IP Multicast:

Deployment and Applicability to “Push” Applications

Mark Handley,

USC/ISI

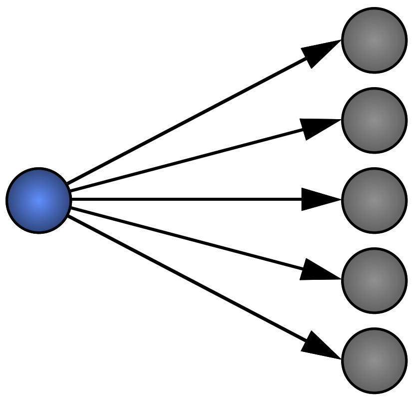
Outline

- What is IP multicast?
  - service model
  - multicast routing
- Deployment of multicast
  - status, problems and limitations
- Multicast and push applications
  - data streaming
  - bulk data distribution

What is IP multicast?

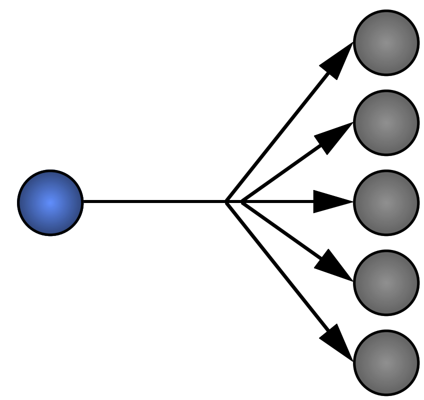
Multicast Requirements

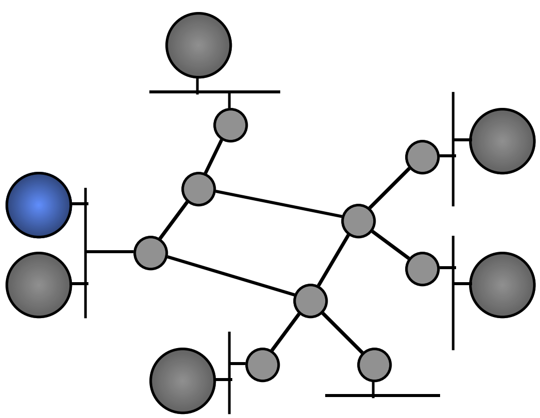
- Support one-to-many and many-to-many data distribution
- Scale to large groups
  - Data should only traverse each link once
- Robustness to link and node failures
  - No fate sharing
  - End systems don’t know topology



The user's view of the problem is
that they want to get this data
to all these receivers

The data sender's view is that they
only want to transmit the data once
and let replication happen downstream





But the network has an arbitrary
topology that may change over
time due to changing receivership
and failures.

IP Multicast

Dynamically Constructs efficient, shortest path distribution trees from senders to receivers

IP Multicast Service Model:

- ***Receivers announce interest***
- ***Senders just send***
- ***Routers conspire to deliver sender's data to all interested receivers***

Class D addresses

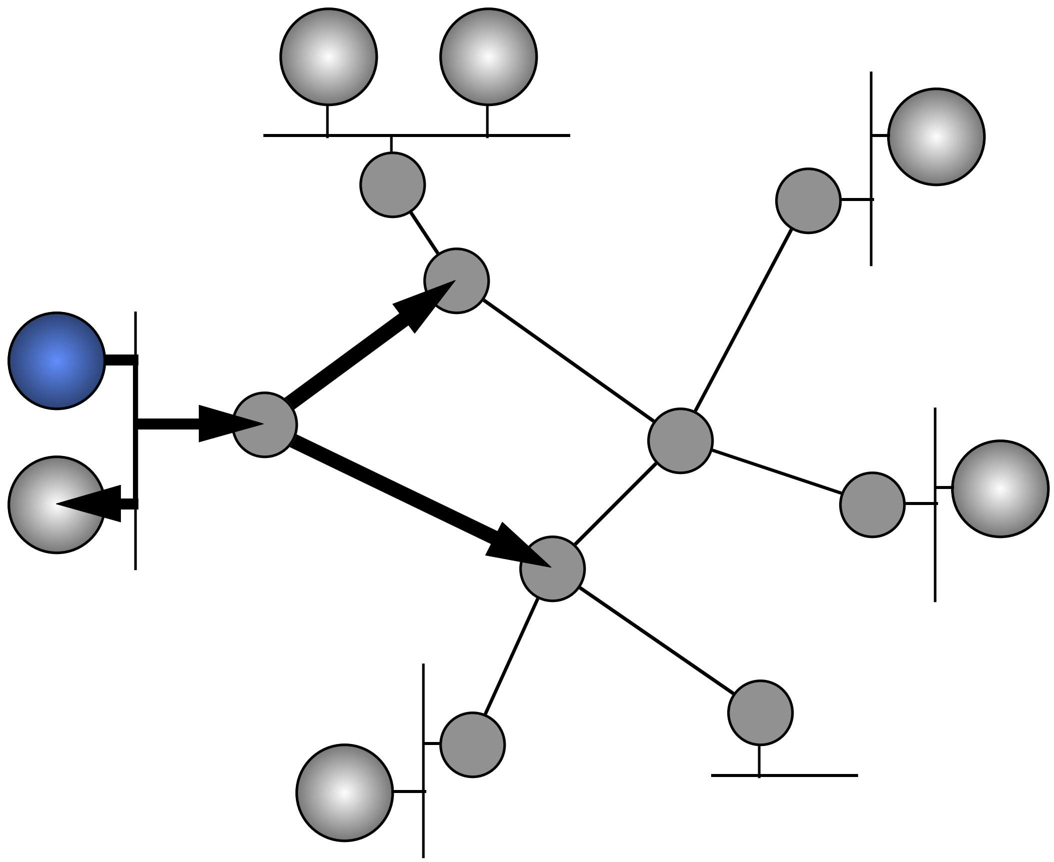
- **Traditionally, IP addresses were divided into classes:**
  - **Class A for companies with millions of hosts**
  - **Class B for companies with thousands of hosts**
  - **Class C for companies with hundreds of hosts.**
- **Class D is multicast group addresses.**
  - **A multicast sender just sends 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.**

Multicast Routing

- Multicast routing performs the task of setting up the distribution trees from senders to receivers.
- DVMRP was the first multicast routing protocol, and is still predominant. But it suffers from scaling problems, and is slowly being replaced.
- PIM is an alternative that is now widely deployed (particularly in Cisco routers). It scales somewhat better.
- BGMP is just being developed to add hierarchy to multicast routing.

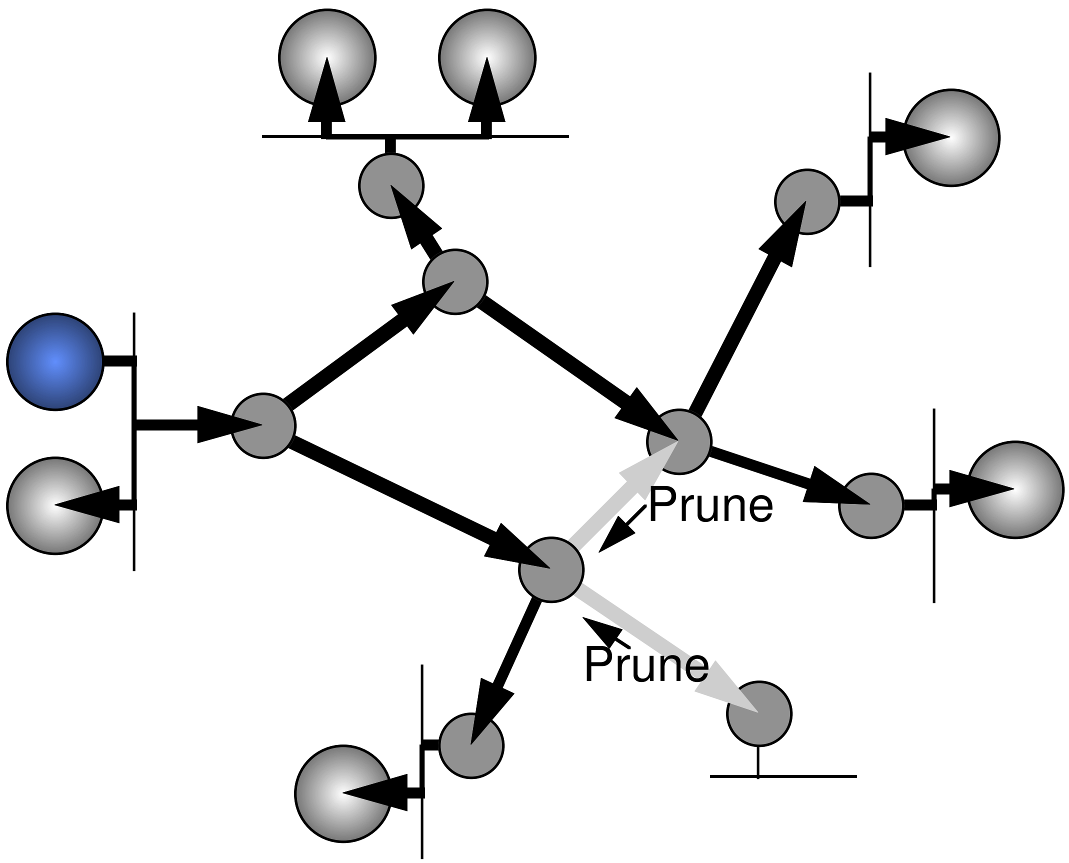
DVMRP

- **Basic Model is "Flood and Prune"**



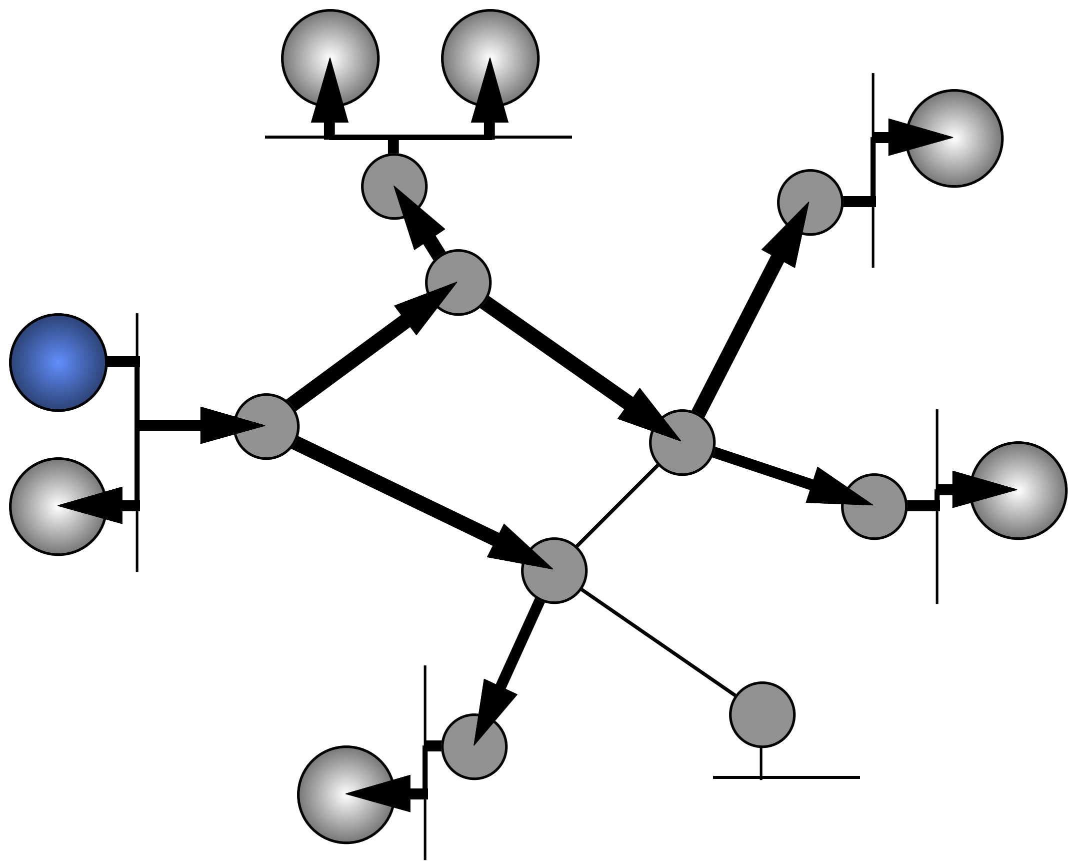
Flood and Prune

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



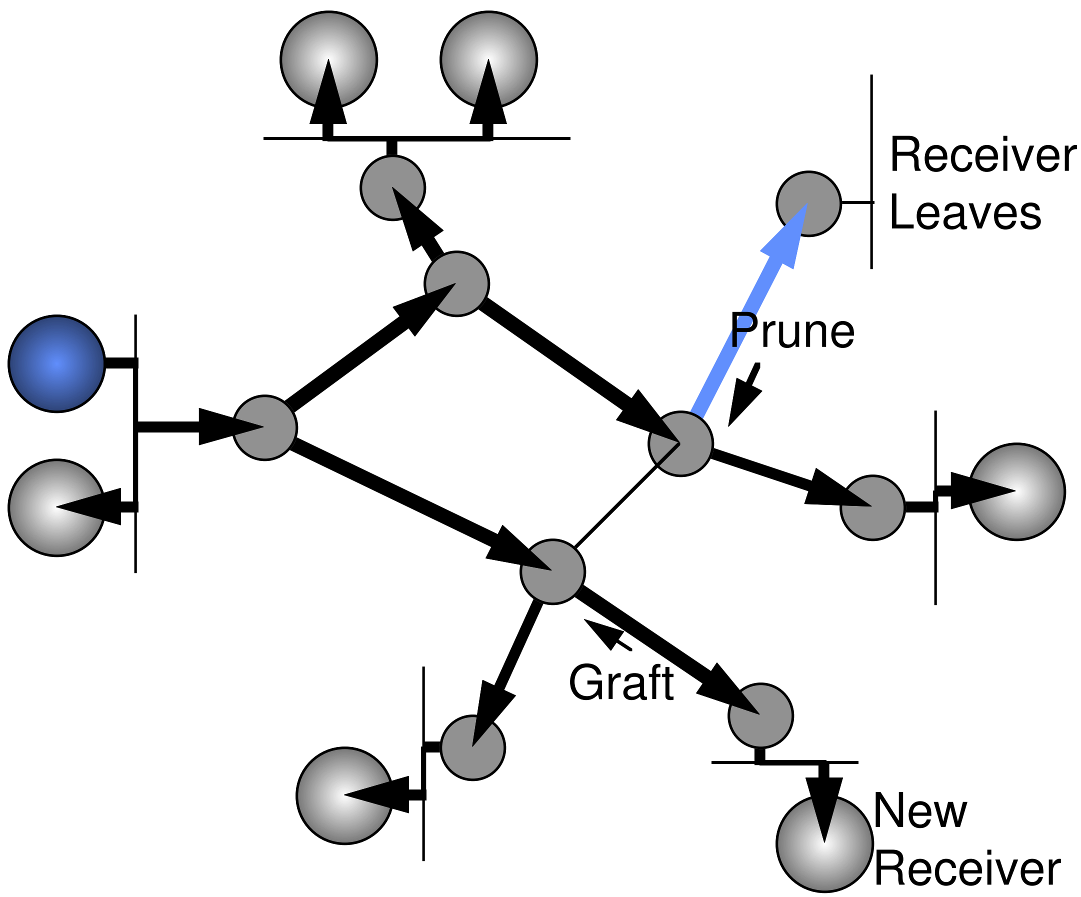
Shortest Path Tree

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



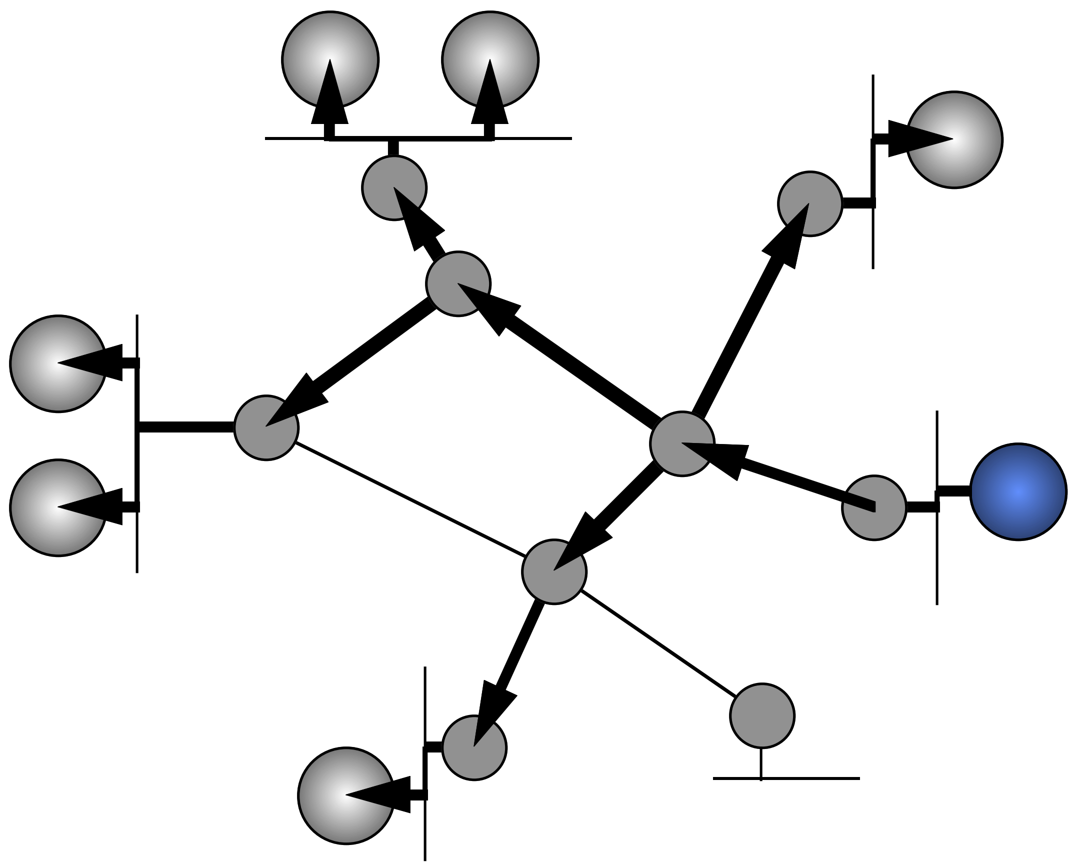
Local Tree Maintenance

Changes in membership are handled efficiently and locally via grafting and pruning:



A Shortest Path Tree per Sender

Distribution trees are per-{sender,group}, and are triggered by data packets:

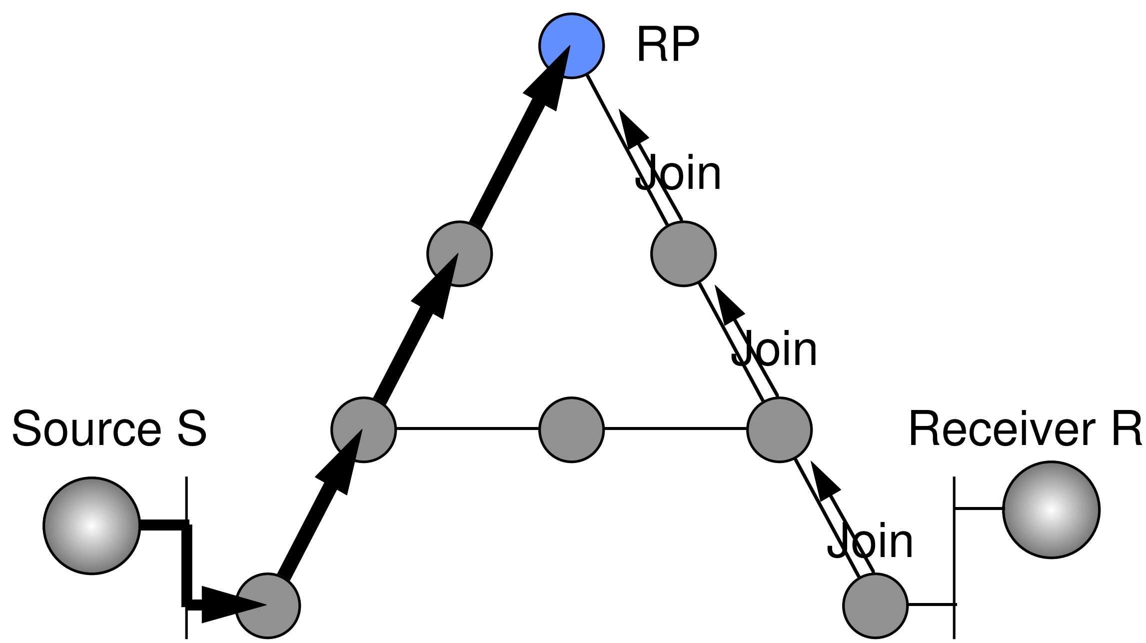


SM-PIM

- **Instead of flooding and pruning to directly build a shortest path tree, Sparse-mode PIM 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.**

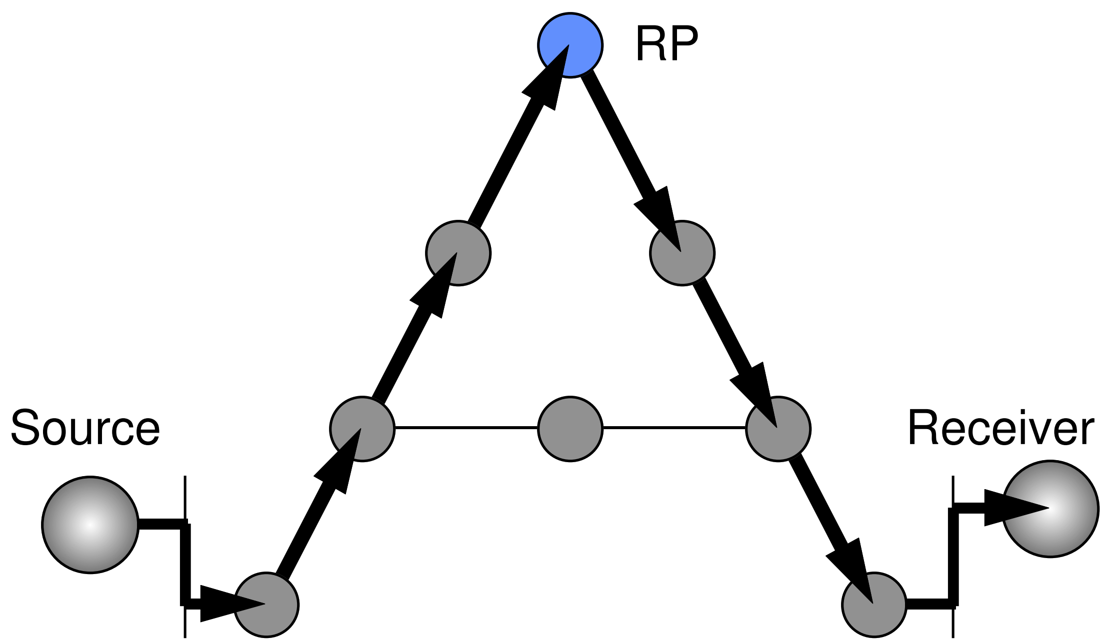
SM-PIM: Building a Shared Tree

- Source S sends traffic to group G
- S's local router forwards group G traffic to the Rendezvous Point (RP) for group G
- R's local router sends a join message for G to RP



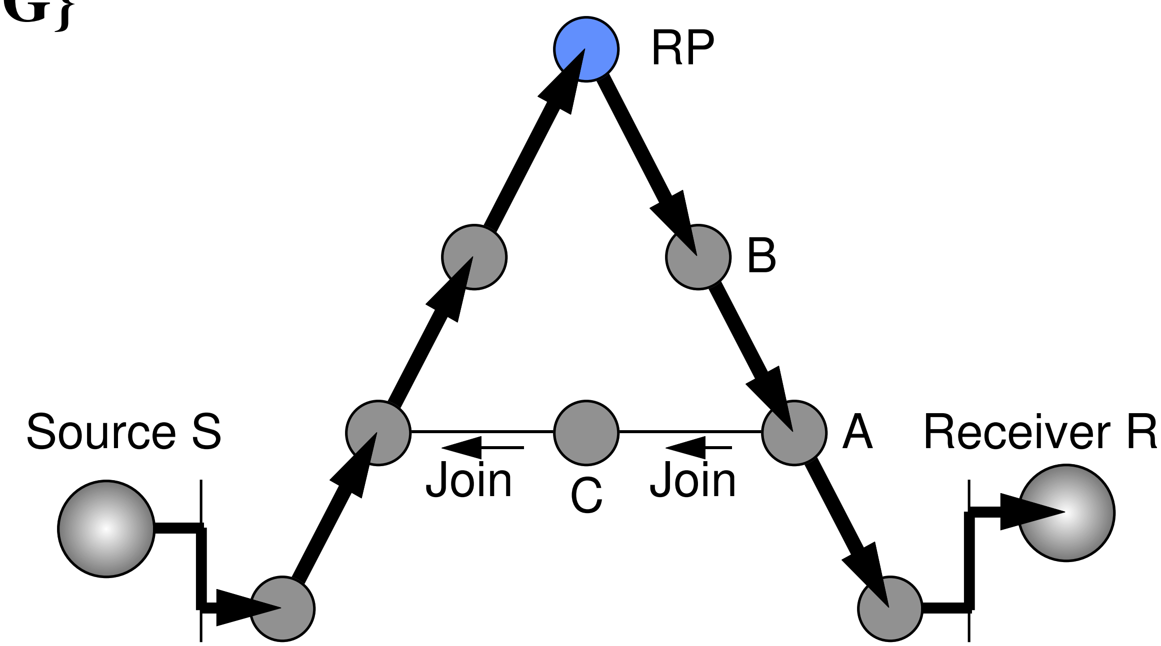
SM-PIM Shared Tree

- **Initially data flows over RP tree**



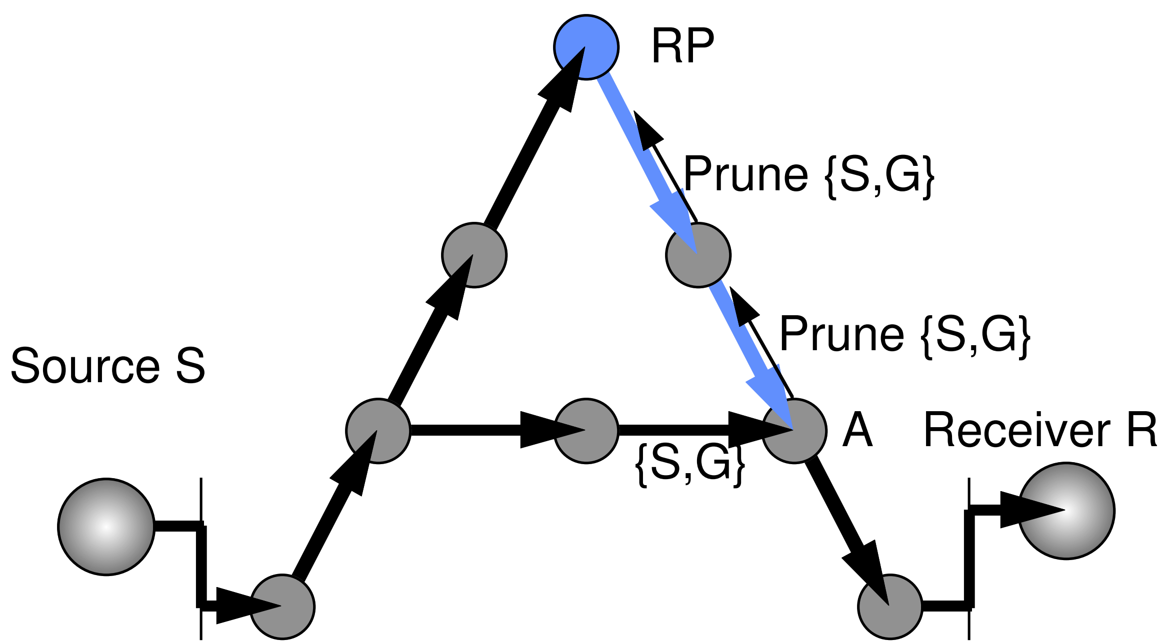
Transfer to SP Tree

- Router A sees traffic from S arriving over its B interface, when its unicast route to S is over its C interface.
- If traffic warrants it, A sends a join towards S for {S,G}



Transfer to SP Tree (cont)

- When the Join towards S is successful, group G traffic from S starts to flow along the shortest path tree. A now receives the traffic twice.
- A then sends a prune for {S,G} towards the RP for group G, and the transfer to SP tree is complete.

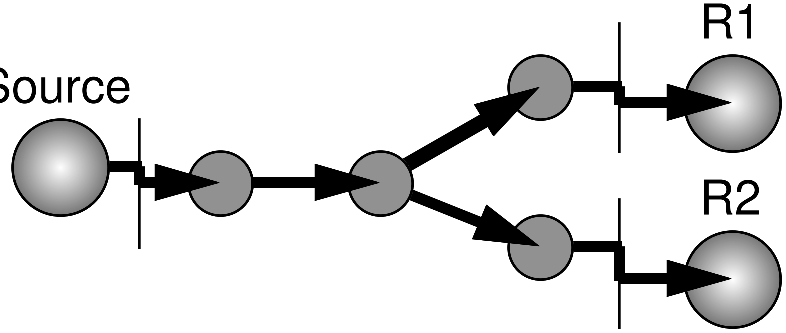


Current Status

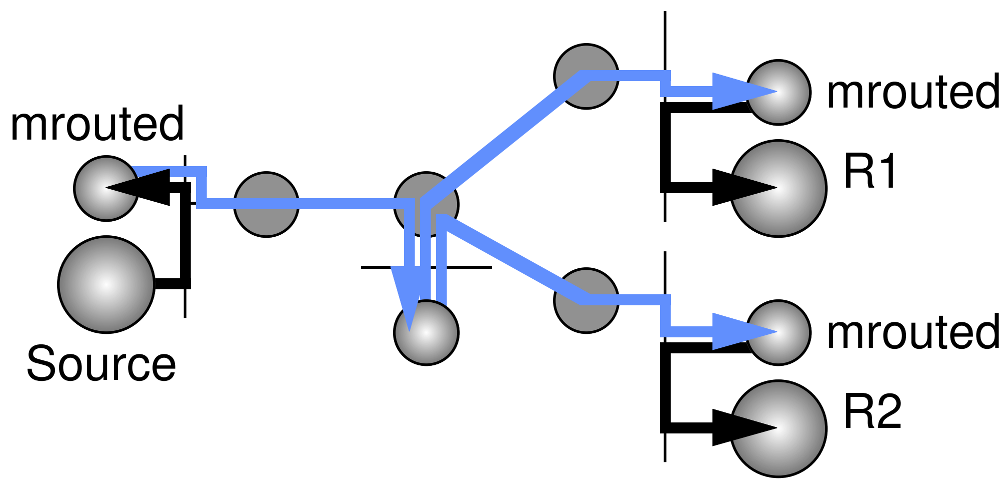
- DVMRP and PIM are what is out there currently.
- Approximately 8000 subnetworks connected to the Mbone.
- Sites are typically connected by “tunnelling” multicast traffic across parts of the internet that don’t yet support multicast, but this is starting to change as native router code supporting multicast is deployed.

Multicast Tunnels

Native Multicast in Routers:

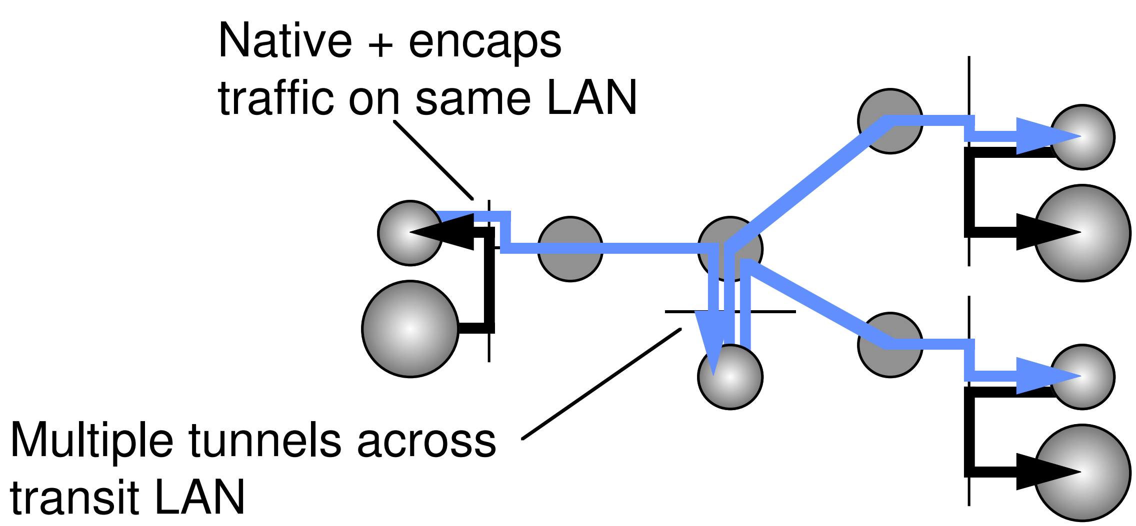


Multicast Tunnelling in workstation mrouters:



Multicast Tunnelling (cont)

- Multicast can be tunnelled across multicast unaware links, but this can limit performance.
- Workstations are not ideal routers
- At best, multicast traverses your LAN twice
- At worst, many tunnels traverse your LAN
- Tunnels require manual configuration



Multicast Myth:
The Mbone can only support 512Kbps total

When the Mbone was small, pruning didn't work, and most multicast was tunnelled, typically a host on an ethernet would support up to 6 tunnels.

- **Typical realistic 10M ethernet bandwidth = 4Mbps**
- **512Kbps * (1 incoming tunnel, 5 outgoing tunnels, and 1 local fanout) = 3.5Mbps**
- **Hence a recommended limit of 512Kbps.**

Also this was 1/3 of a T1, which was the predominant backbone link back in 1992.

This limit no longer applies!!!

Multicast Problems (1):

Scaling problems

- **8000 subnets is a lot for a distance vector multicast routing protocol! This leads to instability...**
- **The MBone is in transition from being an experiment to being a ubiquitous service as DVMRP improves (route aggregation), PIM is used more, and as we start to introduce hierarchy in the routing protocols.**
- **Watch out for BGMP - an upper level multicast routing protocol leveraging BGP unicast routing mechanisms.**

Multicast Problems (2):
Lack of support for dialup users

- **This is no longer the case - several manufacturers now support multicast for dialup.**
- **Getting ISPs to enable it a potential problem.**

Multicast Problems (3):
ISP's are confused about the business case

- **actually multicast will mostly help them**
- **without users there is no business case**
- **if there's user demand, ISP's MUST provide multicast or customers go elsewhere.**
- **probably intranets are the first place multicast will become ubiquitous.**

An Example of Multicast for "PUSH"

- **Live multicast audio and video at 192Kbps has been transmitted to around 1000 simultaneous users.**
- **1000 is a small user base, but to do this using "conventional" push technology would be extremely difficult.**
  - **a single server would need to source 192Mbps**
  - **we only needed a SparcStation 10**
- **There is no technical reason why this 1000 couldn't have been 1,000,000, with no additional investment by the senders or receivers.**

Multicast Problems (4):
not everyone gets acceptable quality

- **For audio and video there are techniques now being deployed that use layered codecs to let the receivers decide appropriate quality, and add low bandwidth redundancy to provide some measure of protection against packet loss.**

Beyond audio and video.

- **Whilst multicast is *required* for large-scale distribution of audio and video, probably various forms of data distribution will actually have more impact in the long run. For live data we have:**
  - **shared whiteboards**
  - **a shared text editor**
  - **voting and chat tools**
- **For stored data, applications have been written that use multicast to distribute:**
  - **still images (particularly weather satellite images)**
  - **usenet news**
  - **large email lists**
  - **earthquake data**

Multicast Problems (5):
Multicast doesn't reach all our customers

- **Clearly it never will unless there's demand.**
  - **or when IPv6 takes over from IPv4.**
- **Multicast usenet news is a good example of one approach:**
  - **Multicast is their primary means of distribution, but their multicast transport protocol is unreliable.**
  - **Standard unicast NNTP provides the reliability - messages that were distributed by multicast don't need to be sent via unicast, but those that were dropped are filled in by the standard protocol.**
  - **NNTP lends itself quite well to this optimisation - not all protocols are so simple.**

Push-pull usage:
A simple design for an inter-cache protocol

- Every "site" in an ISP (POP,large custoner, whatever) runs a conventional web cache.
- The web caches talk to each other using multicast. Each cache multicasts at a low rate indexes of all the data in the cache together with checksum, and expiry and last modified timestamps. This serves to inform the other servers about the cache item to update timeouts and to revoke modified data in other caches.
- When a client requests an item not stored in the local cache, if it is in any other cache, it can be retreived directly from the best connected cache using standard HTTP 1.1.
- The distributed caches now act as one cache without needing a fragile configured cache hierarchy.
- 10 caches communicating indexes of 100,000 compressed URLs every 8 hours requires a single multicast channel of only 11Kbps.

Multicast for server discovery

Multicast is great for application-level self-organising topologies because senders and receivers are "joined together" by the routing protocol. They don't need a-priori knowledge of each other.

Typically these can be supplimented by less flexible configured hierarchies when multicast is unavailable.

Multicast Problem (6)
Reliability

This isn't really a problem for many applications.

- **Receiver-based reliability scales well.**
  - **no way to tell that all receivers have all the data though, so application architectures must be designed with this in mind.**
- **Sender-based reliability does not scale**
  - **although it probably scales better than unicast.**

The best mechanisms appear to be a hybrid of:

- **Scalable Reliable Multicast (SRM - see Proc ACM Sigcomm 95) - receiver based retransmission.**
- **Packet-level FEC (see Proc ACM Sigcomm 97) to cope with uncorrelated packet loss.**

Multicast Problem (7)
Congestion Control

- **This is the hardest technical problem.**
  - **TCP congestion control does not extend trivially to multicast.**
  - **To avoid response implosions, feedback needs to be delayed.**
  - **Congestion control mechanisms acting on timescales slower than TCP and competing with TCP are not well understood at this time, but there is significant active research in this area.**
- **Note: it is not appropriate to perform sender-based congestion control for all flows:**
  - **Eg: stock price feed has a fixed data rate - if the network is insufficient for receivers to keep up, they should leave as no network friendly protocol (including TCP) is going to be able to do anything useful!**
- **However, for elastic flows where the appropriate data rate is "as fast as possible", congestion control is a *requirement.***

Multicast Problem (8)
Multicast Address Allocation

- This has been a research issue for several years, but in the context of BGMP an architecture is now starting to emerge.
- Expect an IETF WG to be formed at the next IETF to standardise multicast address allocation mechanisms.
- The first part to be standardised will be how a client requests an address, so application writers have an API to program to.
  - This will be based on DHCP.

Summary

Problems:	Solutions
Scalable multicast routing	PIM (L1) + BGMP (L2)
Dialup support	Starting to be available
ISP business case	??? probably demand driven
Acceptable quality .	Solutions depend upon the application but many do exist
Reaching enough users	Hybrid applications
Reliability	SRM + packet-level FEC
Congestion Control .	Depends on application - solutions are just starting to emerge.
Address Allocation	Standardisation has now started

Internet Research Task Force (IRTF) *Research* Group on Reliable Multicast

- **A new open IRTF RG aimed at bringing together research in this area so that the IETF can then start the standardisation process.**
  - **co-chairs: Allison Mankin, Mark Handley, Ran Atkinson**
- **Mailing list:**
  - **rm@irtf.cs.berkeley.edu**
  - **(to subscribe, rm-request@irtf.cs.berkeley.edu)**
- **Meetings twice a year - next meeting is after Sigcomm 97 (19th and 20th Sept) in Cannes.**
- **Expect some standardisation work to begin in Q1 1998...**