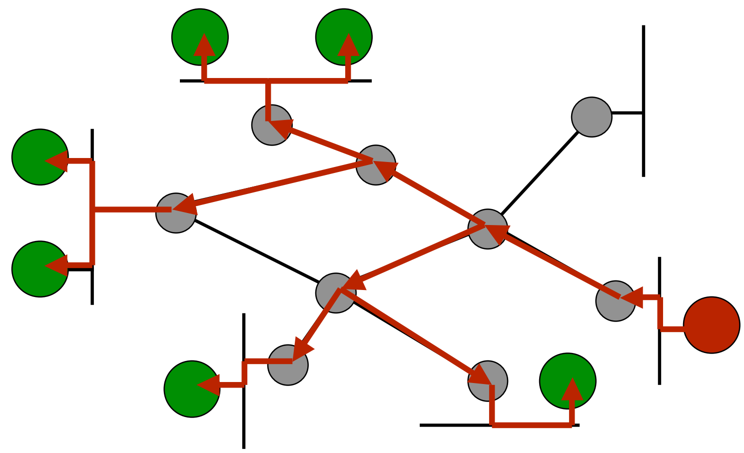
Resulting in a shortest path distribution tree, rooted at the sender

Local Tree Maintenance



Changes in membership are handled efficiently and locally

A Shortest Path Tree per Sender

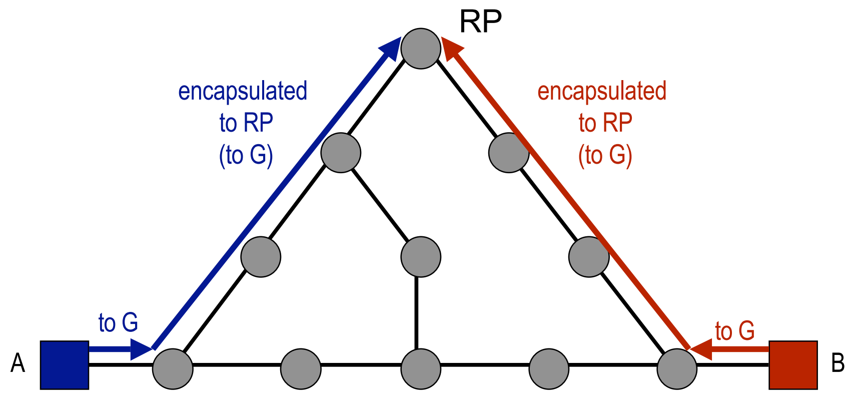


Distribution trees are per-{sender,group}, triggered by data pkts

Sparse-Mode PIM

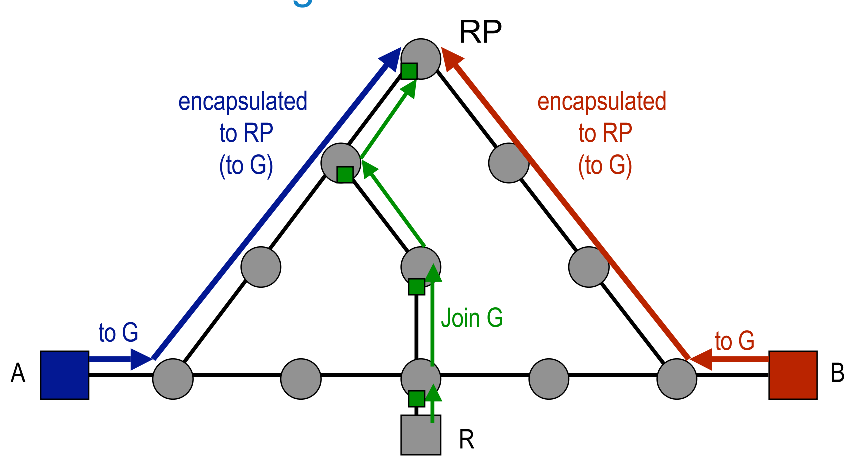
- Instead of flooding and pruning to directly build a shortest-path tree, PIM-SM initially builds a shared tree.
- The shared tree is built by sending *join messages* towards a *Rendezvous Point* (RP)
- Once data is flowing, the shared tree can be converted to a shortest-path tree if required.

PIM-SM: Building the Shared Tree



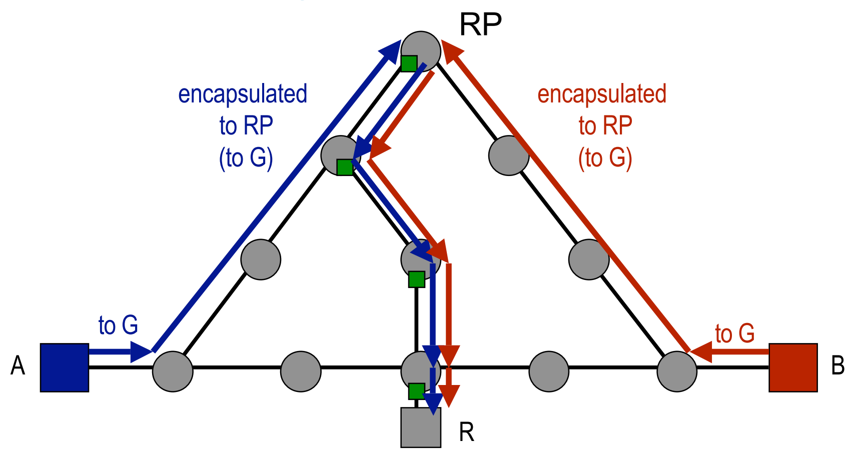
- Sources A and B send data
  - their local routers encapsulate it to the RP

PIM-SM: Building the Shared Tree



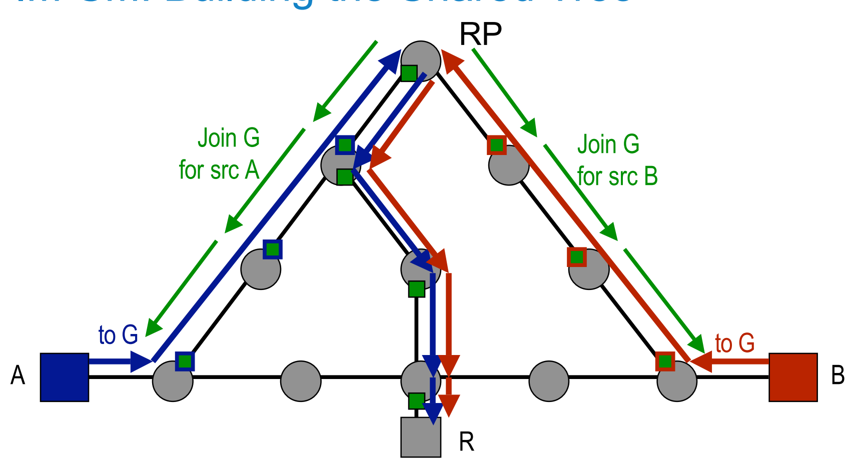
- Receiver R joins the group
  - It's local router sends a join message towards the RP

PIM-SM: Building the Shared Tree



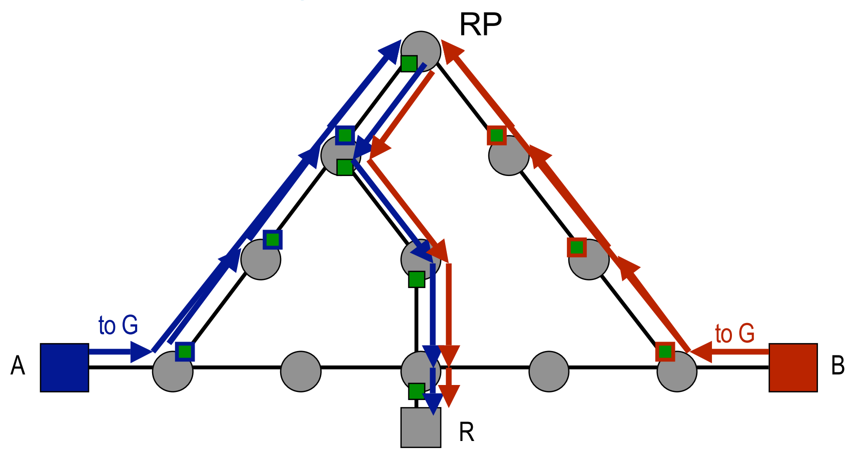
- Data starts to reach R from A and B

PIM-SM: Building the Shared Tree



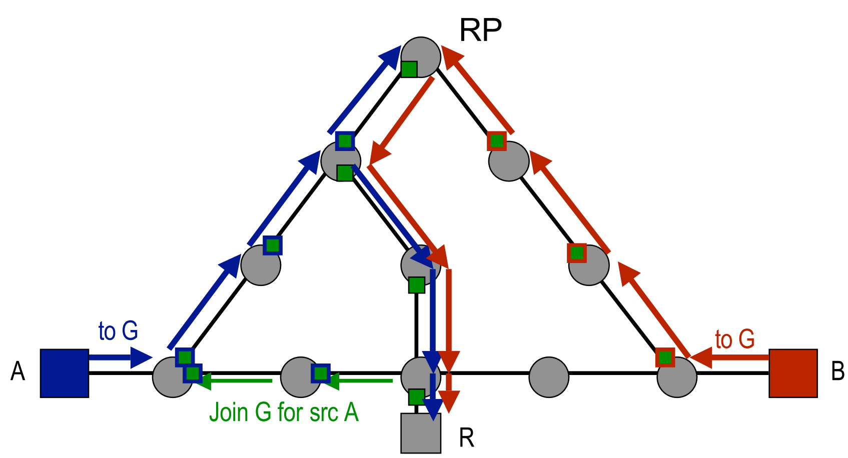
- The RP also sends join messages back to senders A & B
  - This will end the encapsulation.

PIM-SM: Building the Shared Tree



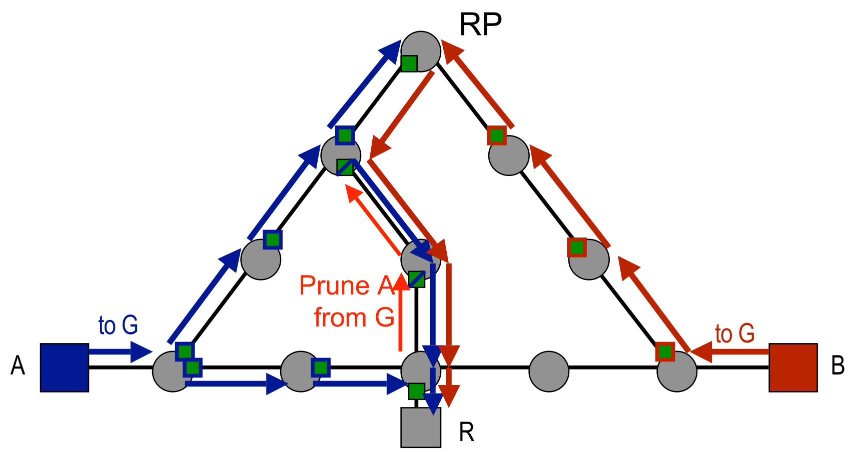
- The data now flows natively on the shared tree

PIM-SM: Shortest Path Tree



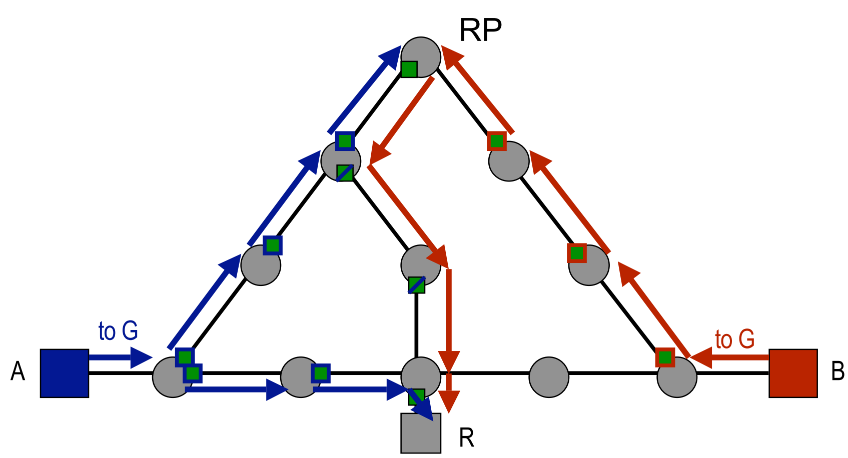
- A is sending high rate traffic
  - R's local router decides to switch to the shortest path tree
  - It sends an (source,group) join towards A.

PIM-SM: Shortest Path Tree



- Traffic from A now reaches R via the shared and shortest path trees.
  - This triggers an (S,G) prune to be send towards the RP.

PIM-SM: Shortest Path Tree



- R receives traffic from A on the shortest path tree.
- R receives traffic from B on the shared tree.

PIM-SM: Finding the RP

- The problem with PIM-SM is how the local routers discover which RP to use for which group.
  - Manual configuration.
  - Hashing the group address into a dynamically generated list of candidate RPs.
- Neither mechanism scales well, so PIM-SM is restricted to use as an intra-domain multicast routing protocol.
- Multicast Source Discovery Protocol (MSDP) is a hack to allow inter-domain shortest path trees to be built directly.
  - Also has significant scaling problems.

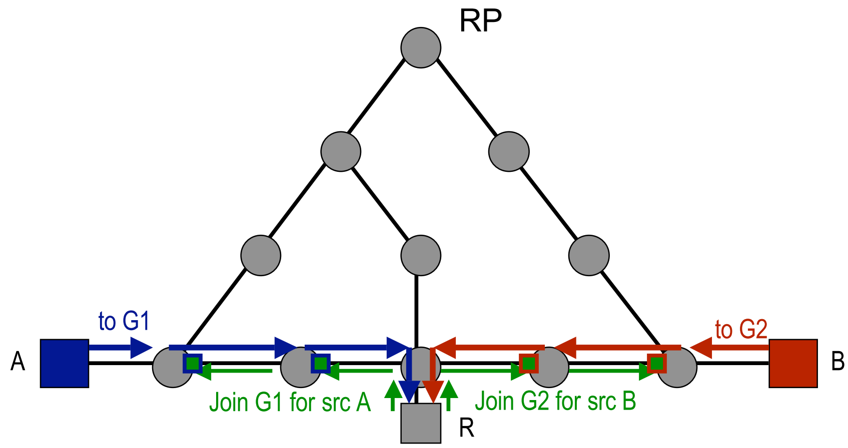
Hindsight

- After more than ten years of multicast research, where are we now?
- **We solved the wrong problem!**
  - For intra-domain, the any-source multicast model is fine.
  - For inter-domain, it's too difficult.
    - Difficult to debug
    - Difficult to charge for.
    - Vulnerable to denial-of-service attack.
    - Multicast address allocation is difficult.

The right problem

- Source-specific multicast (SSM)
  - Sources just send (as before)
  - Receivers directly join a source, specifying the group to join.
- Much simpler.
  - Less vulnerable to DoS.
  - Can re-use PIM-SM mechanisms.
  - Multicast addresses allocated by sender.
  - Needs IGMPv3 in hosts and routers.

PIM-SSM: Joining Directly



- R discovers the senders through some application-level mechanism
  - R signals it's local router to join A and B for groups G1 and G2.

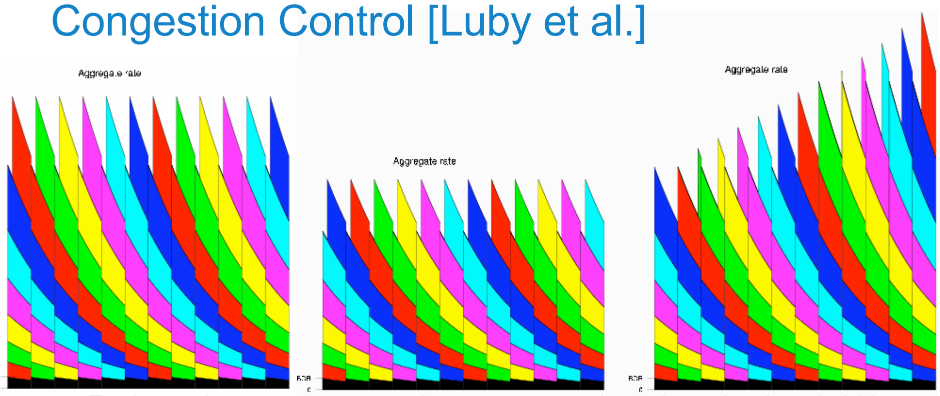
Reliable Multicast

- NACK based:
  - negative acknowledgments only
  - use where network is quite reliable
- Tree-based ACK (TRACK):
  - use intermediate nodes to handle ACKs and re-transmit.
  - use across heterogeneous networks with differing QoS
- Forward error correction:
  - FEC coding in content
  - suitable for real-time applications – re-transmit not needed
  - can use layered codecs

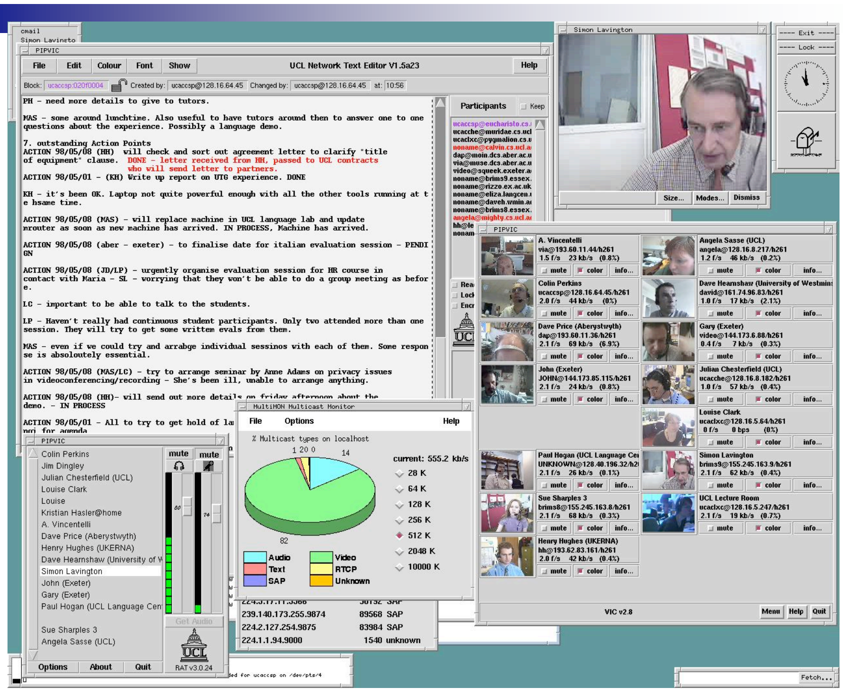
Multicast Congestion Control

- Equation-based congestion control
  - Receivers calculate the desired Tx rate, sender goes at rate of slowest receiver.
- TCP-emulation at the receiver [Rhee et al]
- PGMCC [Rizzo]
  - AIMD from a single representative
- Layered approaches
  - RLM [McCanne], RLC [Rizzo, Vicisano]
  - Wave & Eqn Based CC [Luby]

Wave & Equation-based Multicast Congestion Control [Luby et al.]



- Each multicast group transmits an exponentially decaying bandwidth.
- Receiver needs to keep joining the next group at at same place in each wave to keep constant bandwidth. Joining earlier increases bandwidth, later decreases bandwidth. Leaves group when it decays to zero, before it goes to max bandwidth again.



Multicast conferencing

- Floor control:
  - who speaks?
  - chairman control?
  - distributed control?
- Loose control:
  - one person speaks, grabs channel
- Strict control:
  - application specific, e.g.: lecture
- Resource reservation:
  - not widely supported
  - ~500Kb/s per conference (using video)
- Per-flow reservation:
  - audio only
  - video only
  - audio and video

Summary

- For really large scale distribution, multicast is a huge win.
  - Can duplicate to some extent at the application level.
- Lots of hard pieces:
  - Routing
  - Reliability
  - Congestion Control.
- After 15 years, multicast is finally maturing.
  - Does anyone care anymore?