Living Streaming and Overlay Multicast

David Harrison
harrisod@eecs.berkeley.edu

EE290T Spring 2004

Video and Image Processing Lab, EECS
University of California, Berkeley
http://www-video.eecs.berkeley.edu

Some slides and diagrams lifted from I. Stoica, A. Parekh, and P. Mehra

Outline

Media Streaming Problem

Background

Application-Layer Multicast

End-system multicast: Narada

BREAK

Scalability via Distributed Hash Tables

DHT-based multicast: Splitstream

Infrastructure-based Multicast: Scattercast (if time)

Media Streaming Problem

- Stream live audio/video to many, large audiences.
- Streaming audio:
  - Top 5 online broadcasters:
    MusicMatch, AOL Radio, Yahoo launchcast, Live365, Virgin Radio had est. tot 207000 average simultaneous listeners in 2/04 [Arbitron]
  - Virgin Radio had 4200 average numbers listeners in 2/04 [Arbitron]
  - Live 365 claims 10,000's simultaneous stations.
- Video streaming
  - Rush Limbaugh's Dittocam (hundreds? thousands? simultaneous viewers).

Outline

Media Streaming Problem

Background

Application-Layer Multicast

End-system multicast: Narada

BREAK

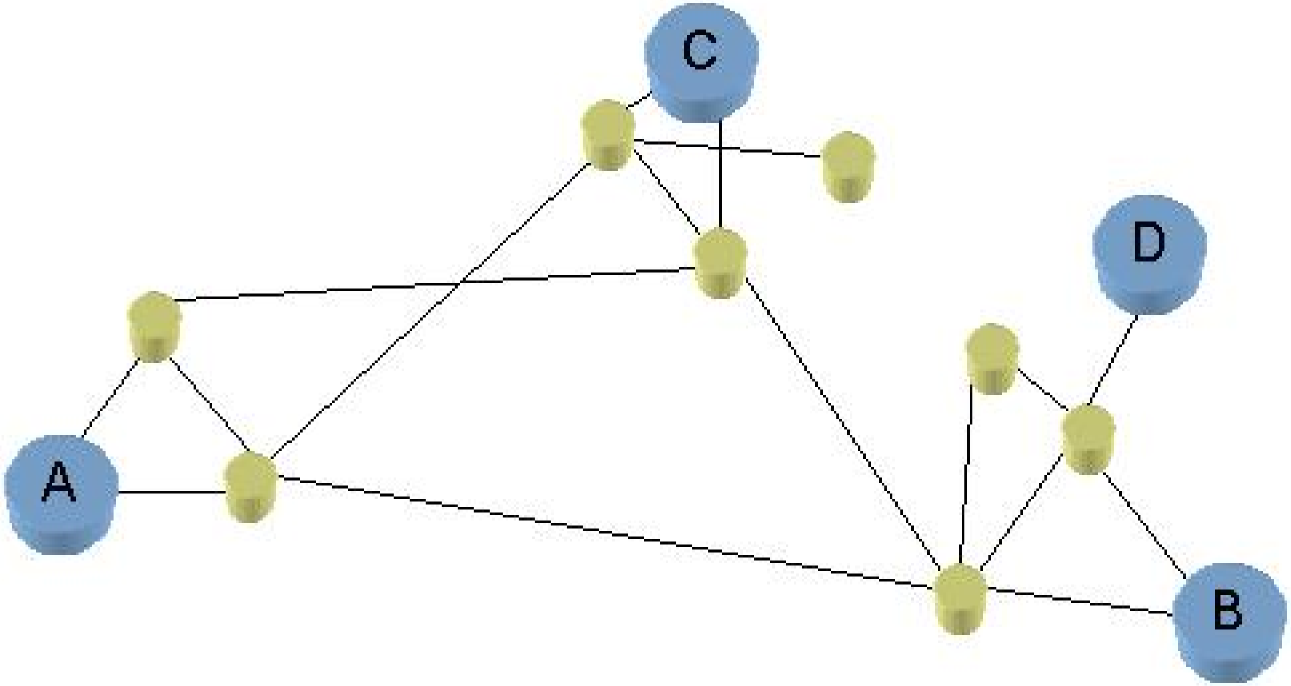
Scalability via Distributed Hash Tables

DHT-based multicast: Splitstream

Infrastructure-based Multicast: Scattercast (if time)

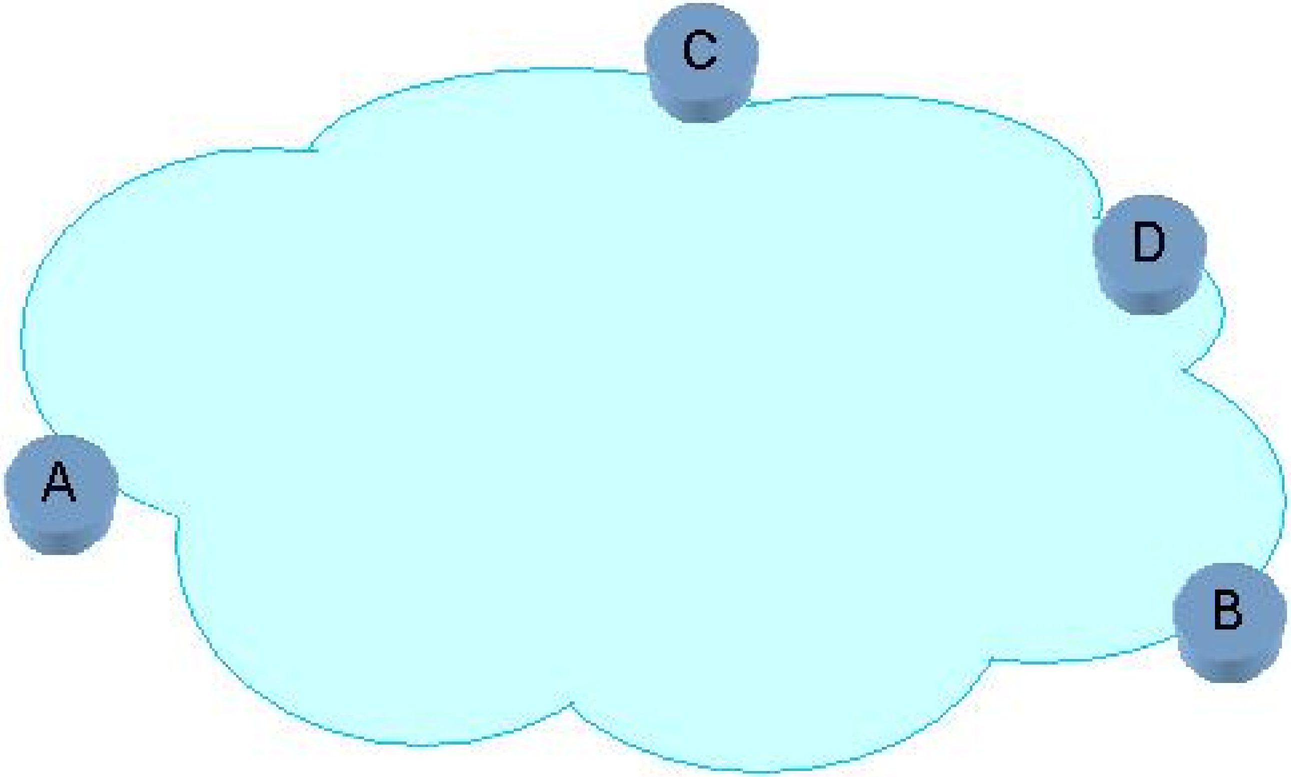
What is Overlay Multicast?

- Subset of IP nodes engage in multicast.
- Other nodes are oblivious. Just see unicast traffic.



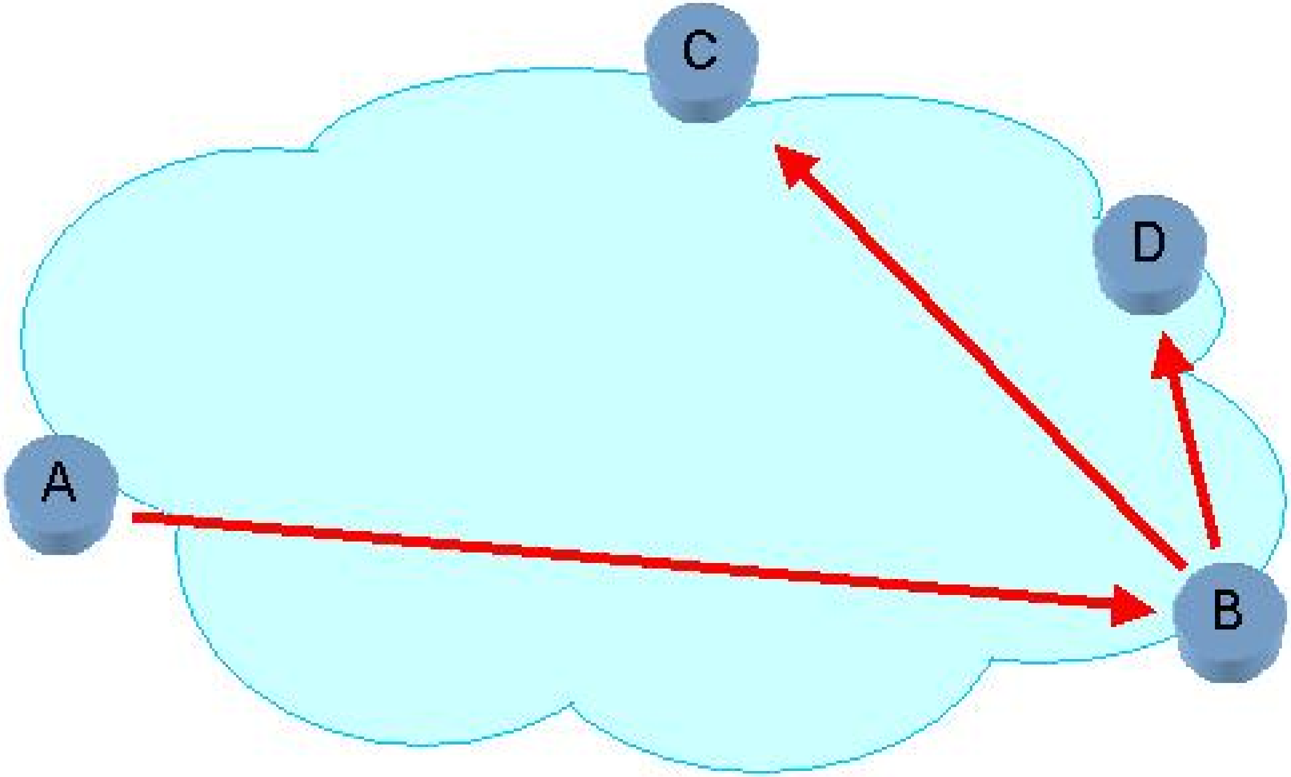
What is Overlay Multicast?

- Subset of IP nodes engage in multicast.
- Other nodes are oblivious. Just see unicast traffic.



What is Overlay Multicast?

- Subset of IP nodes engage in multicast.
- Other nodes are oblivious. Just see unicast traffic.



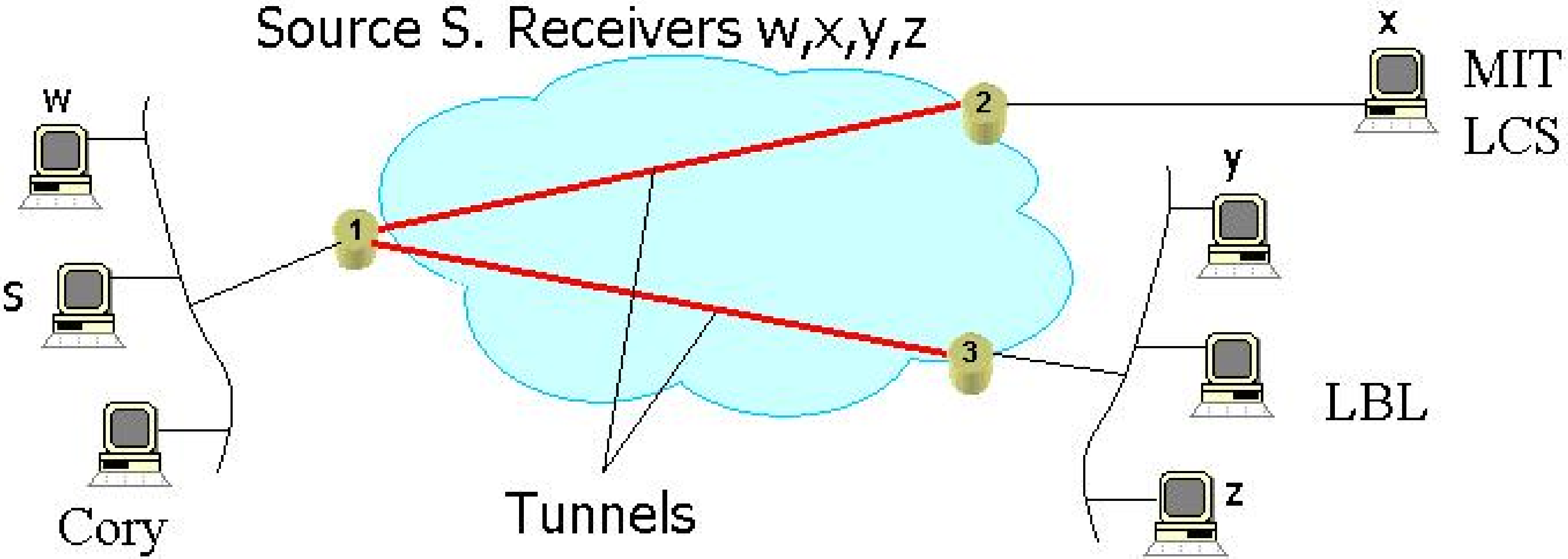
What’s wrong with IP Multicast?

- Not deployed. But why?
- Routing table explosion.
  - Routers maintain per-group routing table entry.
  - Difficult to aggregate multicast addresses.
- Reliability and congestion control are difficult.
  - Potentially every receiver has a different rate.
  - NAK implosion.
- Christophe Diot adds:
  - Multicast address allocation
  - Lack of support for network management
  - Group management (receiver/sender authorization, group creation).
- Difficult to Monitor Performance

Why not BIG servers?

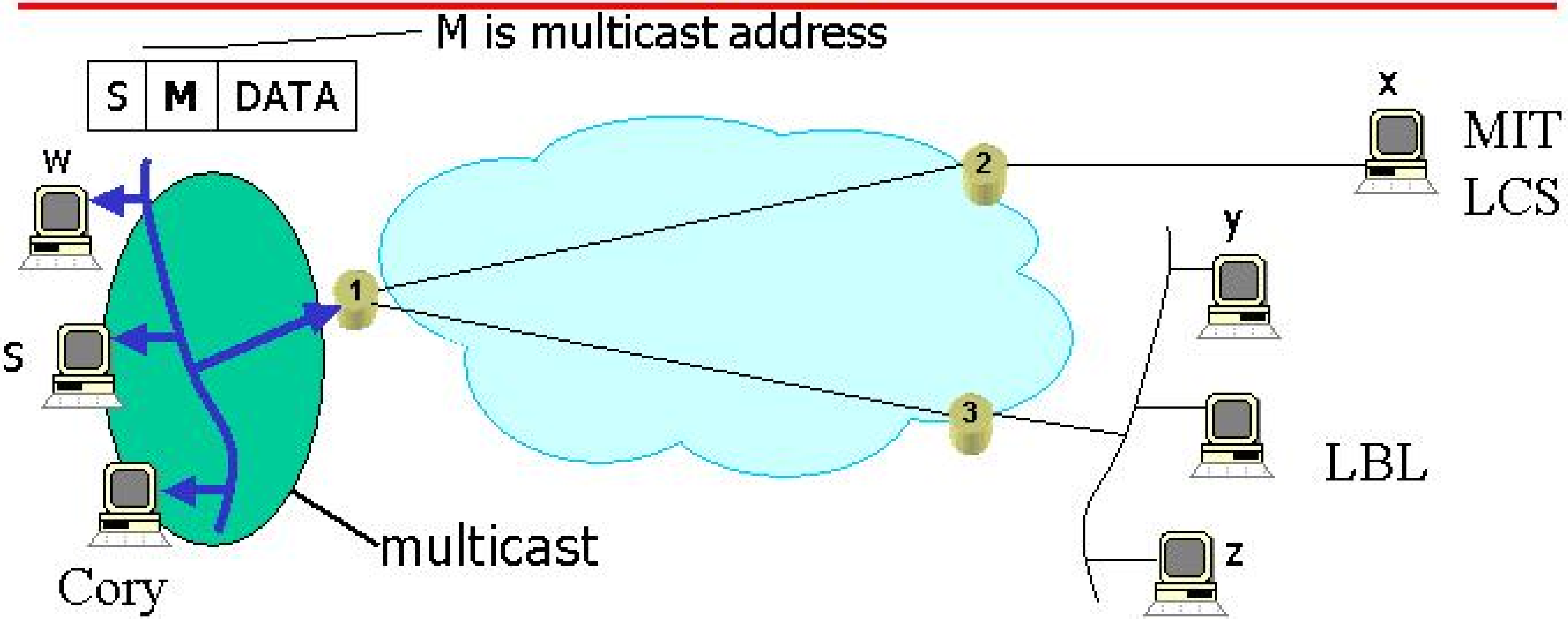
- Use TCP or UDP+TFRC from server to each receiver.
- Current method for streaming video.
- Server load, state, bandwidth, cost grows *linearly* with number of receivers n.
- Inefficient.
  - Same data transferred O(n) times over access link
- Server farms scale to larger audiences but still **O(n).**

Partial solution: IP Tunneling and the MBONE



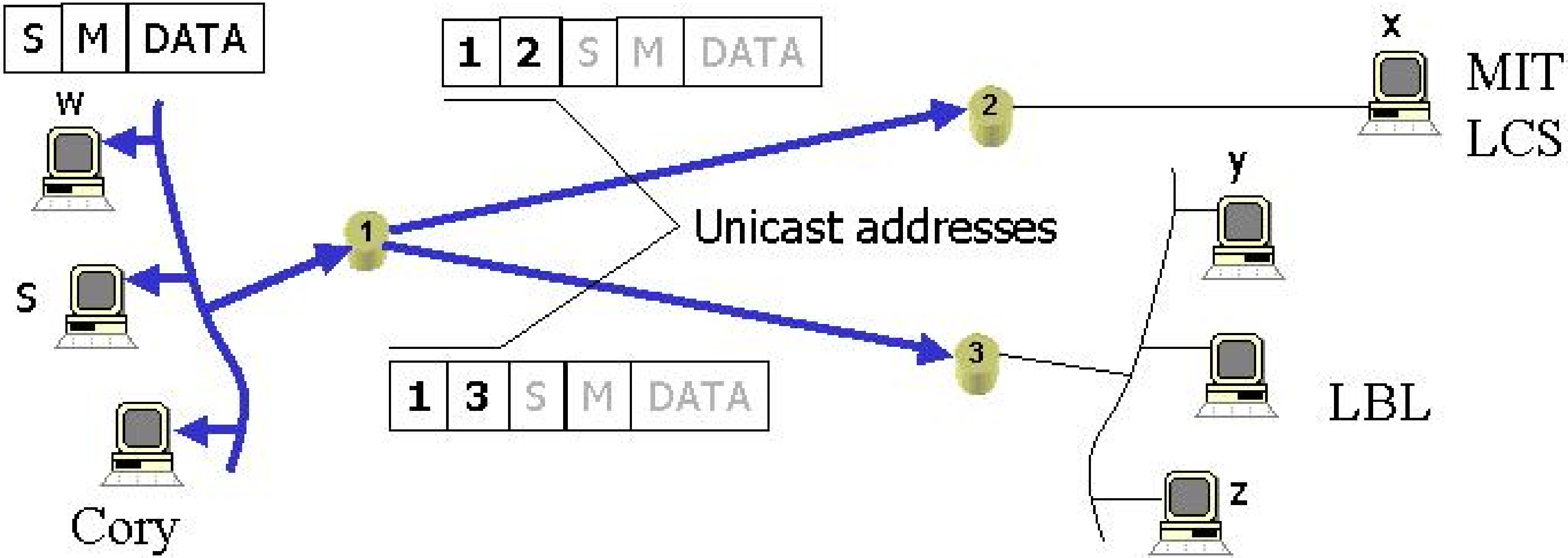
- Connect multicast-enabled networks (campus, LANs) via IP tunnels.
- First example of overlay multicast. *Tunnels overlay core.*
- Solves routing table explosion in core.

Partial solution: IP Tunneling and the MBONE



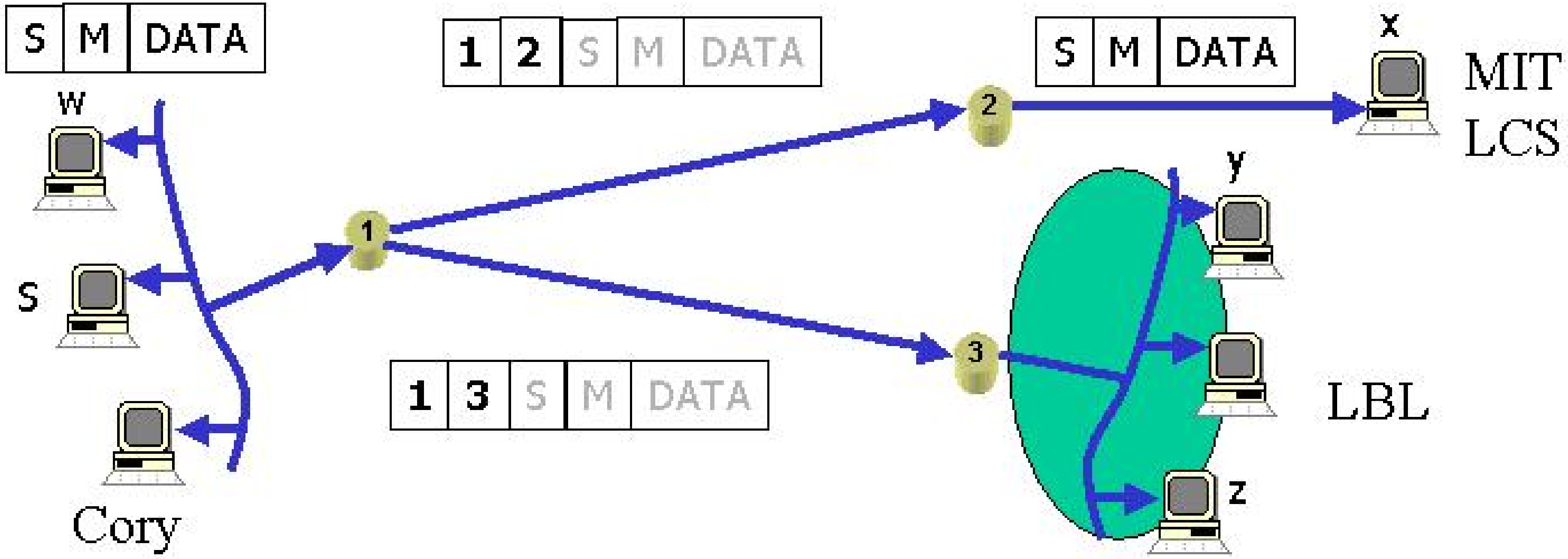
- Connect multicast-enabled networks (campus, LANs) via IP tunnels.
- First example of overlay multicast. *Tunnels overlay core.*
- Solves routing table explosion in core.

Partial solution: IP Tunneling and the MBONE



- Connect multicast-enabled networks (campus, LANs) via IP tunnels.
- First example of overlay multicast. *Tunnels overlay core.*
- Solves routing table explosion in core.

Partial solution: IP Tunneling and the MBONE



- Connect multicast-enabled networks (campus, LANs) via IP tunnels.
- MBONE is in current Internet as a working testbed.
- First example of overlay multicast.
- Solves routing table explosion in core.

Why not IP tunneling?

- Perfect when small number of sites with dense viewership within each site.
- Must configure each tunnel endpoint.
- Tunnel endpoints must maintain state for every tunnel terminating at a tunnel endpoint.
- Does not scale when many sites.
  - Consider when # sites is O(n),
  - Tunnel endpoints must maintain O(n) routing state.

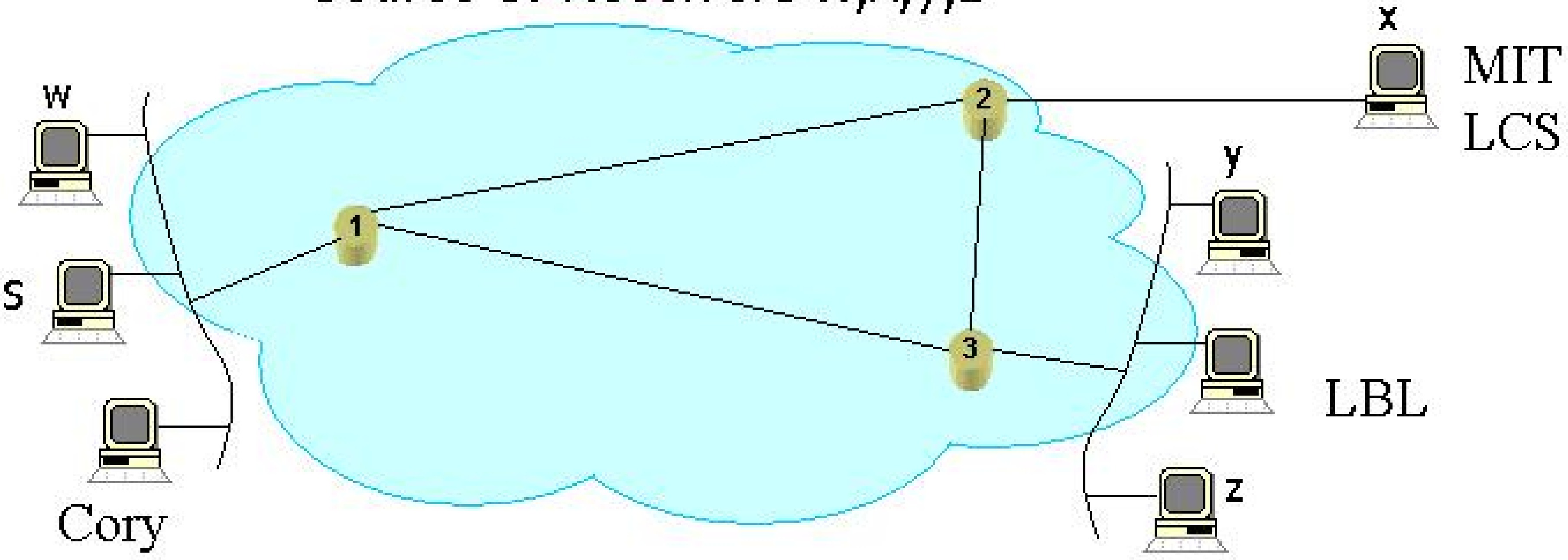
Outline

- Media Streaming Problem
- Background
- **Application-Layer Multicast**
- End-system multicast: Narada
- BREAK
- Scalability via Distributed Hash Tables
- DHT-based multicast: Splitstream
- Infrastructure-based Multicast: Scattercast

What is Application-Layer Multicast (ALM)?

- Move IP Multicast into Application Layer.
- Ex: End-system Multicast (Peer-to-peer)

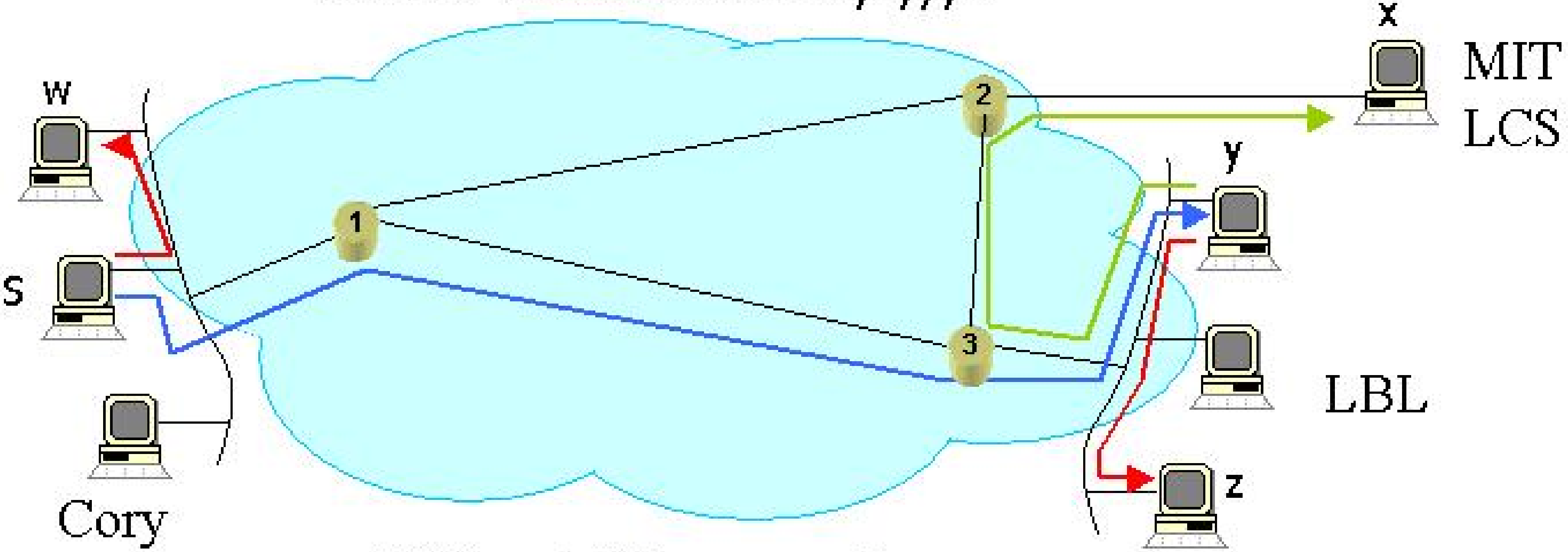
Source S. Receivers w,x,y,z



What is Application-Layer Multicast (ALM)?

- Move IP Multicast into Application Layer.
- Ex: End-system Multicast (ESM), a.k.a., Peer-to-peer

Source S. Receivers w,x,y,z



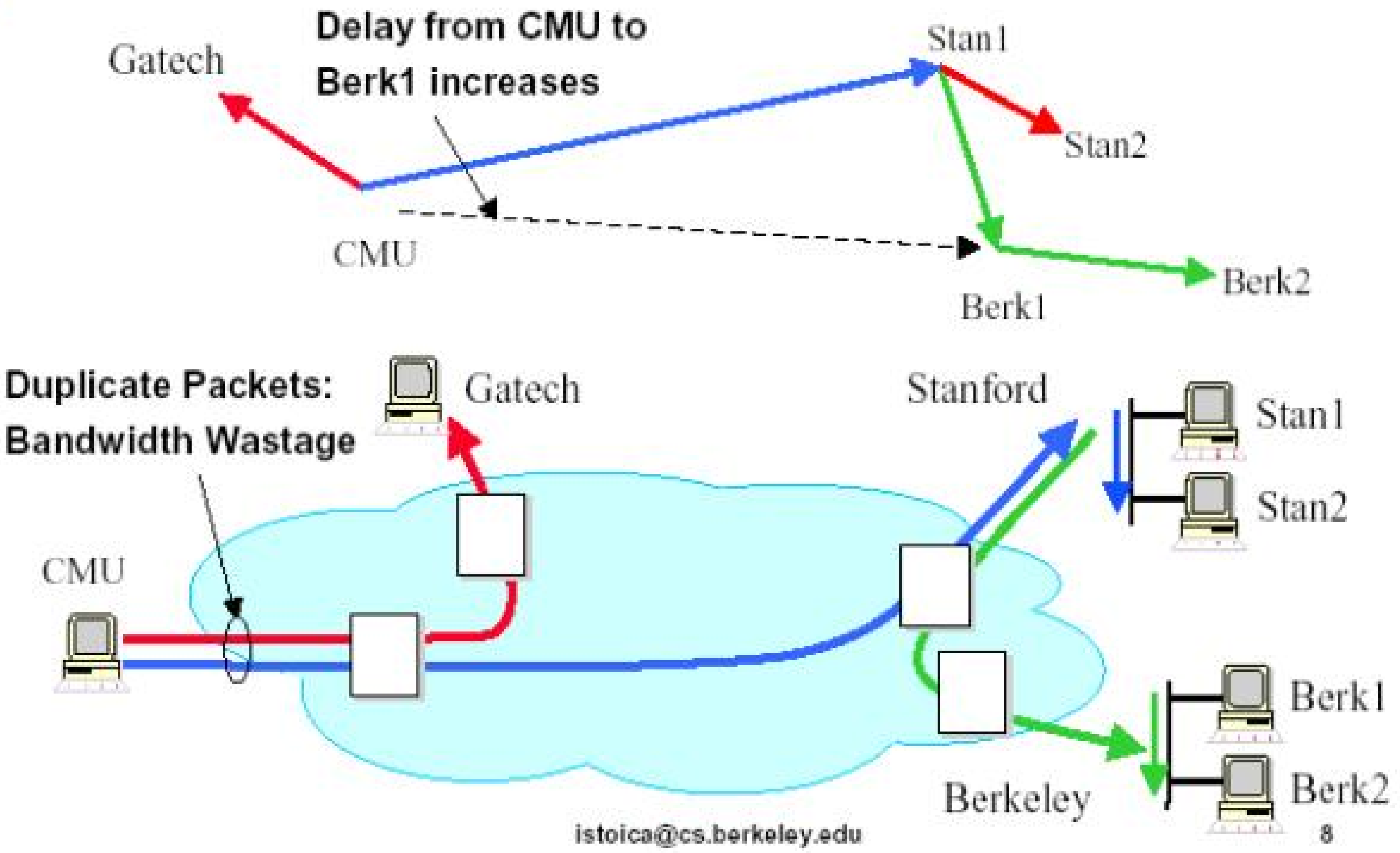
TCP or UDP connections

Why End-System Multicast?

- Scalability
  - Routers maintain no per-group state.
  - End-systems do, but they participate in few groups.
- Easier to deploy
- Potentially simplifies support for higher level functionality
  - Leverage computation and storage of end systems.
  - For example, for buffering packets, transcoding, ACK aggregation
  - Leverage solutions for unicast congestion control and reliability
    - Trivial if use TCP.
    - Or UDP+TFRC
  - Can afford to implement complex security measures.

Adapted from Ion Stoica

Performance Concerns



Delay from CMU to Berk1 increases

Duplicate Packets: Bandwidth Wastage

Adapted from Ion Stoica

Other Challenges facing End-System Multicast

- Small access bandwidth
  - Asymmetric Bandwidth (more down than up)
- End-systems often unwilling to forward
- End-systems typically less trustworthy than router
  - Substitute/Garbage content

- (We won't discuss these further)

Outline

Media Streaming Problem

Background

Application-Layer Multicast

End-system multicast: Narada

BREAK

Scalability via Distributed Hash Tables

DHT-based multicast: Splitstream

Infrastructure-based Multicast: Scattercast (if time)

NARADA: Example End-System Multicast

- NARADA [Y. Chu et al JSAC Oct 2002]
- A distributed protocol for constructing efficient overlay
- Self-organizes

- Caveat: assume apps with small and sparse group
  - Around tens to hundreds of members

Why is self-organization hard?

- Fully-distributed
  - Implies no central knowledge
- Dynamic changes in group membership
  - Members may join and leave dynamically
  - Members may die
- Limited knowledge of network conditions
  - Members do not know delay to each other when they join
  - Members probe each other to learn network related information
  - Overlay must self-improve as more information available
- Dynamic changes in network conditions
  - Delay between members may vary over time due to congestion

Adapted from Ion Stoica

NARADA self-organizes in 2 steps

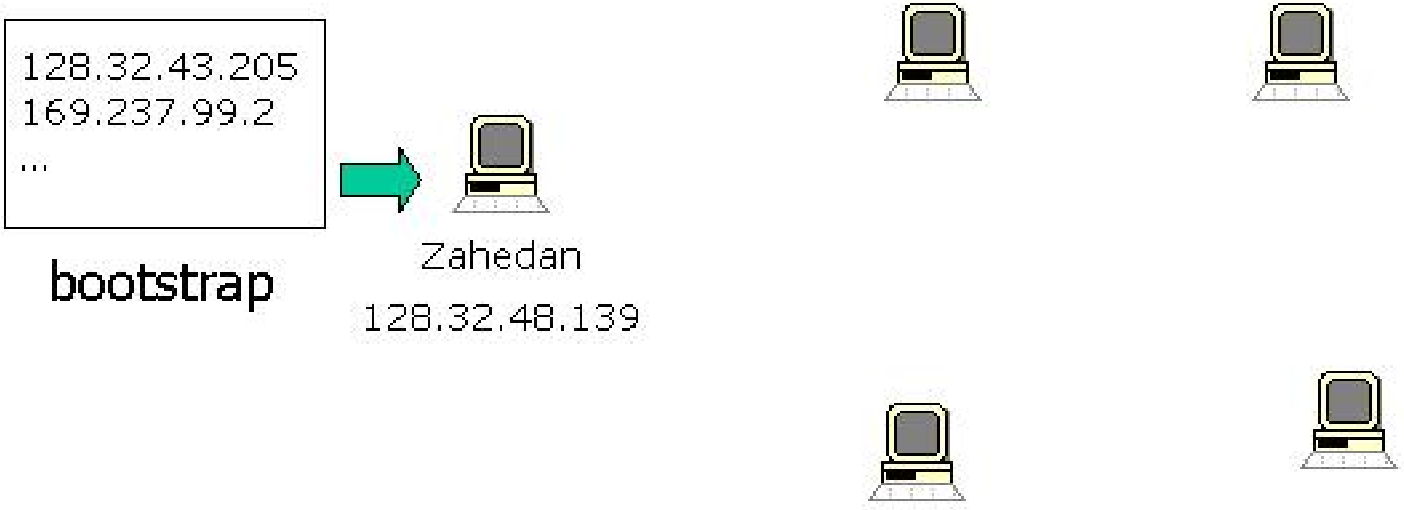
- Build a mesh that includes all participating end-hosts
- Build source routed distribution trees.

NARADA mesh creation

- All nodes can communicate with each other via unicast, but not all paths are good!
- Good mesh has two properties:
  - The quality of the path between any pair of members is comparable to unicast.
  - Each member has limited number of neighbors (commensurate to each nodes bandwidth)
- Mesh created incrementally as nodes join/leave and as nodes exchange state.

NARADA: Member Joins

- Join
  - New node obtains list of members via external mechanism. (can be out-of-date)
  - Node randomly selects neighbors from this list. Reselecting as necessary for non-responders.
  - Each node begins swapping its list of members with its neighbors.



128.32.43.205
169.237.99.2
...

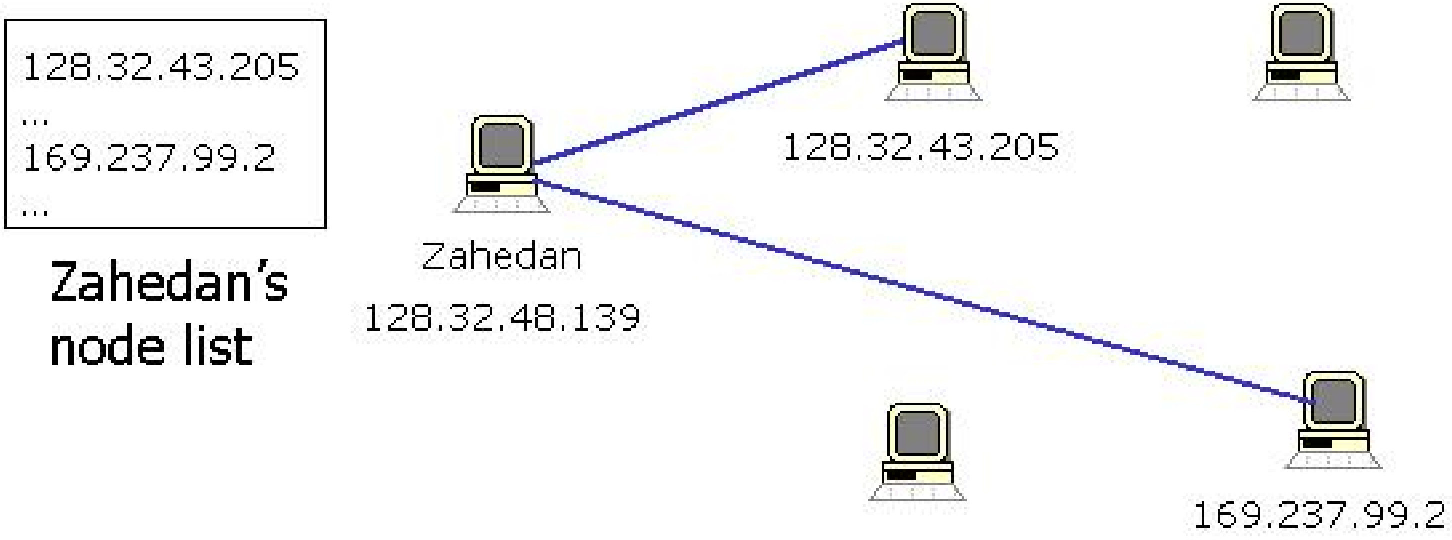
bootstrap

Zahedan

128.32.48.139

NARADA: Member Joins

- Join
  - New node obtains list of members via external mechanism. (can be out-of-date)
  - Node randomly selects neighbors from this list. Reselecting as necessary for non-responders.
  - Each node begins swapping its list of members with its neighbors.

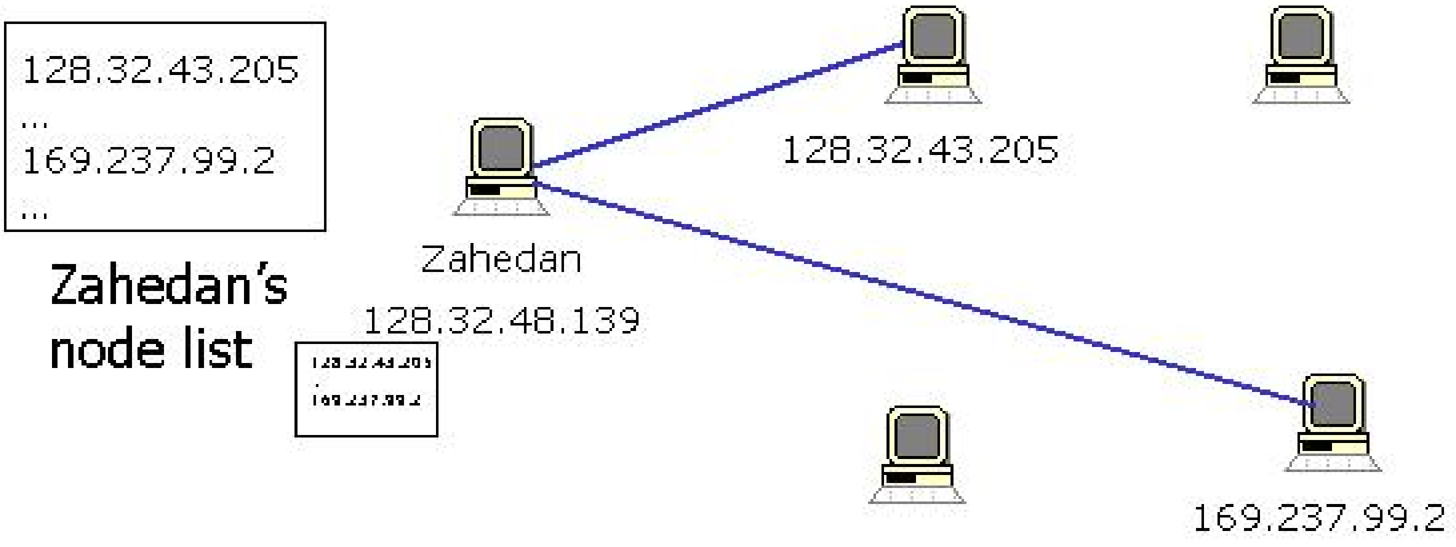


128.32.43.205
...
169.237.99.2
...

Zahedan's node list

NARADA: Member Joins

- Join
  - New node obtains list of members via external mechanism. (can be out-of-date)
  - Node randomly selects neighbors from this list. Reselecting as necessary for non-responders.
  - Each node begins swapping its list of members with its neighbors.



Narada: Leaves/Failures

- Leave
  - When node leaves it notifies neighbors, which propagate this information.
- Failure
  - Neighbor stops responding to probes (pings).
  - Add "dead member" to list of members.
  - Propagate "dead member" to neighbors.

How to scale? Distributed Hash Tables (DHT)

Example DHT: Pastry

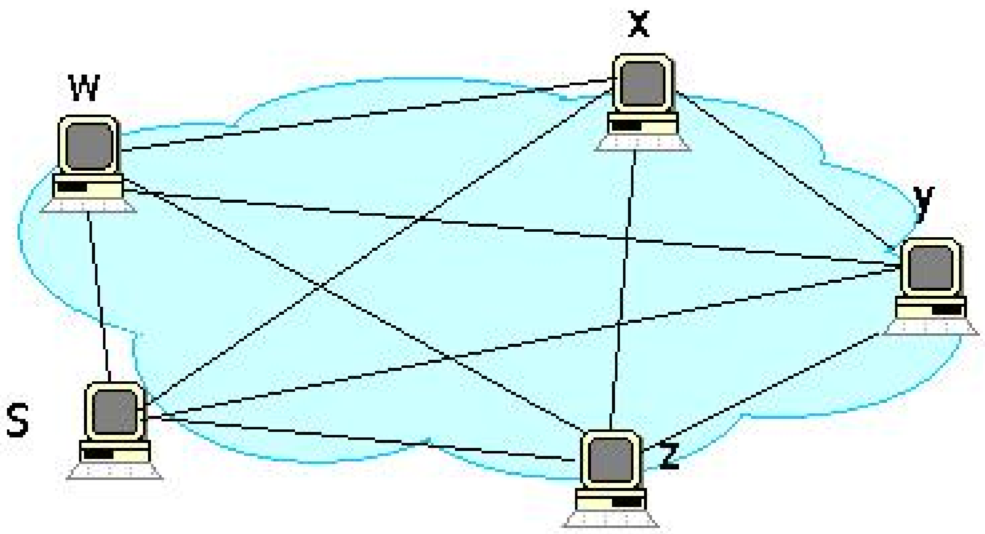
Example DHT Multicast: Scribe

Scribe + Video → SplitStream

Example Proxy-Based Multicast: Scattercast

Logical ESM Overlay Topology

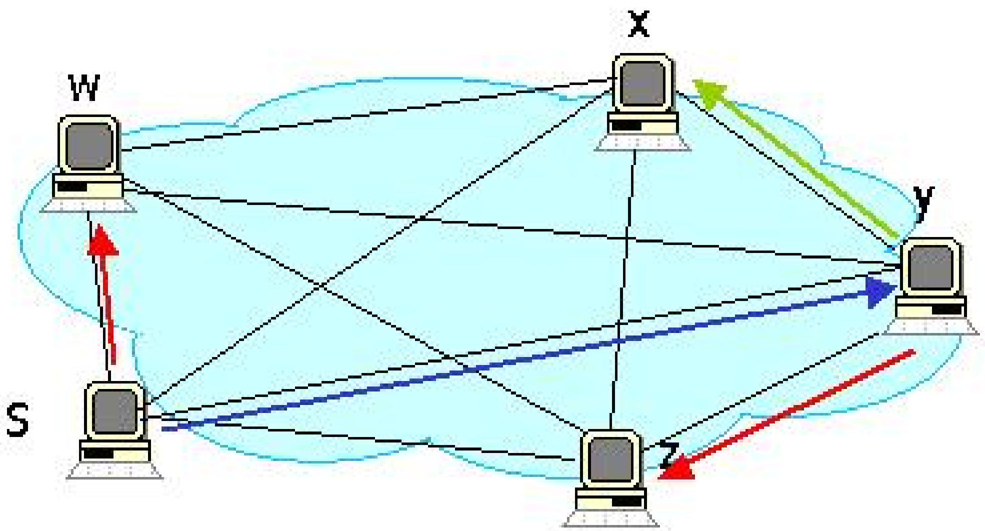
- IP topology is abstracted.
- All overlay nodes can have connectivity to all others.
- Typically only know RTT, loss rate.



TCP or UDP connections

Logical ESM Overlay Topology

- IP topology is abstracted.
- All overlay nodes can have connectivity to all others.
- Typically only know RTT, loss rate.



TCP or UDP connections