End System Multicast

Hui Zhang

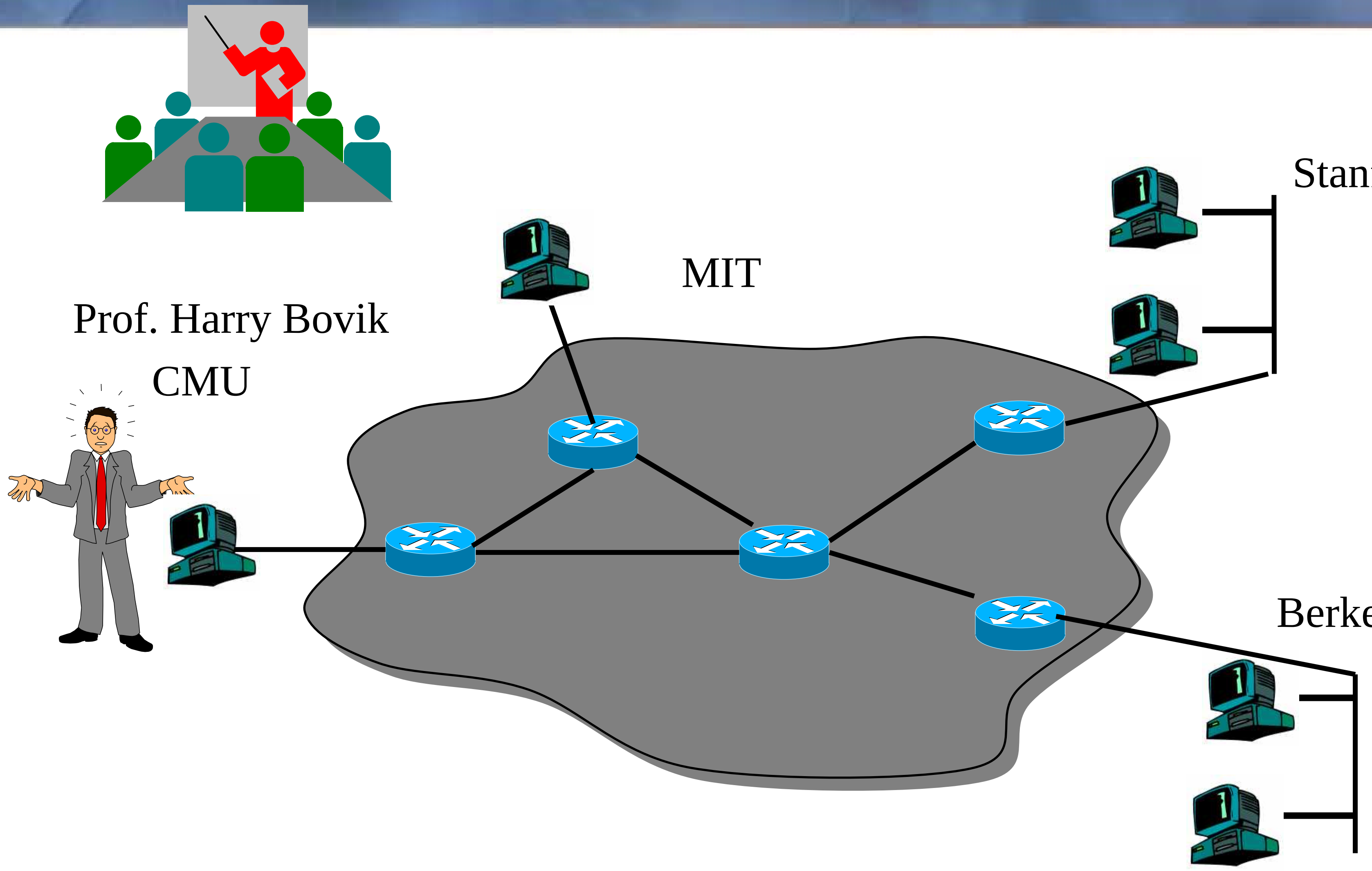
School of Computer Science

Carnegie Mellon University

May 2004

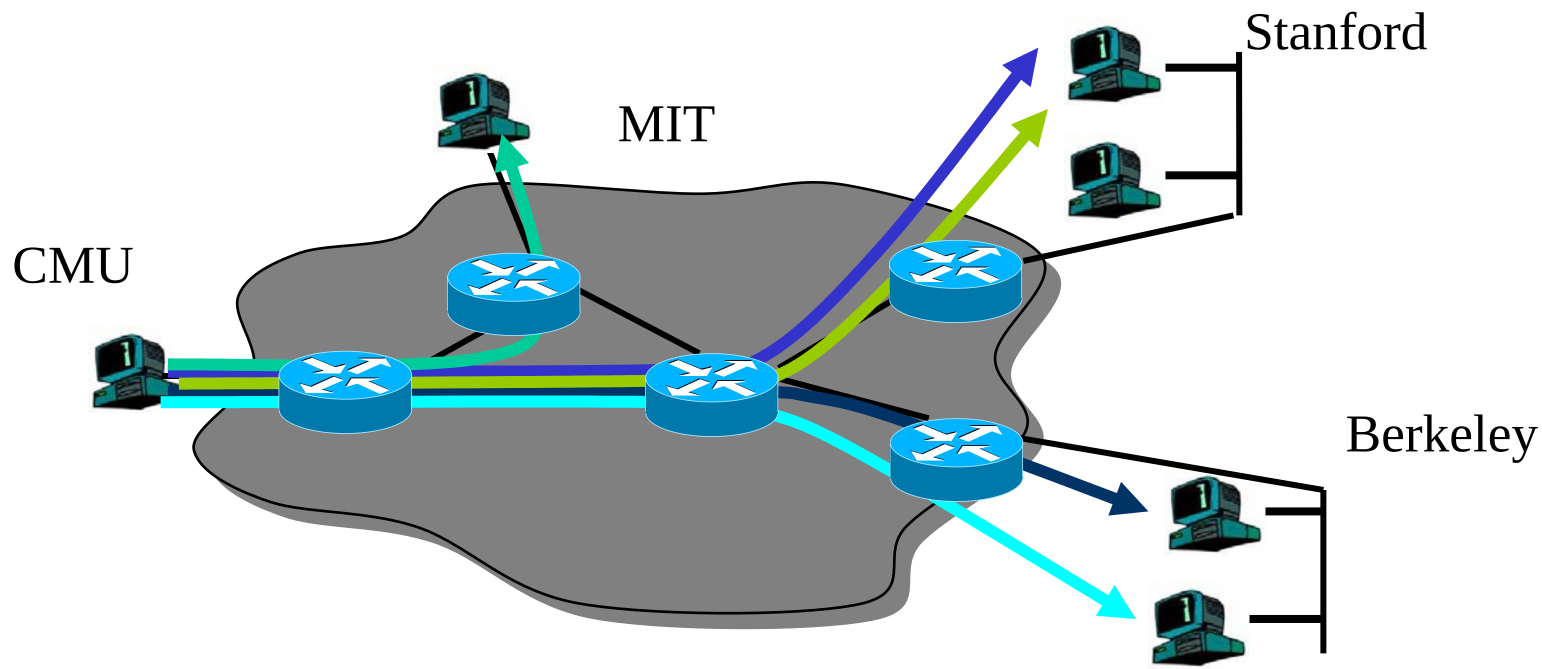
http://esm.cs.cmu.edu/

A Virtual Classroom



http://esm.cs.cmu.edu/

Solution Based on IP Unicast



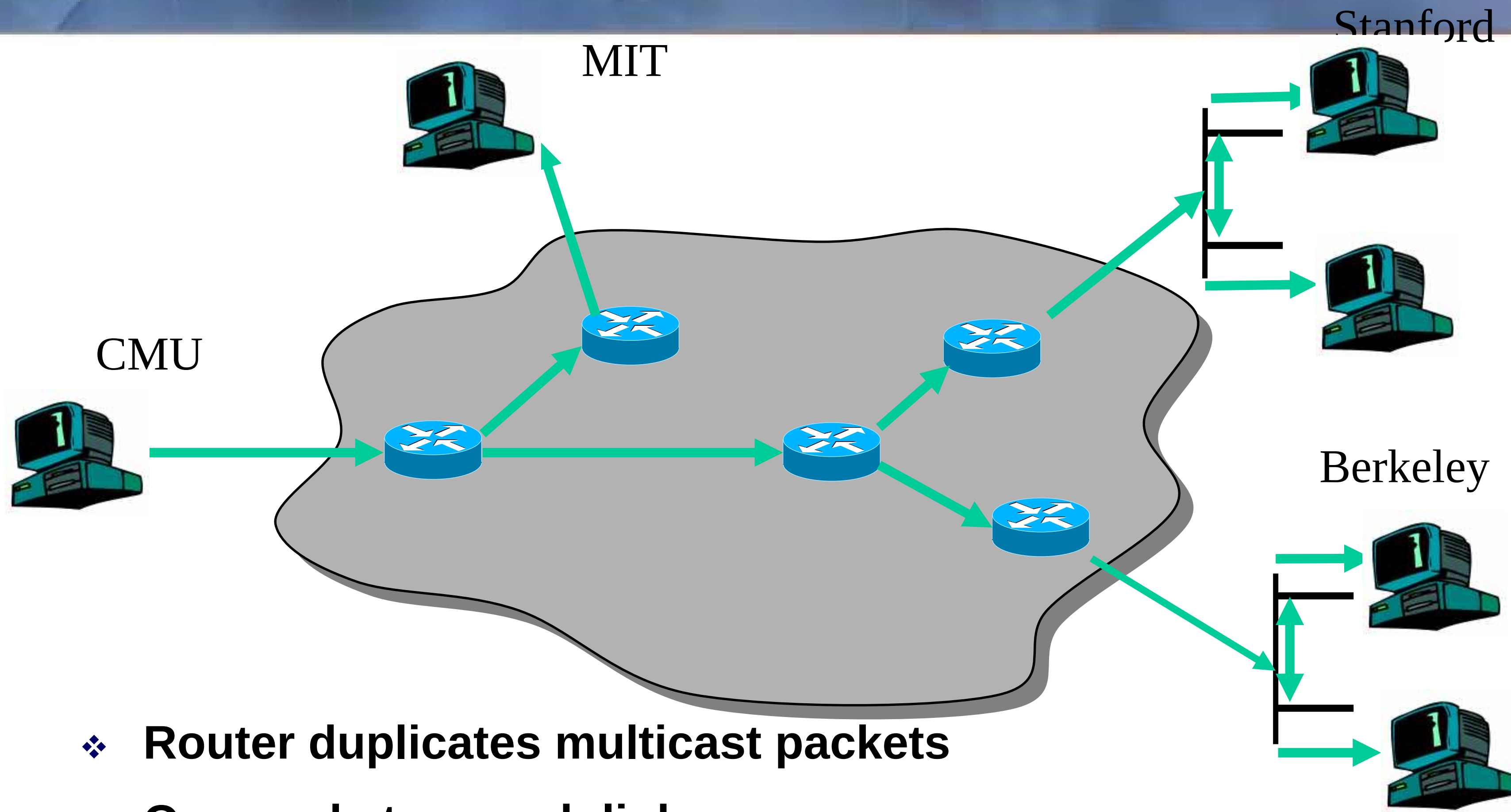
- **Poor performance scalability**
  - delay, throughput
  - sender, network

http://esm.cs.cmu.edu/

The Emerging Internet

- **Multi-party applications**
  - Audio/video conferencing
  - Multi-party games
  - Distributed simulation
  - Broadcast of web cams
  - Subscriber-publisher
- **Consider a world with ...**
  - Tens of millions of simultaneously running multi-point applications
  - Each application with tens to several thousand of end points

IP Multicast



- Router duplicates multicast packets
- One packet on each link
- Good performance scaling property

http://esm.cs.cmu.edu/

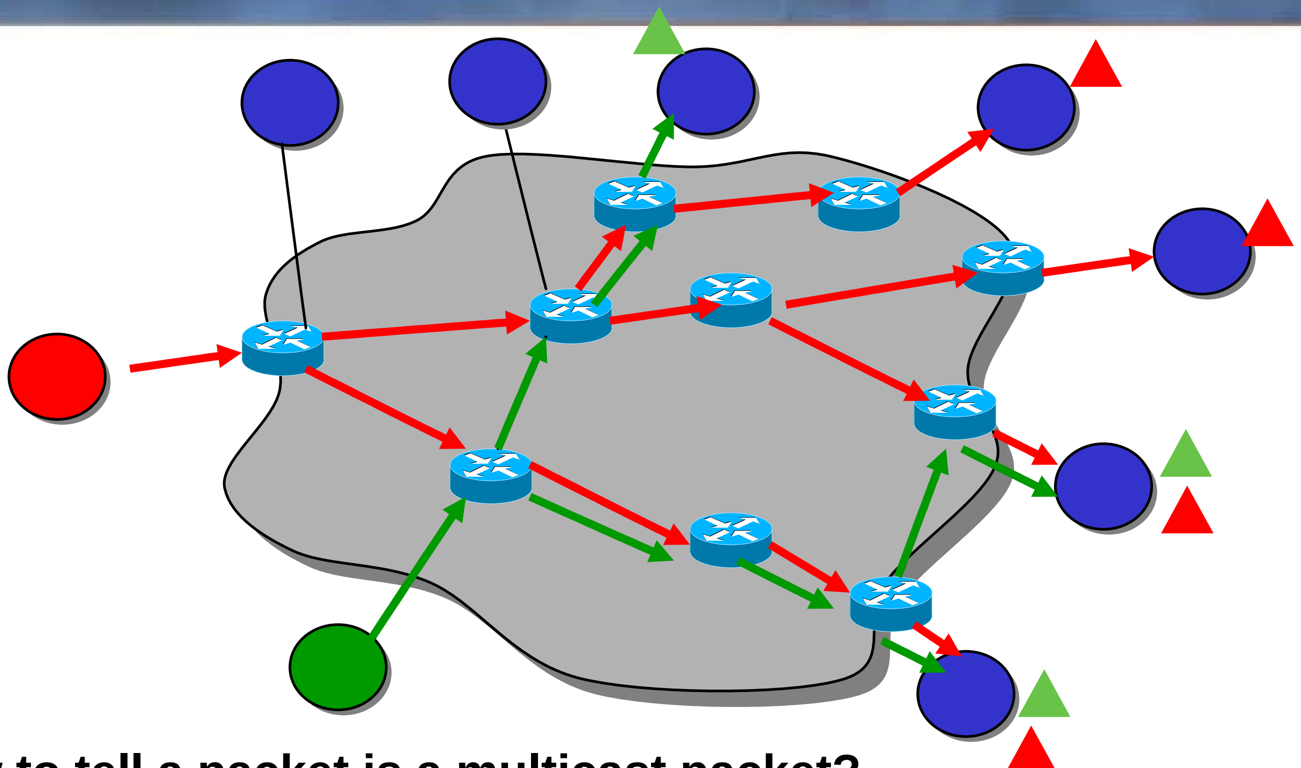
IP Multicast Overview

- **Seminal work by Steve Deering in 1989**
- **Huge amount of follow-on work**
  - Research
    - 1000s papers on multicast routing, reliable multicast, multicast congestion control, layered multicast
    - SIGCOMM, ACM Multimedia award papers, ACM Dissertation Award
  - Standard: IPv4 and IPv6, DVMRP/CBT/PIM
  - Development: in both routers (Cisco etc) and end systems (Microsoft, all versions of Unix)
  - Deployment: Mbone, major ISP's
  - Applications: vic/vat/rat/wb …
- **Situation today**
  - Still not used across the Internet

Many Technical Problems Unsolved

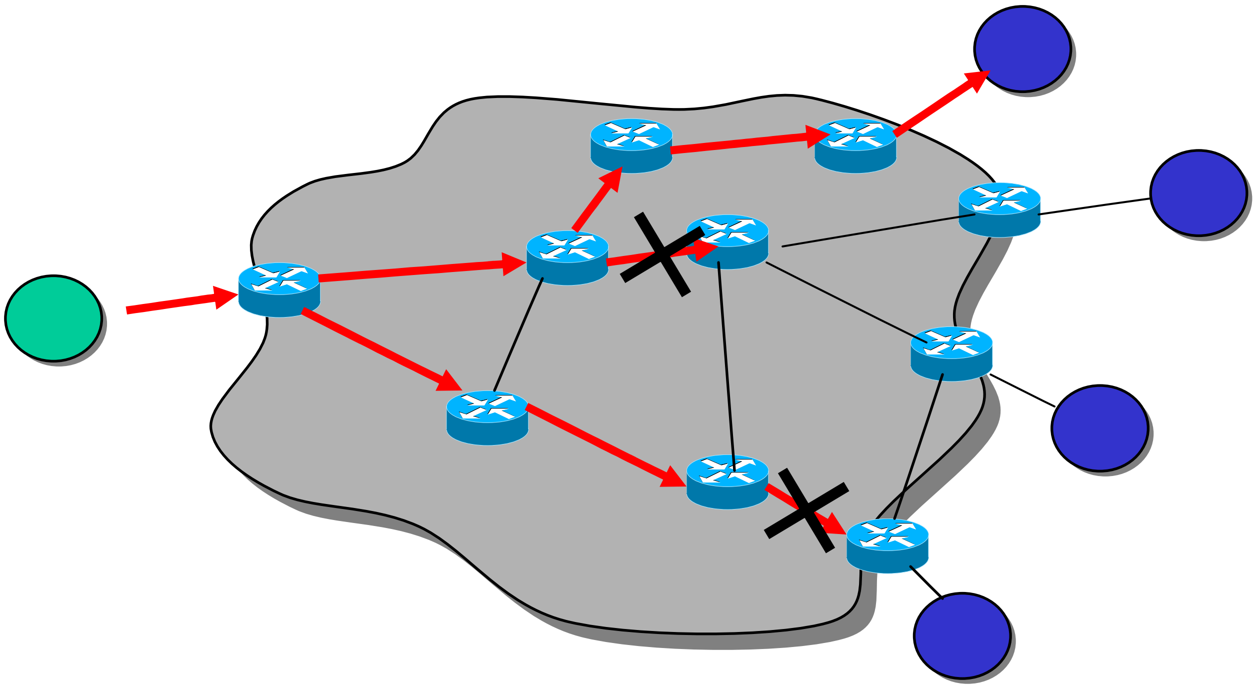
- Poor routing scalability property
- Difficult to support higher functionalities
- Serious security concern
- Address allocation

IP Multicast Scalability



- How to tell a packet is a multicast packet?
  - each group has a group address
- How to tell which hosts are in the group?
- How to decide where and how to branch?
  - routing protocol needs to set up per group state at routers
- Multi-point connection? Scalability and Robustness?

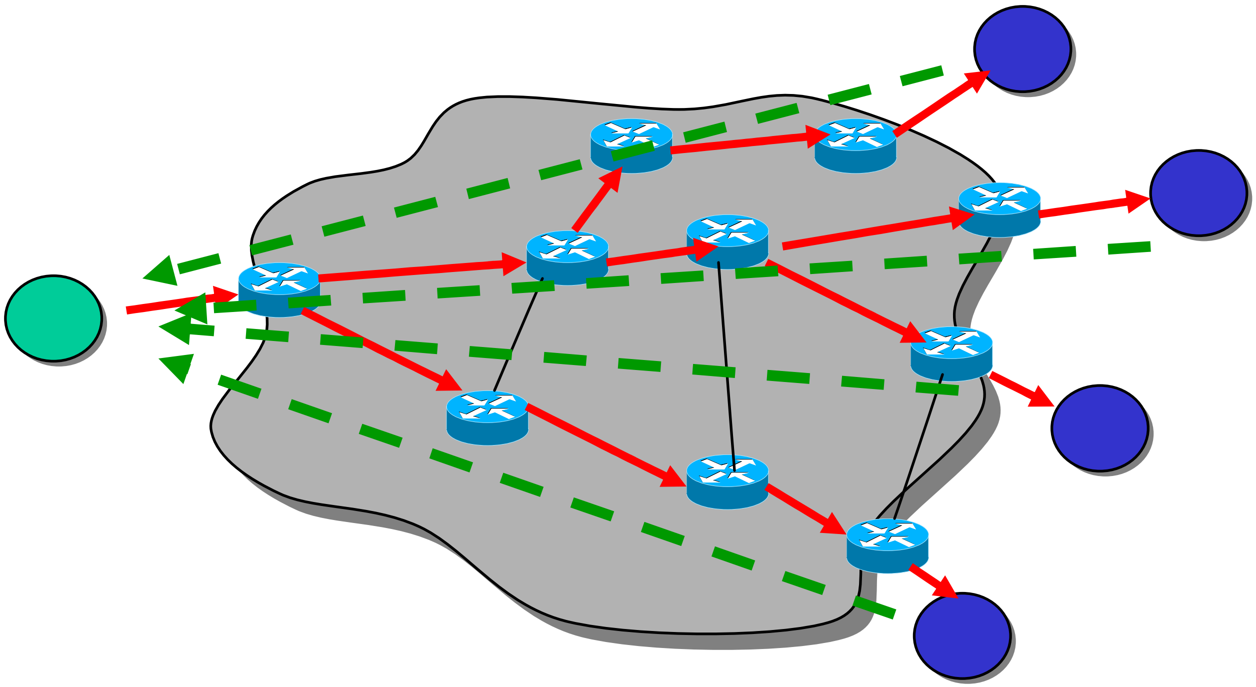
Error Control: Reliable Multicast



- IP is best-effort
- How to achieve reliable delivery?

http://esm.cs.cmu.edu/

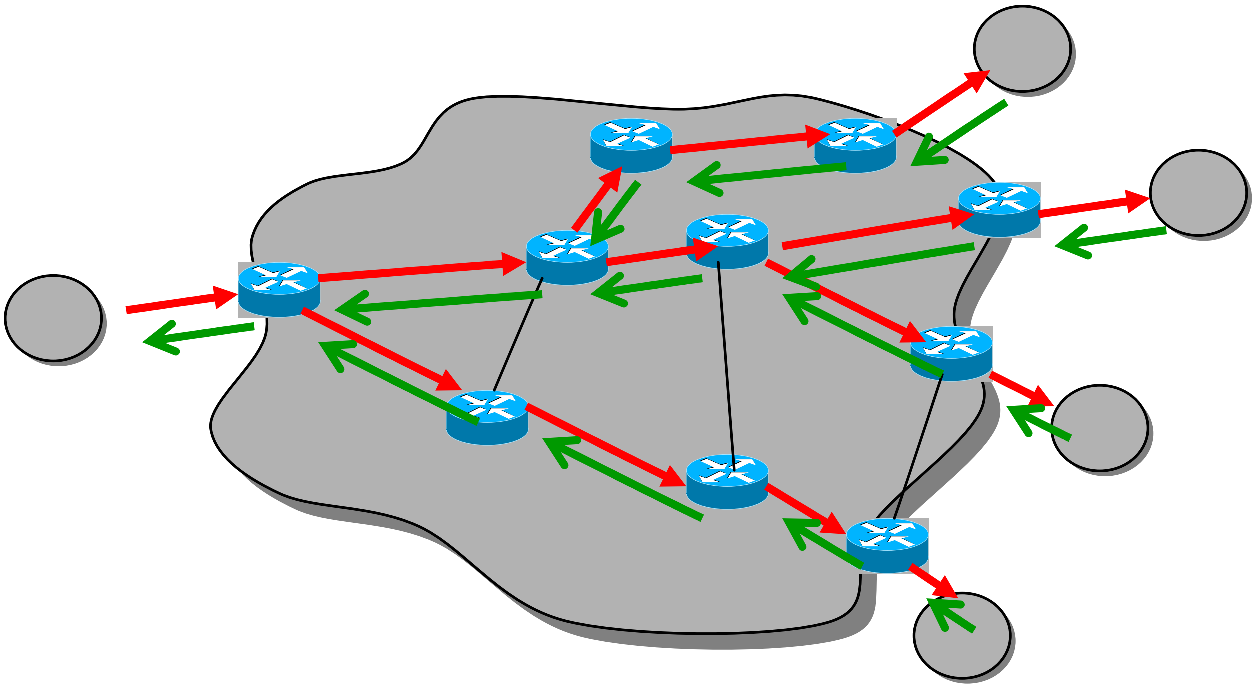
Ack Implosion



- **Scalability: number of acks increase with number of receivers**

http://esm.cs.cmu.edu/

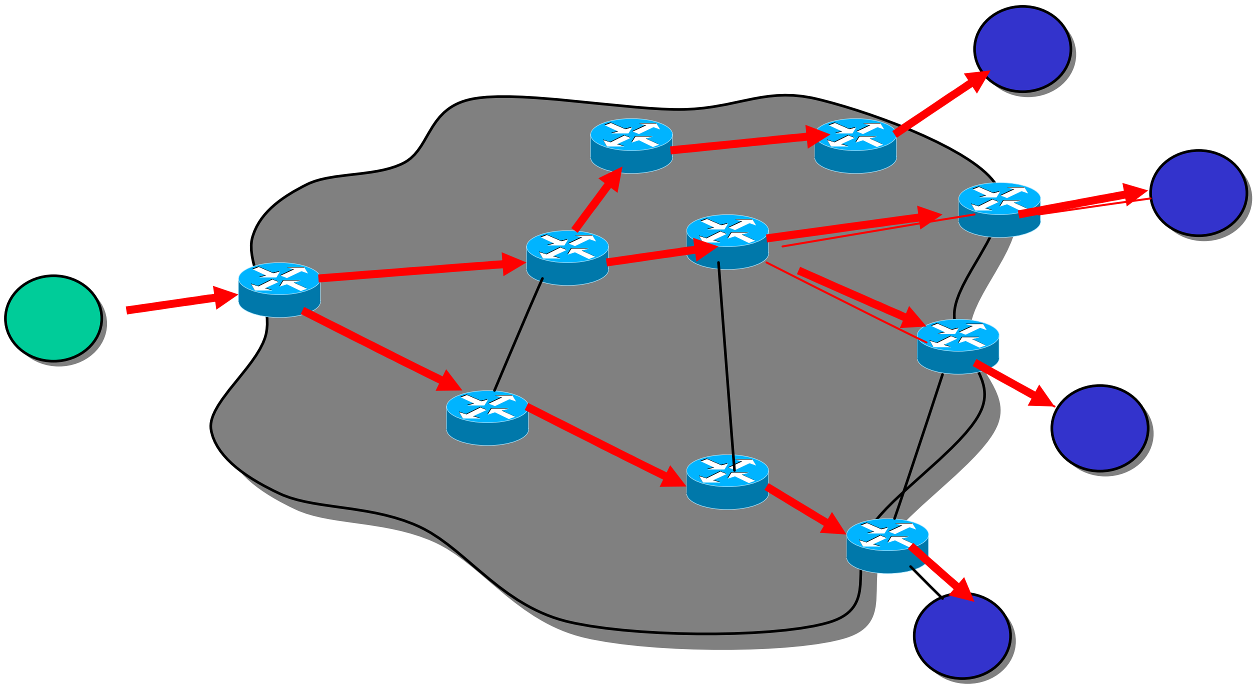
Routers Collect Acks



- **Overload router functionalities**
  - even more per group states

http://esm.cs.cmu.edu/

Congestion/Flow Control



- Diverse link technologies: different rates on each link
- Dynamic network condition: available bandwidth changes on each link
- What rate should sender transmit?

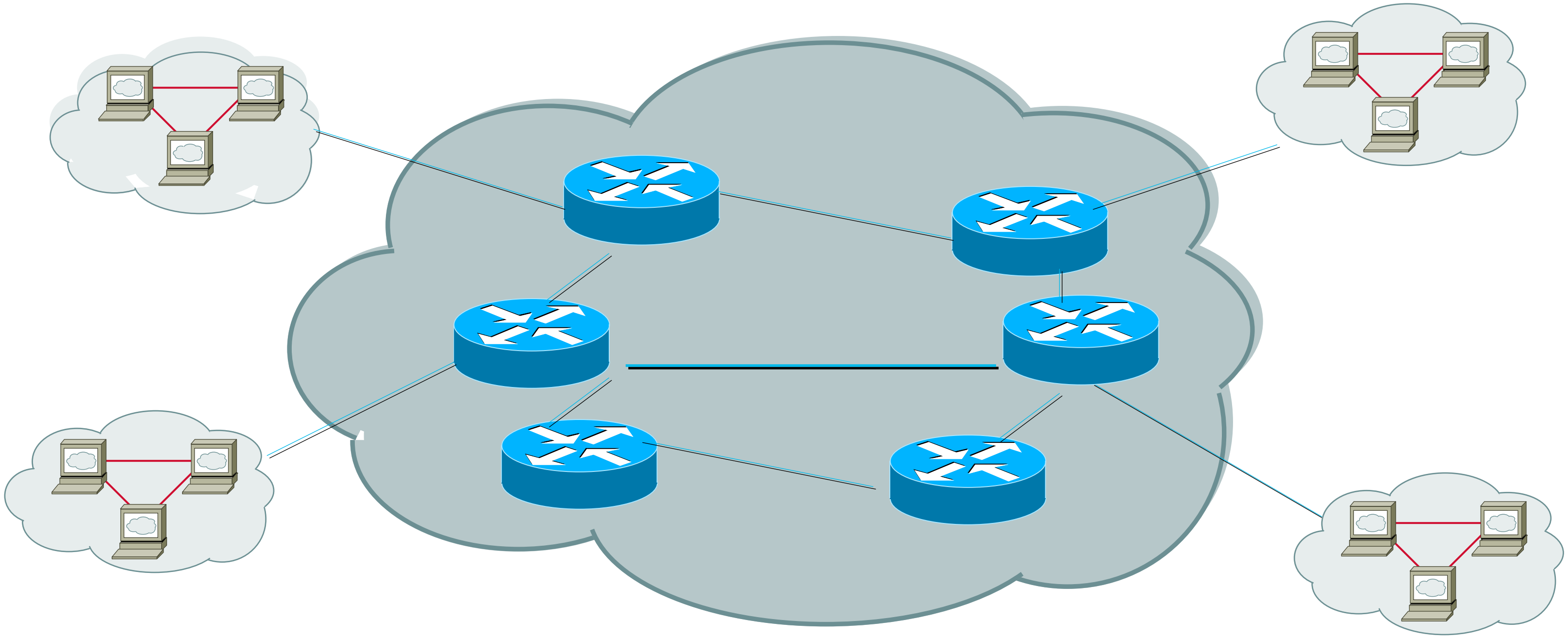
http://esm.cs.cmu.edu/

Many Technical Problems Unsolved

- **Poor routing scalability property**
  - routers need to keep per group/connection state
  - violation of fundamental Internet architecture principle
- **Difficult to support higher functionalities**
  - error control, flow control, congestion control
- **Serious security concern**
  - access control, both senders and receivers
  - Denial of Service attack
- **Address allocation**

End System vs. Network

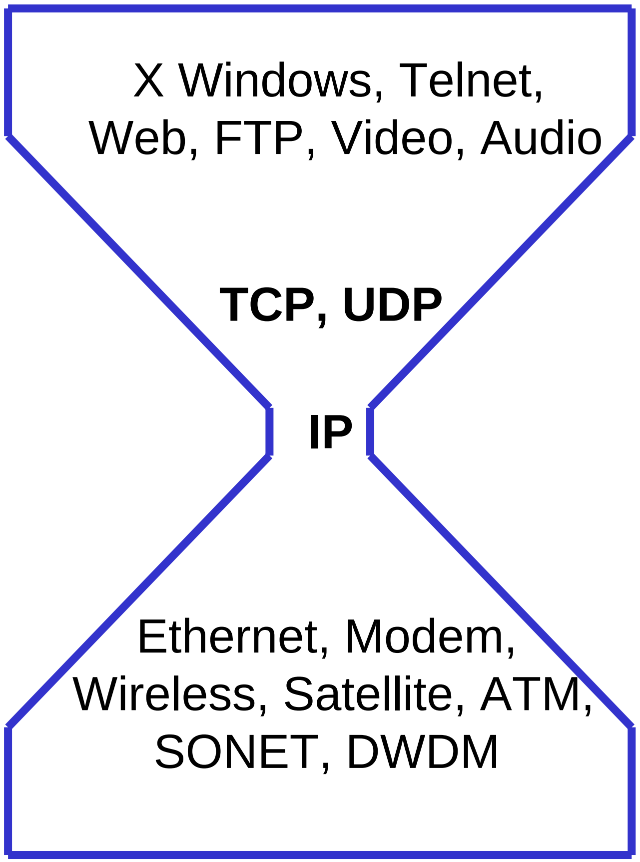
- **One of the most important design decisions in networks**
  - division of functionalities between hosts and routers, or
  - division of functionalities between end systems and networks



http://esm.cs.cmu.edu/

IP Architecture

- **"Dumb" IP layer**
  - minimal functionalities for connectivity
  - Unicast addressing, forwarding, routing
- **Smart end system**
  - transport layer or application performs more sophisticated functionalities
  - flow control, error control, congestion control
- **Advantages**
  - accommodate heterogeneous technologies
  - support diverse applications and decentralized network administration



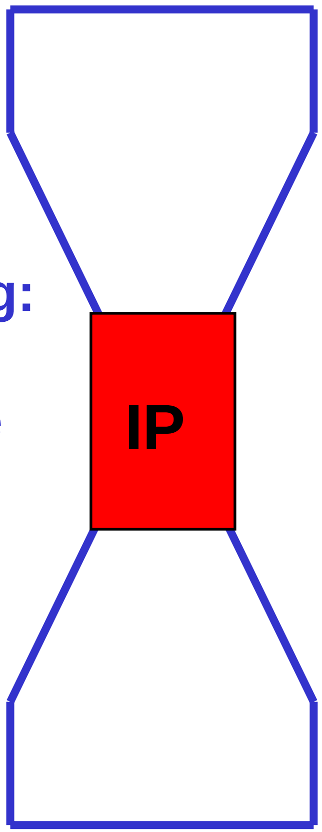
The "Hourglass Model",

http://esm.cs.cmu.edu/

Key Principle: Stateless Architecture

- **Minimalist IP layer maintains no per flow state**
- **IP layer maintains routing state**
  - Highly aggregated
  - 140K routing entries today for hundreds of millions hosts

What New Functionalities Should be Added to IP Layer ?



Steve Deering:

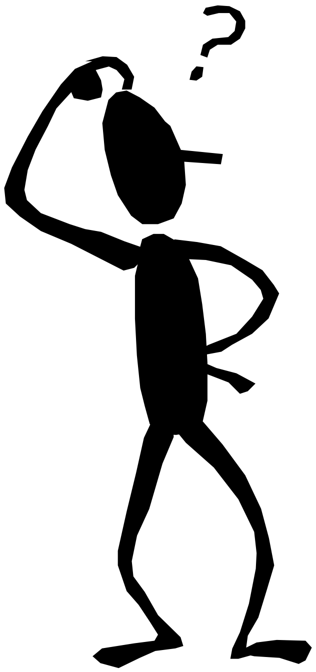
Watch for the Waist of IP Hourglass

- **IP layer functionalities means functionalities that need to be implemented by all routers**
- **New additions to IP**
  - Quality of Service
    - Intserv: per flow state management
    - Diffserv: no per flow state management
  - Multicast
    - Per group state management
- **Others**
  - Mobility, security

http://esm.cs.cmu.edu/

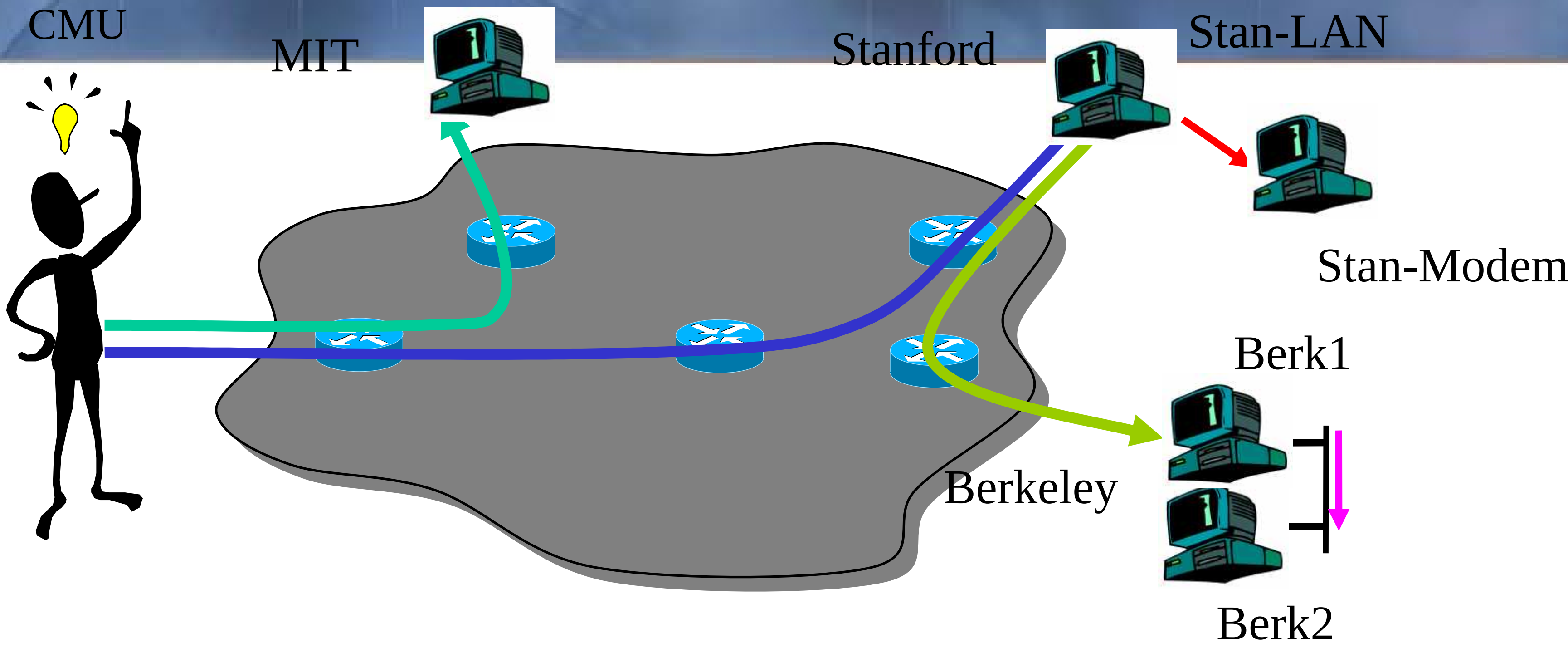
Multicast Revisited

- **Can we achieve**
  - efficient multi-point delivery,
  - without support from the IP layer?



http://esm.cs.cmu.edu/

End System Multicast

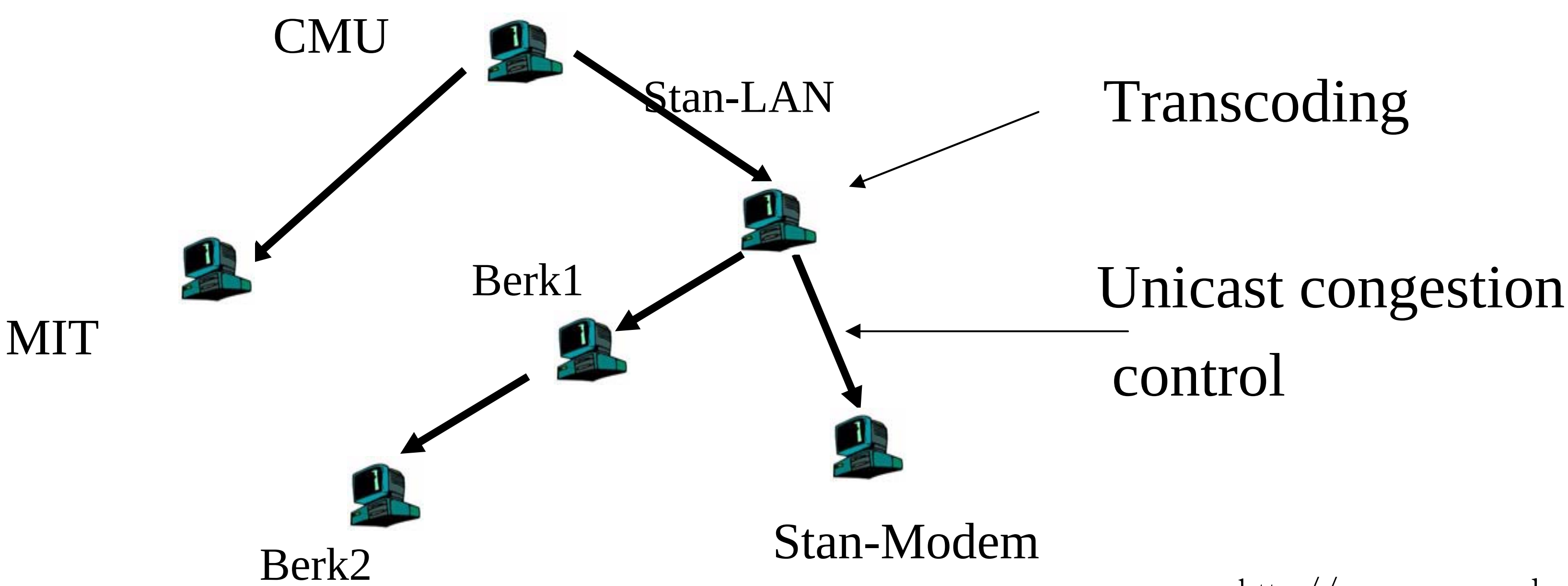


Overlay Tree

http://esm.cs.cmu.edu/

End System Multicast: Benefits

- **Scalability**
  - Routers do not maintain per-group state
- **Easy to deploy**
  - Works over the existing IP infrastructure
- **Can simplify support for higher level functionality**



http://esm.cs.cmu.edu/

ESM: The Unknowns

- **Several potential concerns with ESM**
  - What penalties are involved with an overlay approach?
  - How to organize receivers into efficient overlays?
  - Will users cooperate? ….

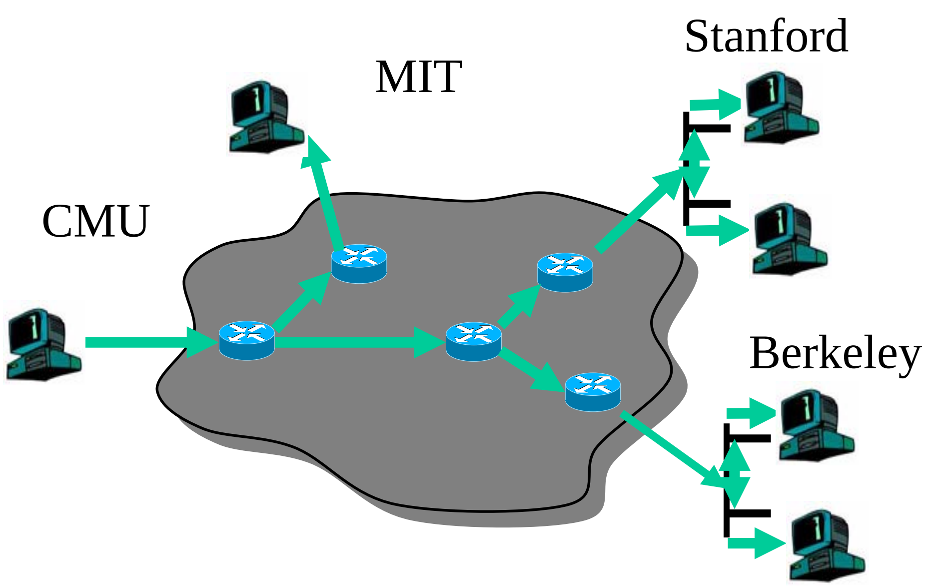
Is ESM viable?

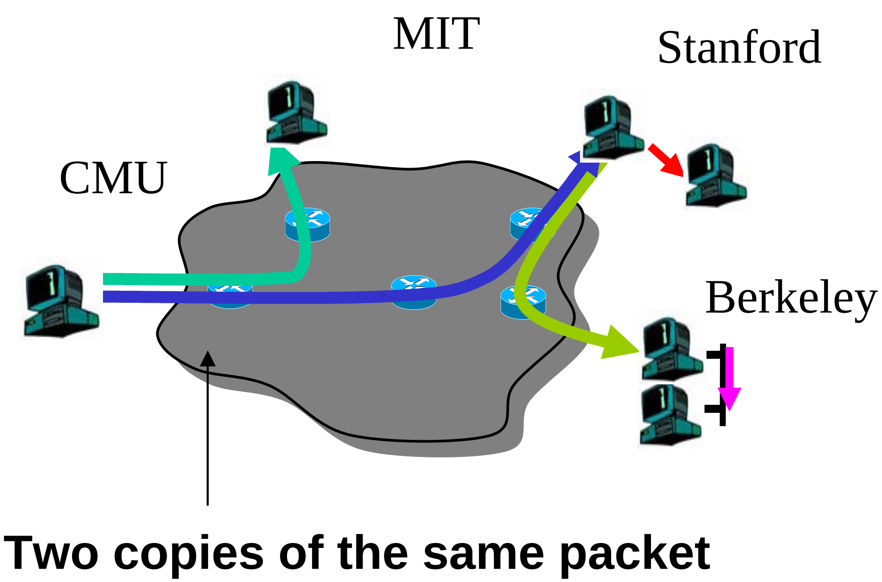
How far and real can we make the architectural vision?

http://esm.cs.cmu.edu/

Performance Challenges

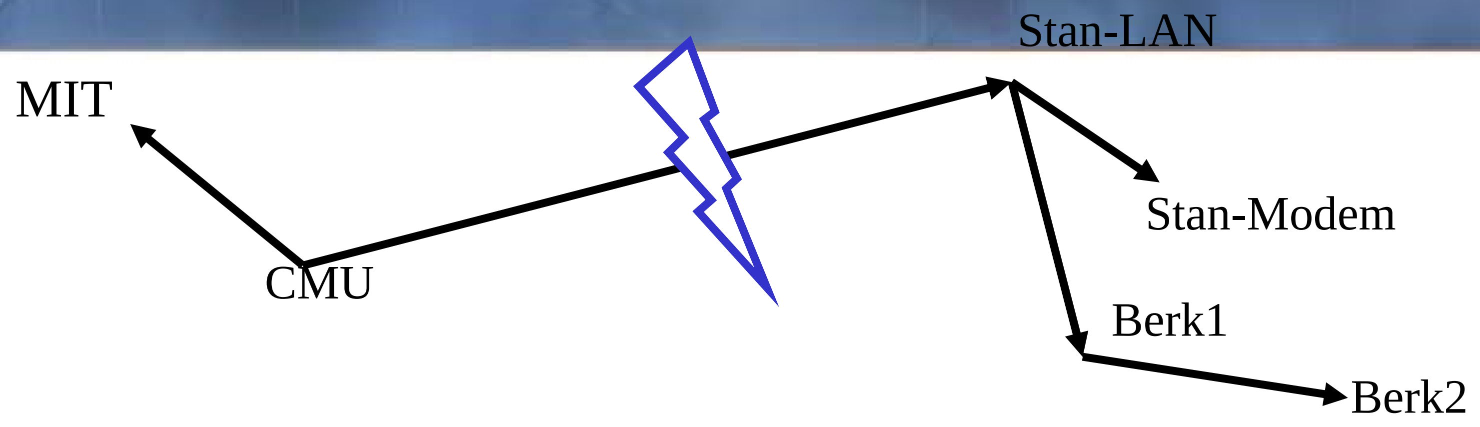
- Degradation in application performance: delay, throughput
- Network overhead: packet duplication over the same link



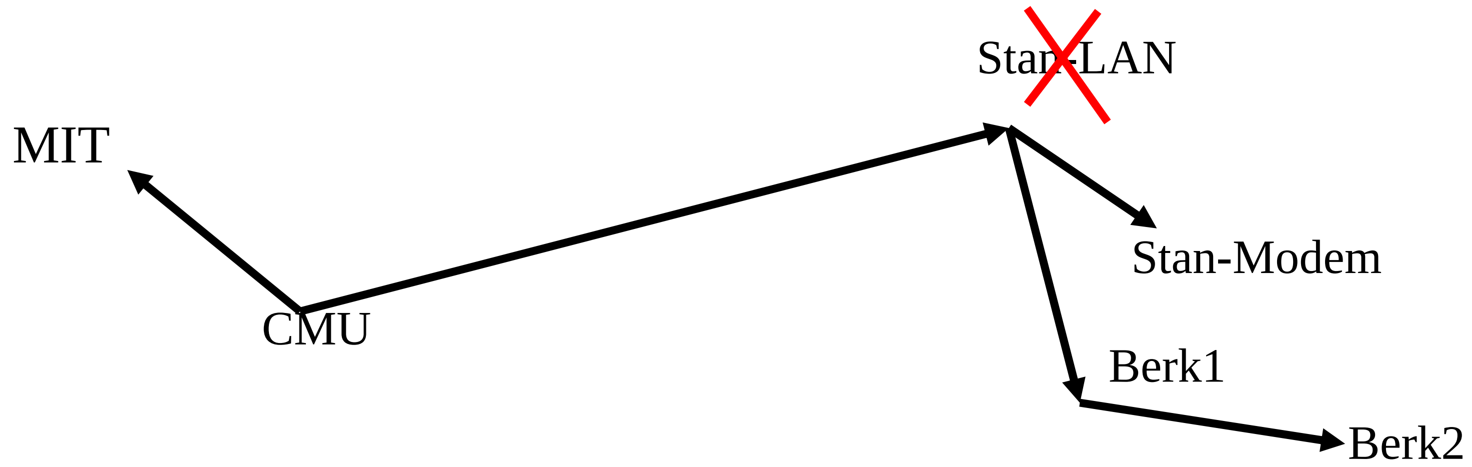


http://esm.cs.cmu.edu/

More Challenges



Overlays must adapt to network dynamics and congestion



Group membership is dynamic: members can join and leave

http://esm.cs.cmu.edu/

CMU ESM Project (1997 – present)

- **Laying the foundation (1997 – 2001)**
  - Self-organizing protocol
  - Simulation and Internet experiments to validate
- **Making it real (2002 – 2003)**
  - Build and deploy Internet video broadcast system based on ESM
- **Refining and Pushing it out (ongoing)**
  - Zero effort Internet video broadcast:
    - any host to any set of hosts
  - Incentive mechanism for end point cooperation
  - Mechanism for resource-constraint environment
  - Better virtual experiences by leveraging on-line features

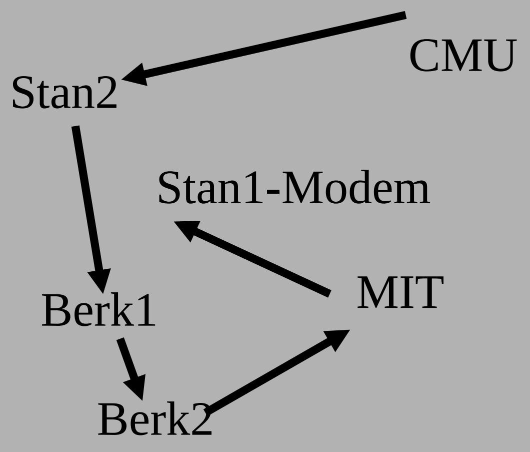
http://esm.cs.cmu.edu/

ESM Protocols

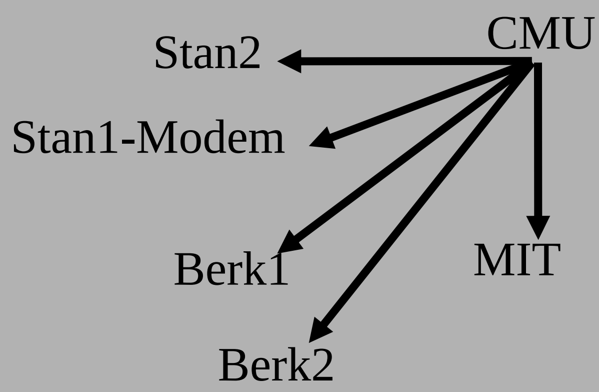
- **Objectives**
  - Self-organizing: adapt to dynamic membership changes
  - Self-improving: automatically evolve into efficient overlays
- **Two versions of protocol**
  - Multi-source, smaller scale conferencing apps
  - Single source, larger scale broadcasting apps

http://esm.cs.cmu.edu/

Inefficient Overlay Trees

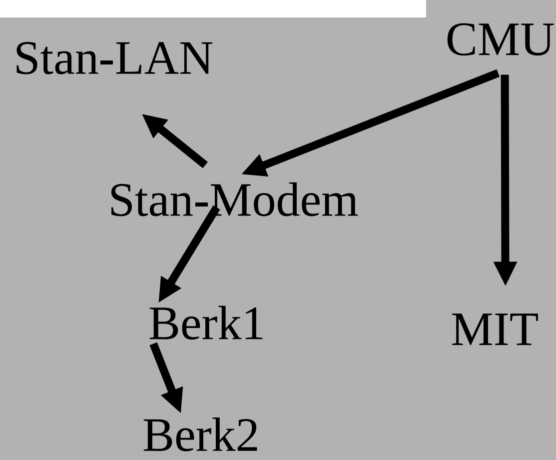


High latency



-Poor network usage

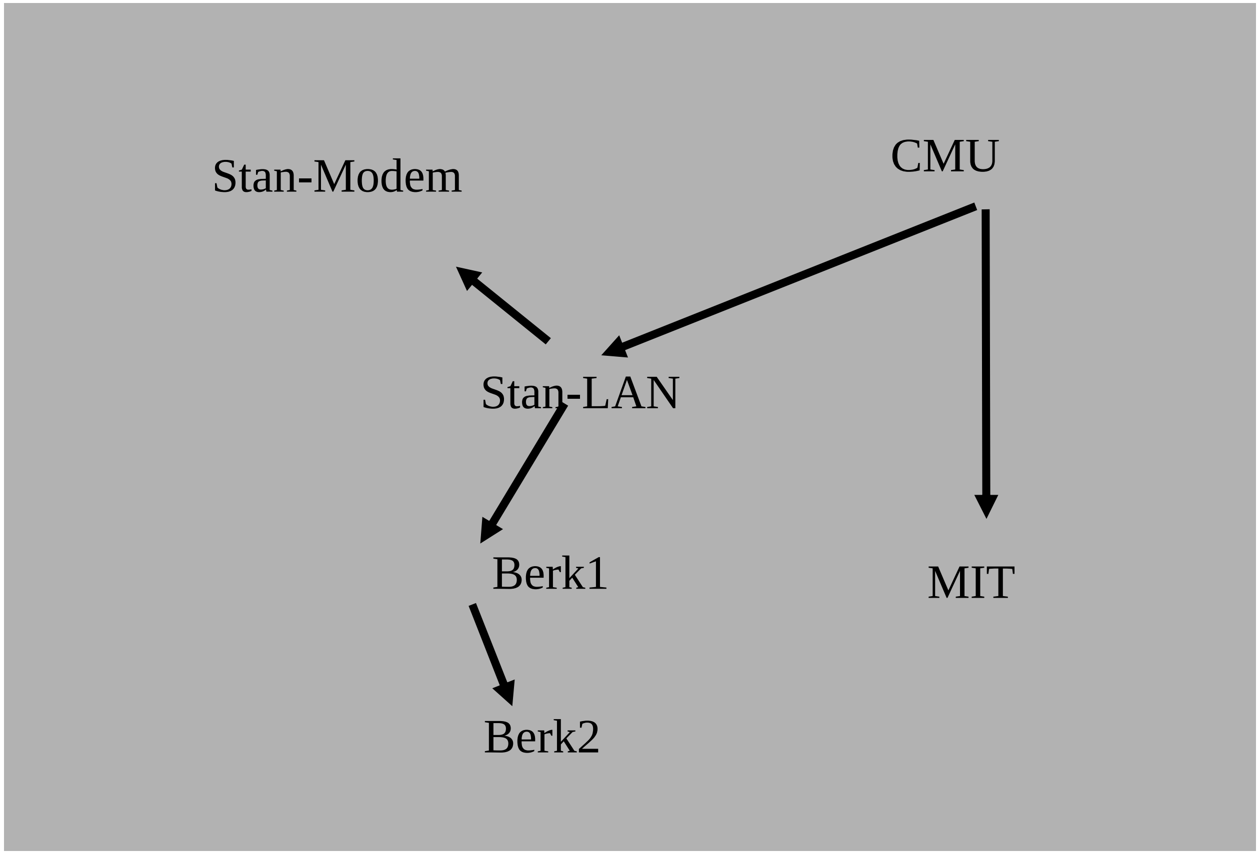
-Potential congestion near CMU



Poor bandwidth to members

http://esm.cs.cmu.edu/

An Efficient Overlay Tree



http://esm.cs.cmu.edu/

Key Components of Protocol

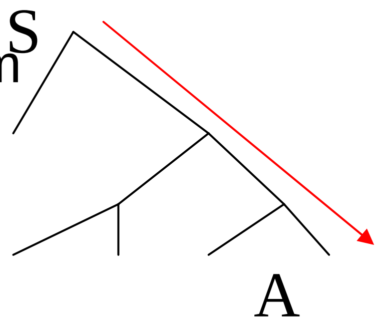
- **Overlay Management:**
  - How many other members does a member know?
  - How is this membership information maintained?

- **Overlay Optimization:**
  - Constructing efficient overlay among members

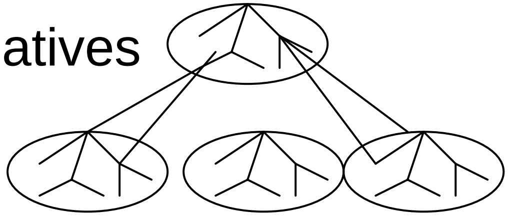
http://esm.cs.cmu.edu/

Group Management

- **Build separate control structure decoupled from tree**
  - Each member knows small random subset of group members
  - Information maintained using gossip-like algorithm
- **Members also maintain path from source**

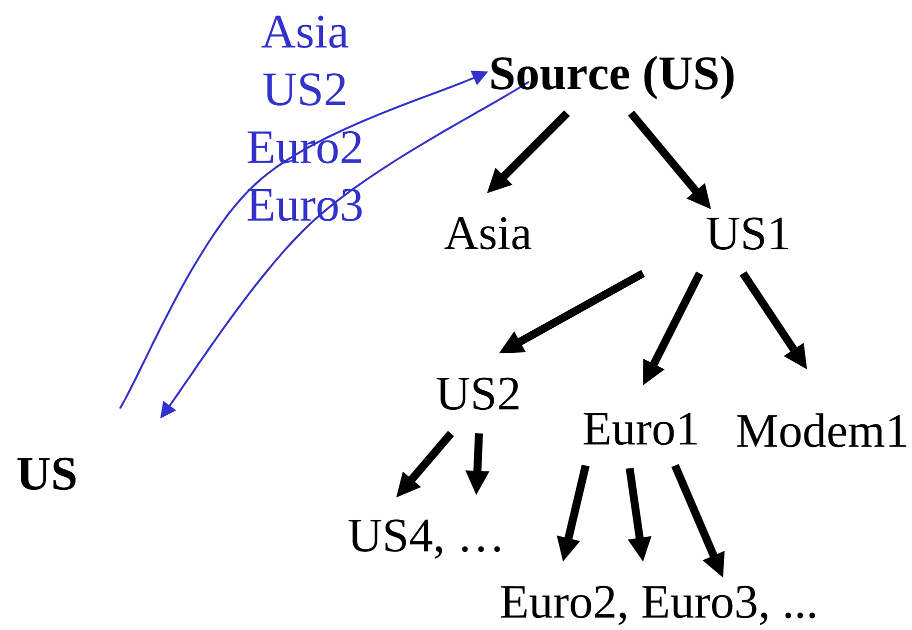


- **Other design alternatives possible:**
  - Example: a hierarchical structure, a DHT
  - No clear winner between design alternatives



http://esm.cs.cmu.edu/

Bootstrap



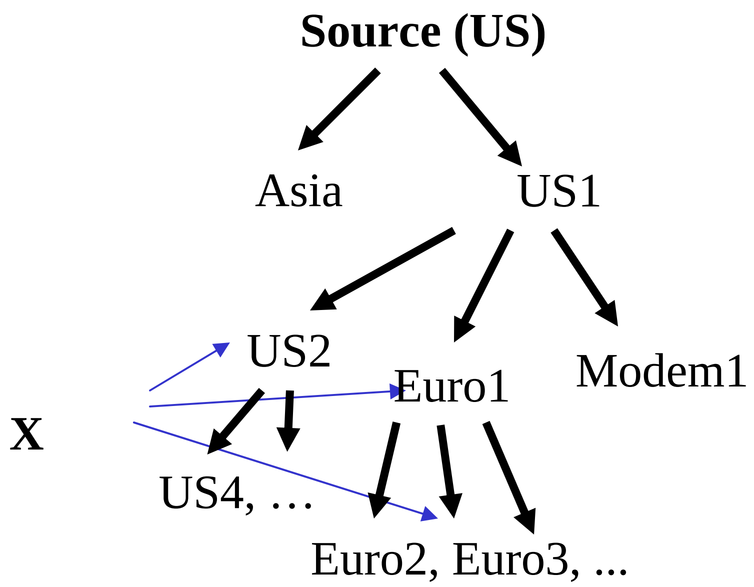
Node that joins:

–Gets a subset of group membership from source

–Finds parent using parent selection algorithm

http://esm.cs.cmu.edu/

Parent Selection

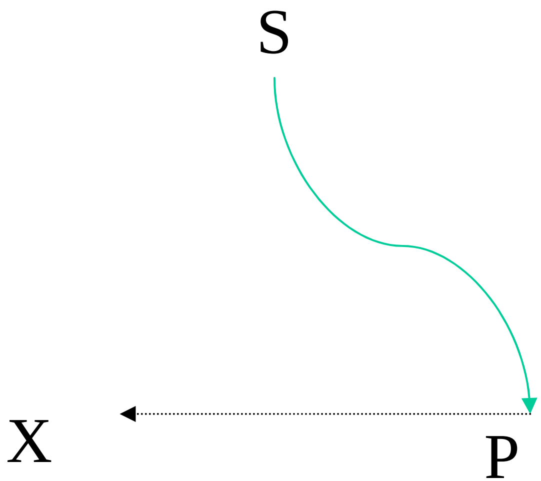


- *X* sends PROBE_REQ to subset of members it knows
- Evaluates remote nodes and chooses a candidate parent

http://esm.cs.cmu.edu/

Factors in Parent Selection

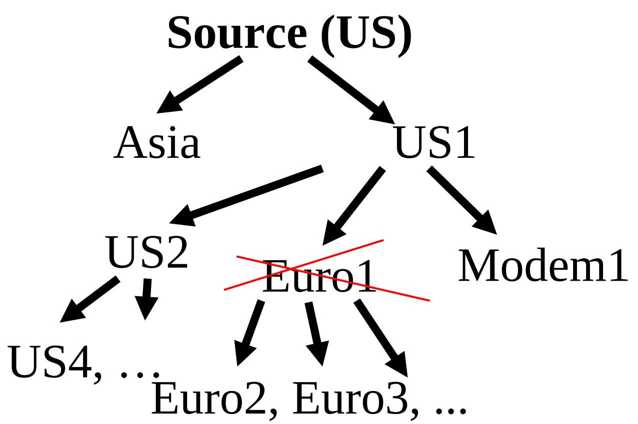
- **Filter out P if it is a descendant of X**
- **Performance of P**
  - Application throughput received by P
  - Delay of path from S to P
- **Saturation level of P**
- **Performance of link P-X**
  - Delay of link P-X
  - TCP bandwidth of link P-X



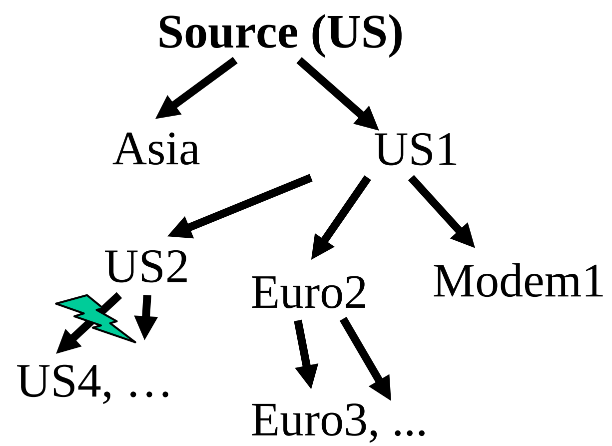
http://esm.cs.cmu.edu/

Causes for Parent Switch

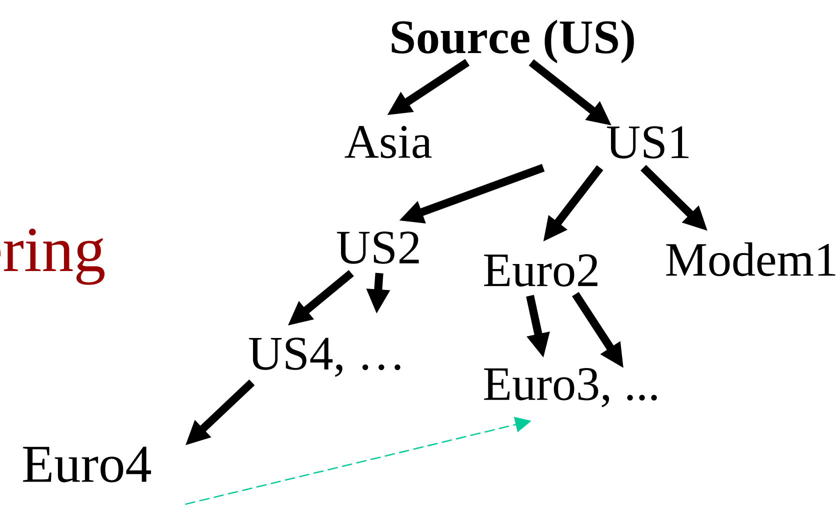
Member Leave/Death



Congestion/ poor bandwidth



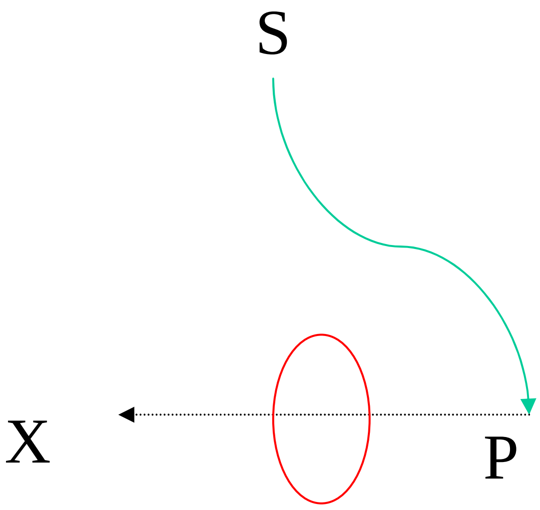
Better Clustering



http://esm.cs.cmu.edu/

Probing Heuristics

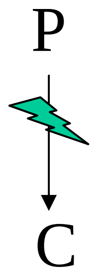
- **Study of light-weight probing heuristics**
  - RTT-probes, 10 KByte transfers, Bottleneck bandwidth
  - Simple RTT probes effective in lowering convergence time
  - Avoid probing hosts with low bottleneck bandwidth
- **History of performance of previously chosen parent**

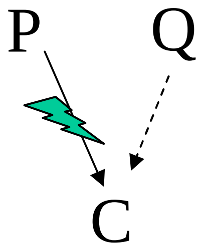


http://esm.cs.cmu.edu/

Bandwidth Adaptation

- **Detection Time: when to adapt to congestion?**
- **Constrained hosts tricky to tackle**
  - Hosts in Asia, behind wireless etc.
  - Need to avoid unproductive parent switches
  - Key difficulty: automatically detecting host is constrained
  - Duplicate parent heuristic could backfire



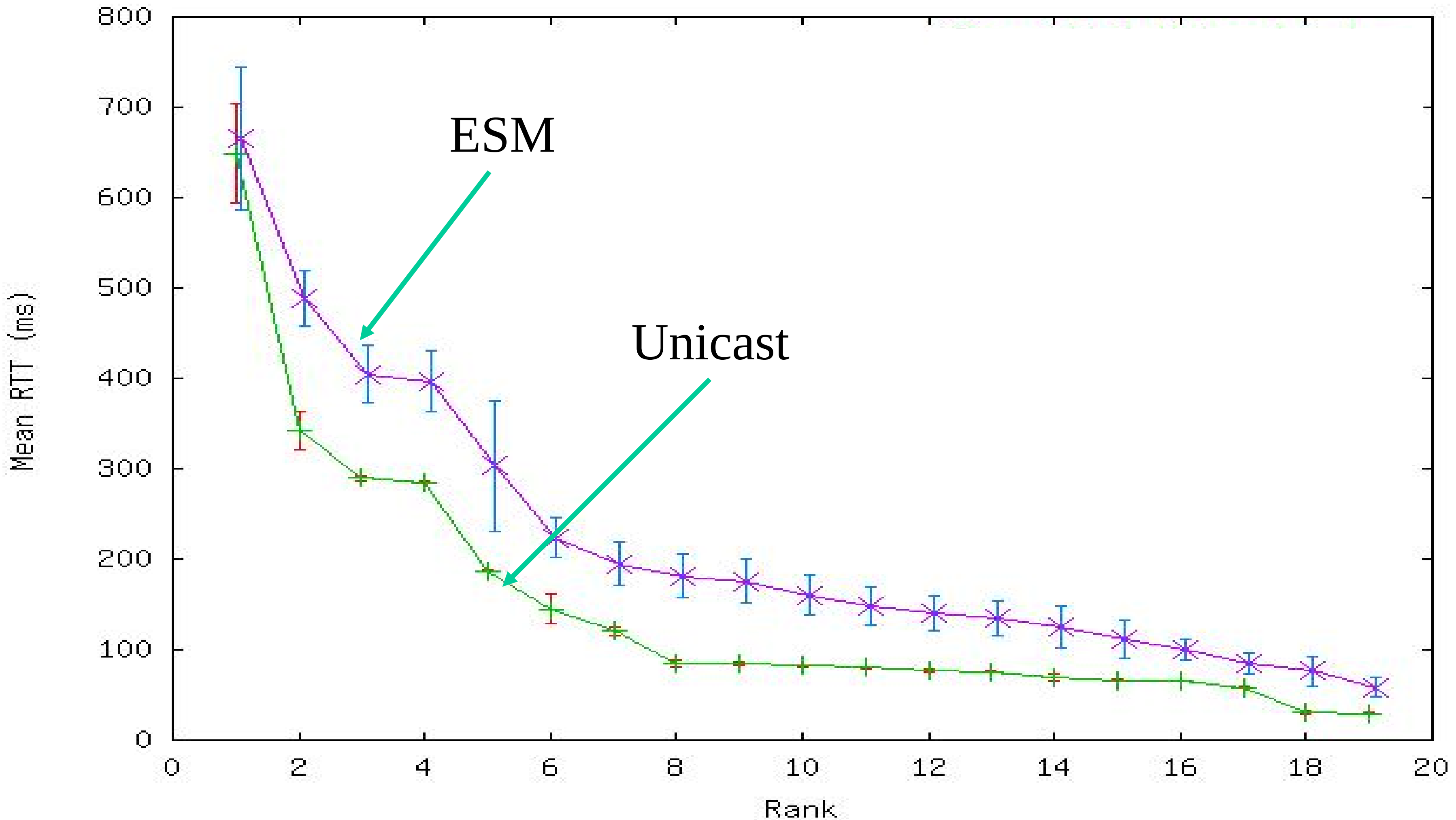


Evaluation

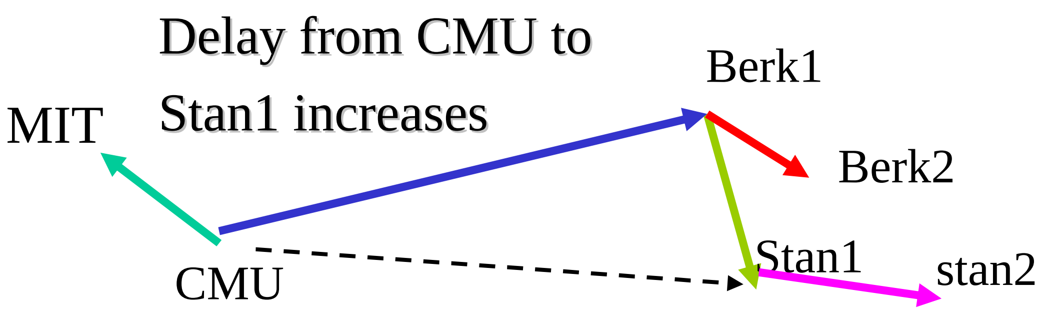
- **Driving Question**
  - Is ESM viable? What are the performance penalties involved?
- **Application level metrics**
  - Latency
  - Throughput
- **Network level Metrics**
  - Stress
  - Resource Usage
  - Protocol Overhead

Internet Test-bed (Sigcomm 2001)

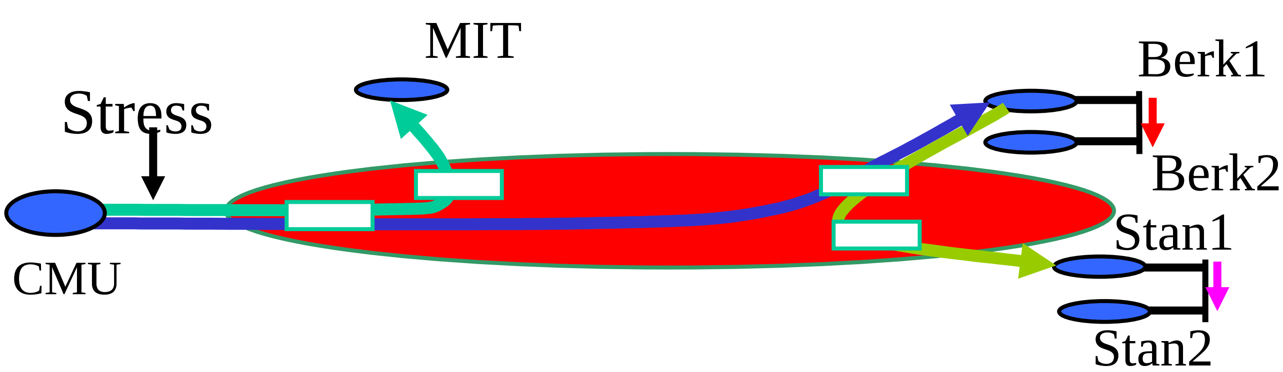
Twenty hosts: 1 DSL host, 3-4 hosts in Asia and Europe



Simulation Experiments
Sigmetrics 2000



Relative Delay Penalty (RDP)



Stress

Typical experiment with 128 members

– 90% of member pairs have RDP less than 4
– Stress reduced by factor of 14 compared to naïve unicast

http://esm.cs.cmu.edu/

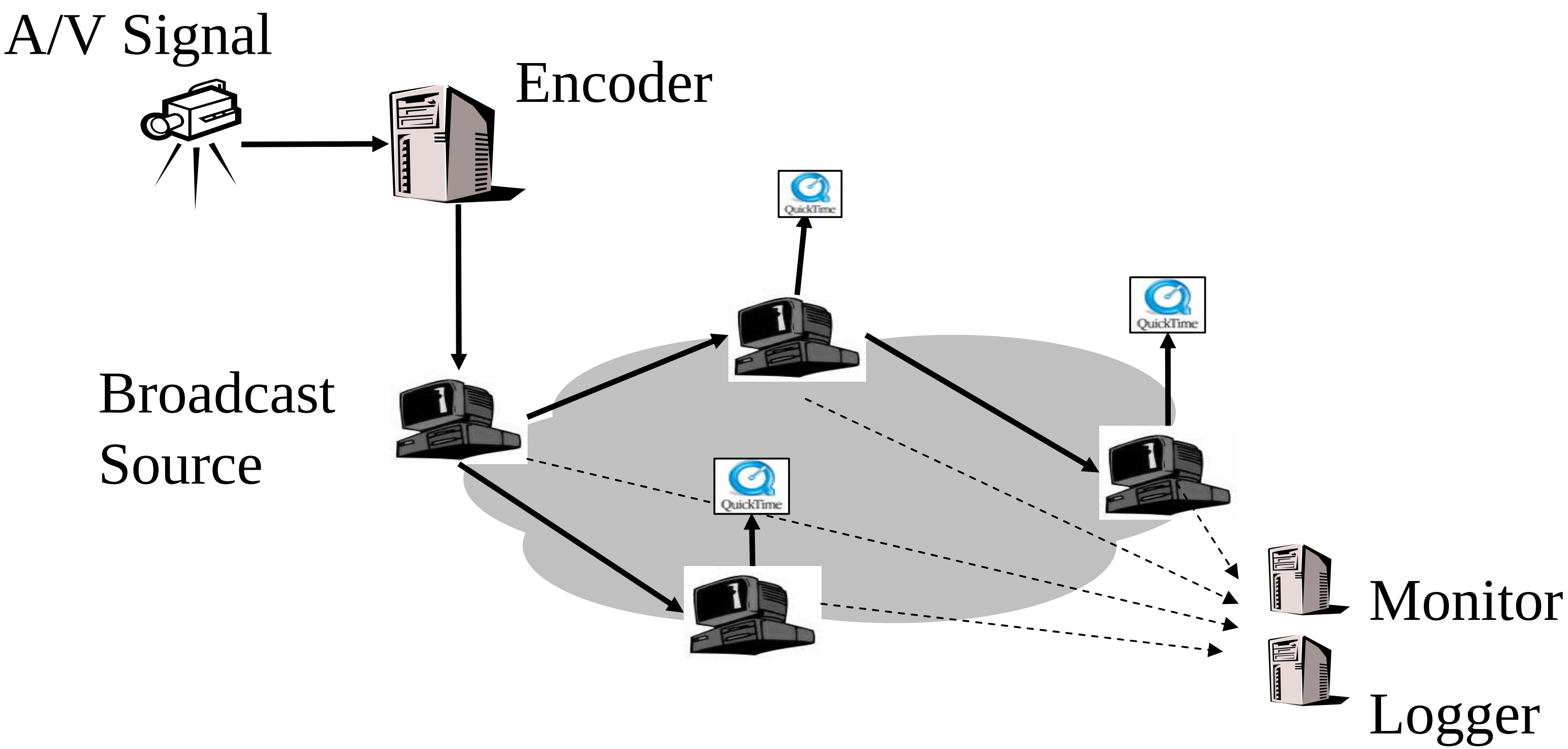
Limitations of Evaluation

- **Internet-based evaluation**
  - Scale limited by availability of expermental end points
  - Bias in end system selections
- **Simulation-based evaluation**
  - Scale limited by computing power and memory size
  - Difficult to model topology
  - Difficult to model dynamic cross traffic
- **Join and leave pattern? Duration?**

The Evaluation Question

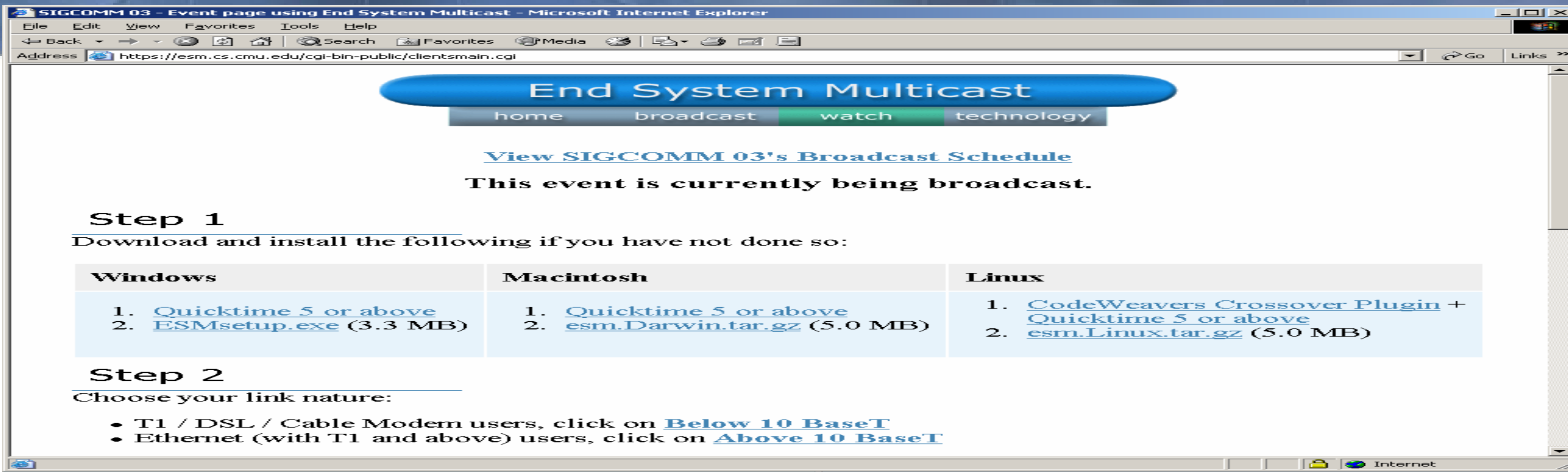
- **Question: how to evaluate Internet-scale systems?**
- **Answer: deploy Internet-scale application and attract real users**
- **Properties wanted**
  - High bandwidth, large number of simultaneous users
  - Free and compelling content
- **Anwer: auido/video webcast**

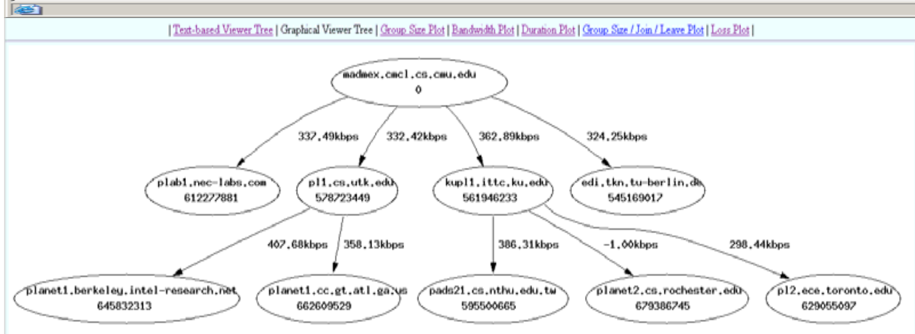
System Overview

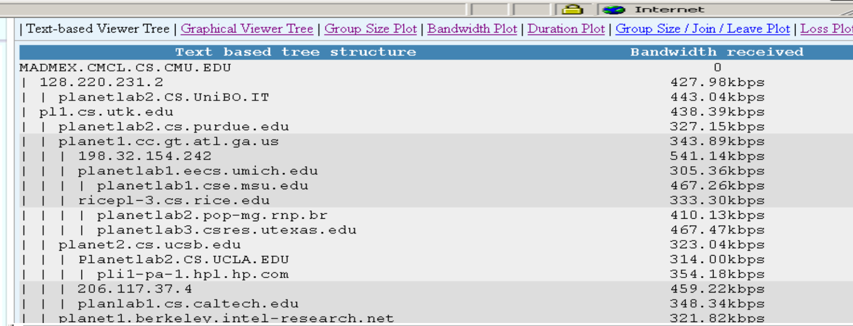


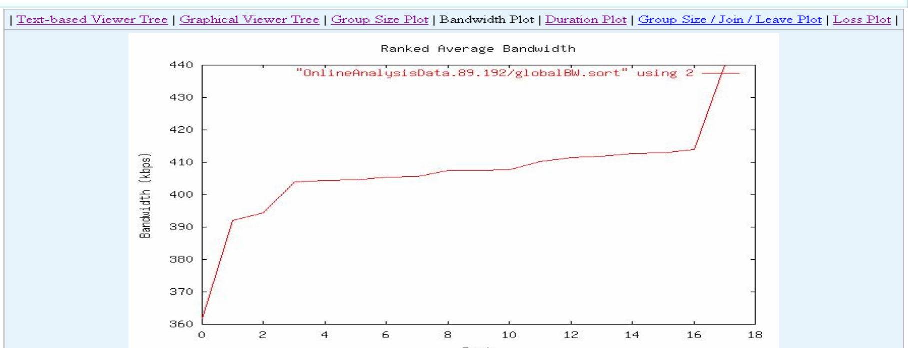
http://esm.cs.cmu.edu/

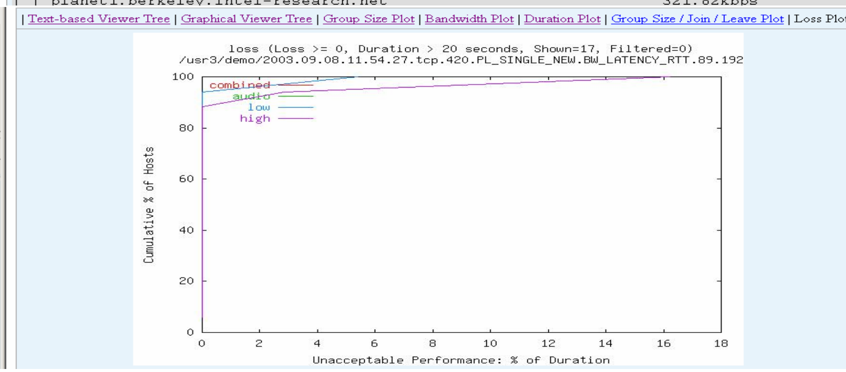
Publisher Toolkit



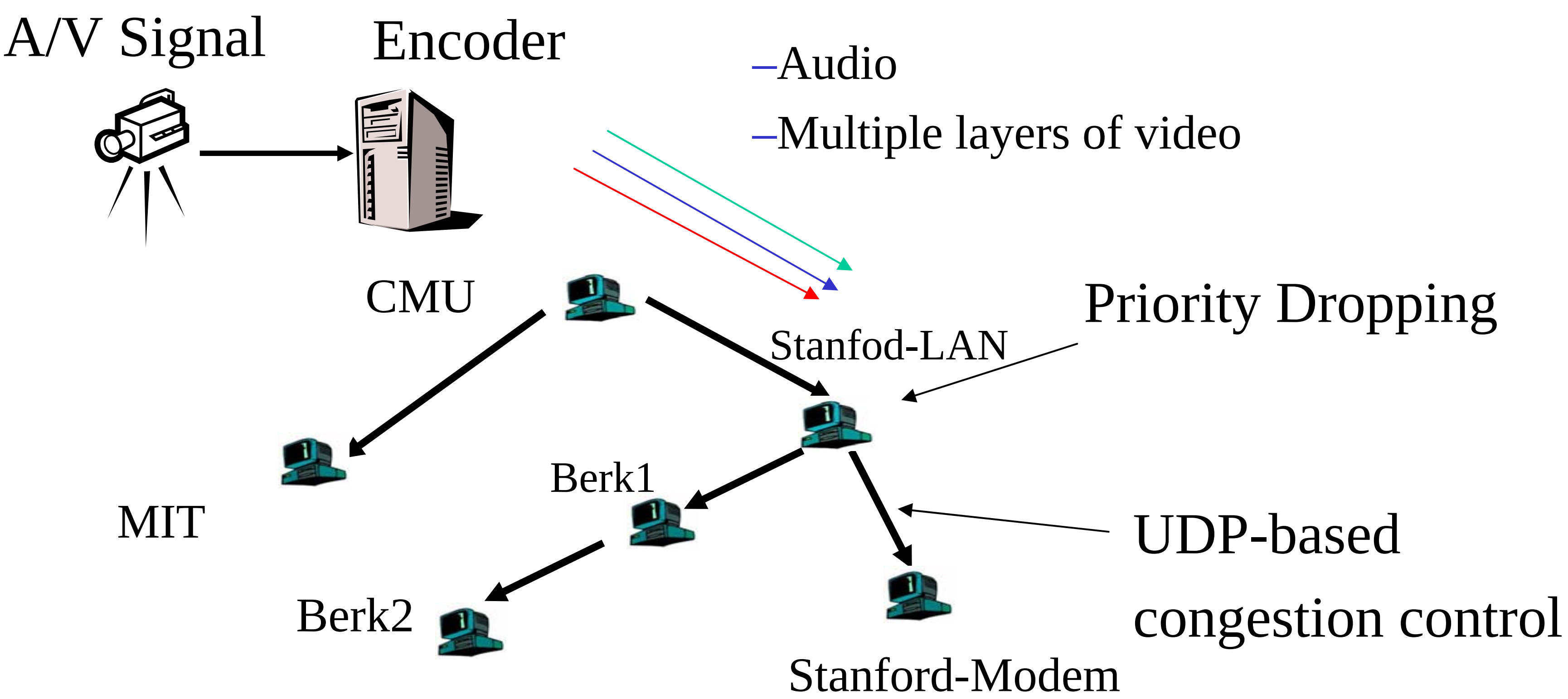








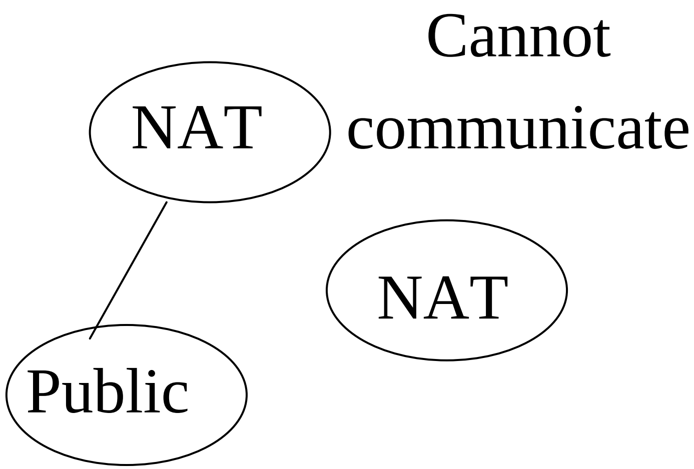
Support for Heterogeneous Receivers



Each receiver: receive as many layers as capacity allows

http://esm.cs.cmu.edu/

Support for NATs



- System supports NATs as children
- Allows NATs to be parents of public hosts
- Public hosts can be parents of all hosts

http://esm.cs.cmu.edu/

Deployment Experience

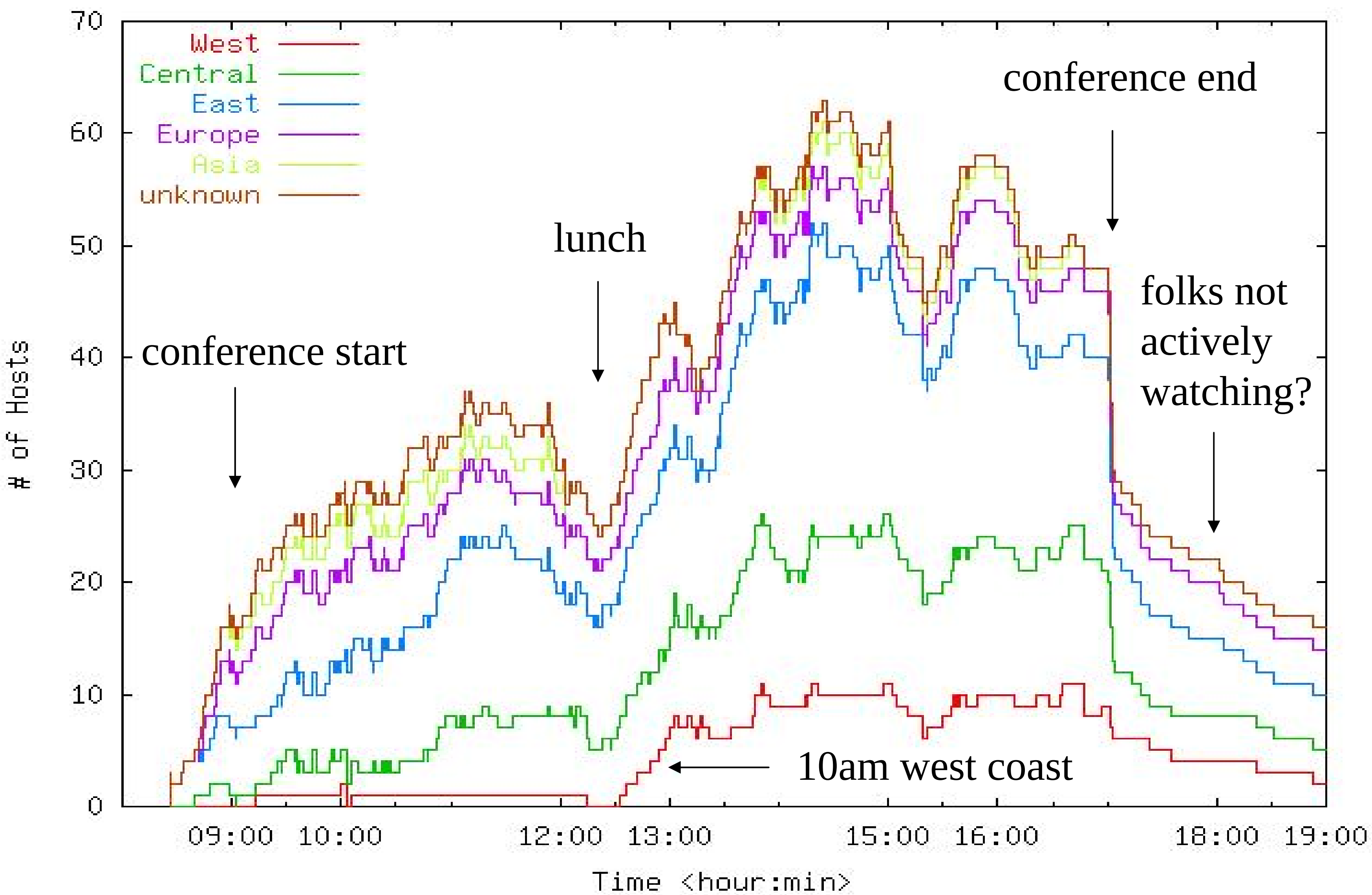
- **First broadcast in Aug ’02: Sigcomm02**
- **Total ~25 events, ~200 operational hours**
  - ~6600+ participants: across 5 continents
  - in home, academic and commercial environments
  - behind various technologies (DSL/cable modem, wireless, T1, T3, Ethernet) and NAT/Firewall.
- **Ease of Use:**
  - Viewer: 2 or 3 Clicks, Download & install software: seconds
  - Publisher: Audio/video/computer equipments: ~ 0.5 -- 3 hours. (depending on the environment and quality requirement)

Major Event Highlight

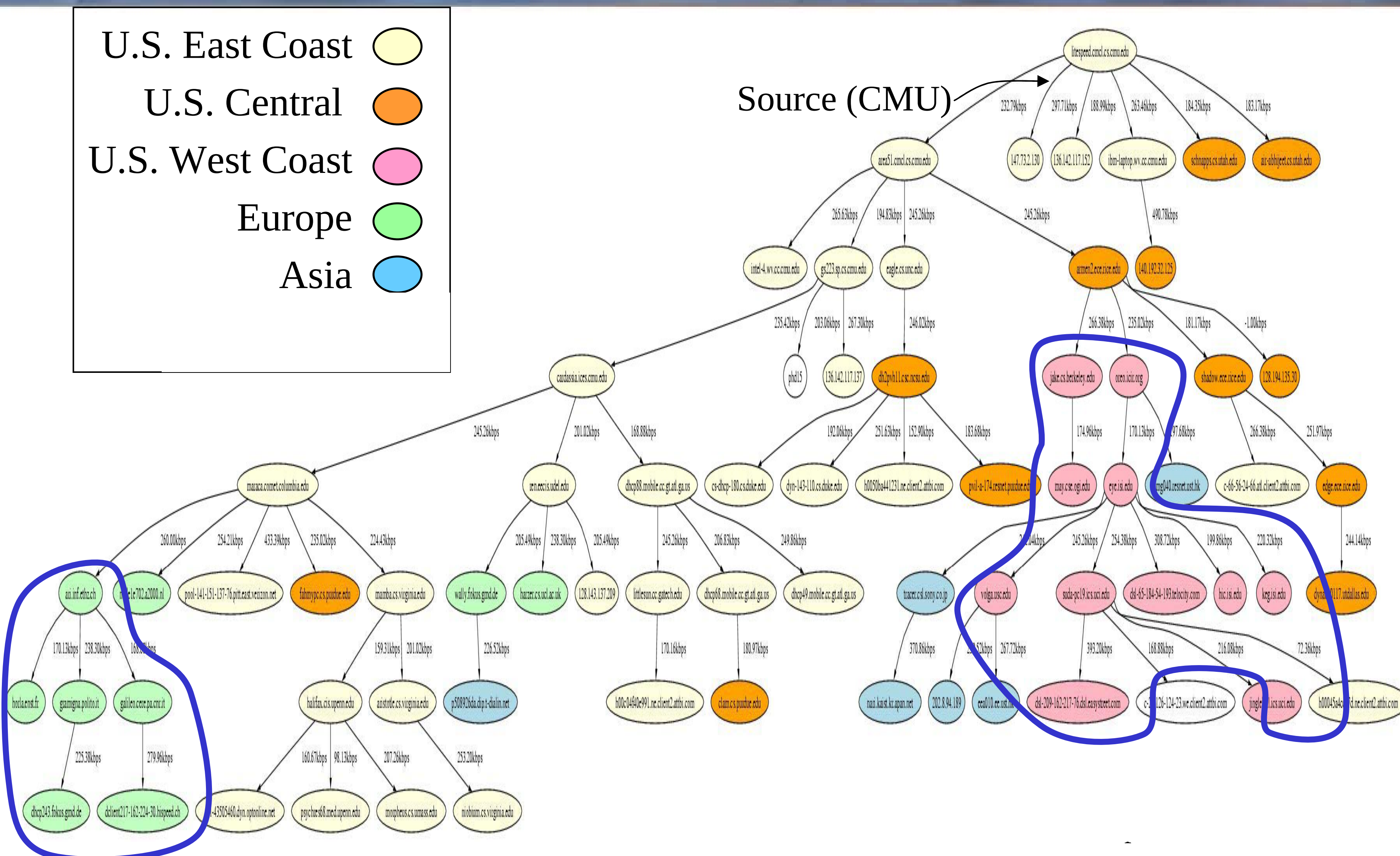
Event	Duration (hours)	Unique Hosts	Peak Size
SIGCOMM ’02	25	338	83
SIGCOMM ’03	72	705	101
SOSP’03	24	401	56
DISC’03	16	30	20
Distinguished Lectures	11	400	80
AID Meeting	14	43	14
Buggy Race	24	85	44
Slashdot	24	1609	160
Grand Challenge	6	2005	280

http://esm.cs.cmu.edu/

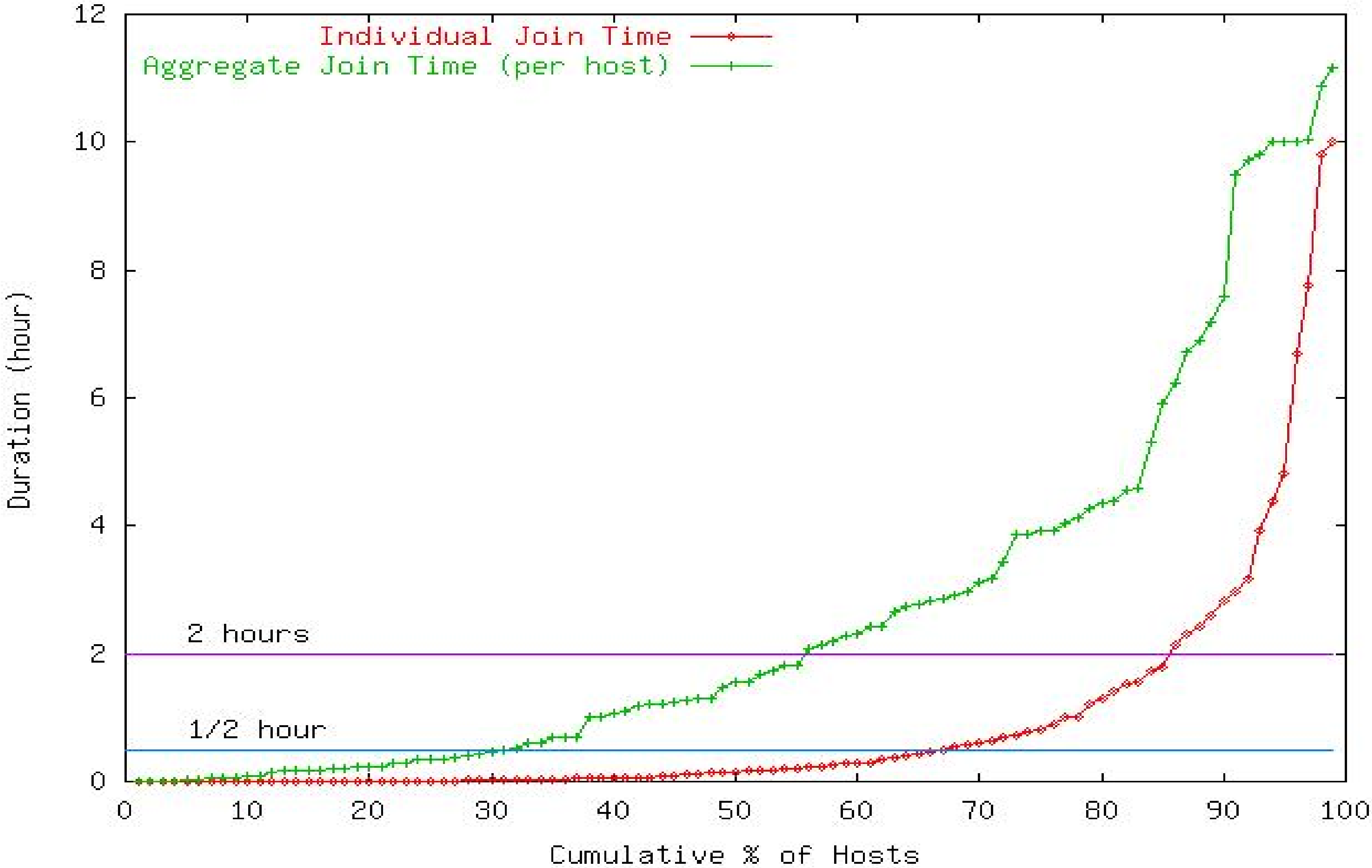
Group Dynamics



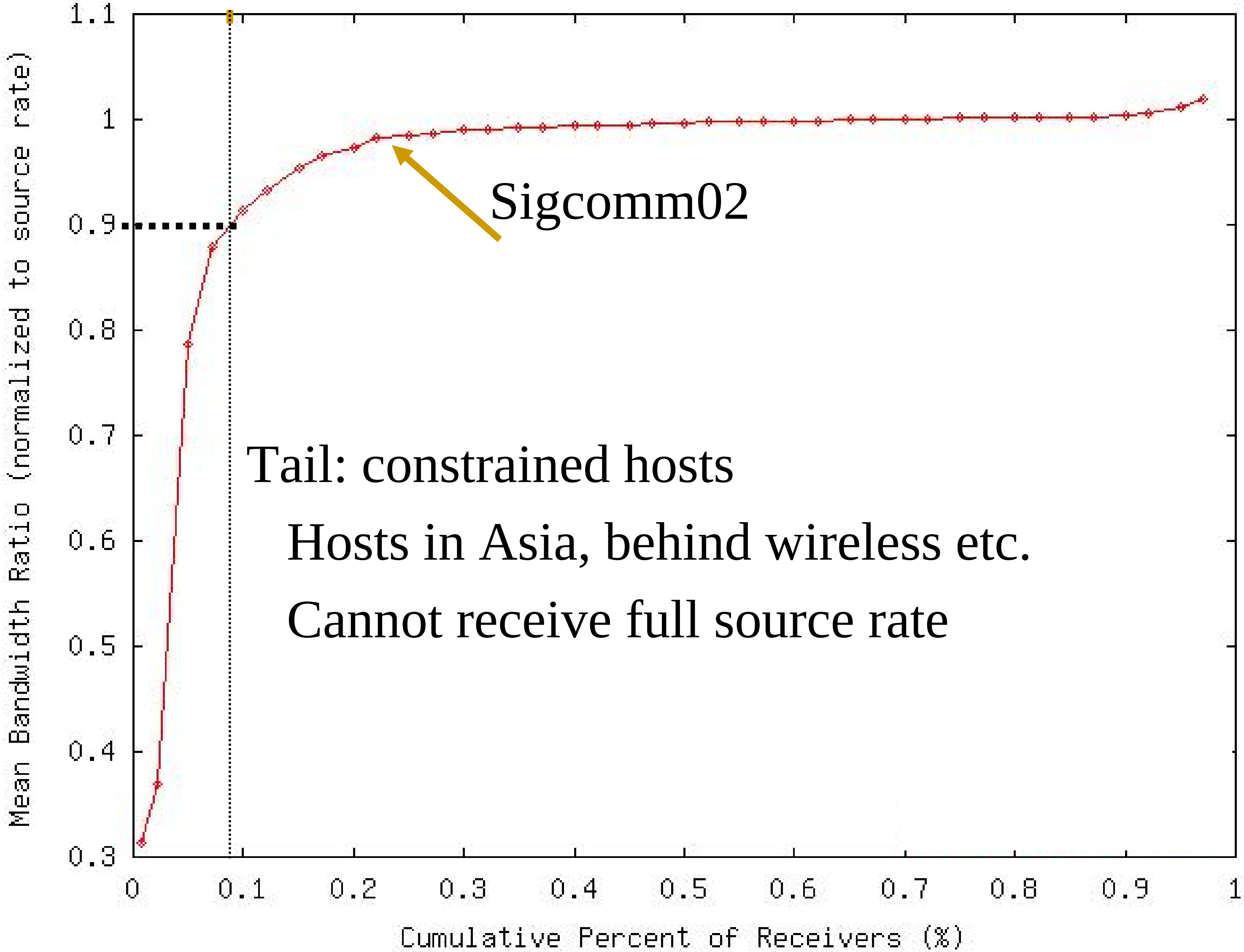
Overlay Tree at 2:09 pm



Duration of Participation

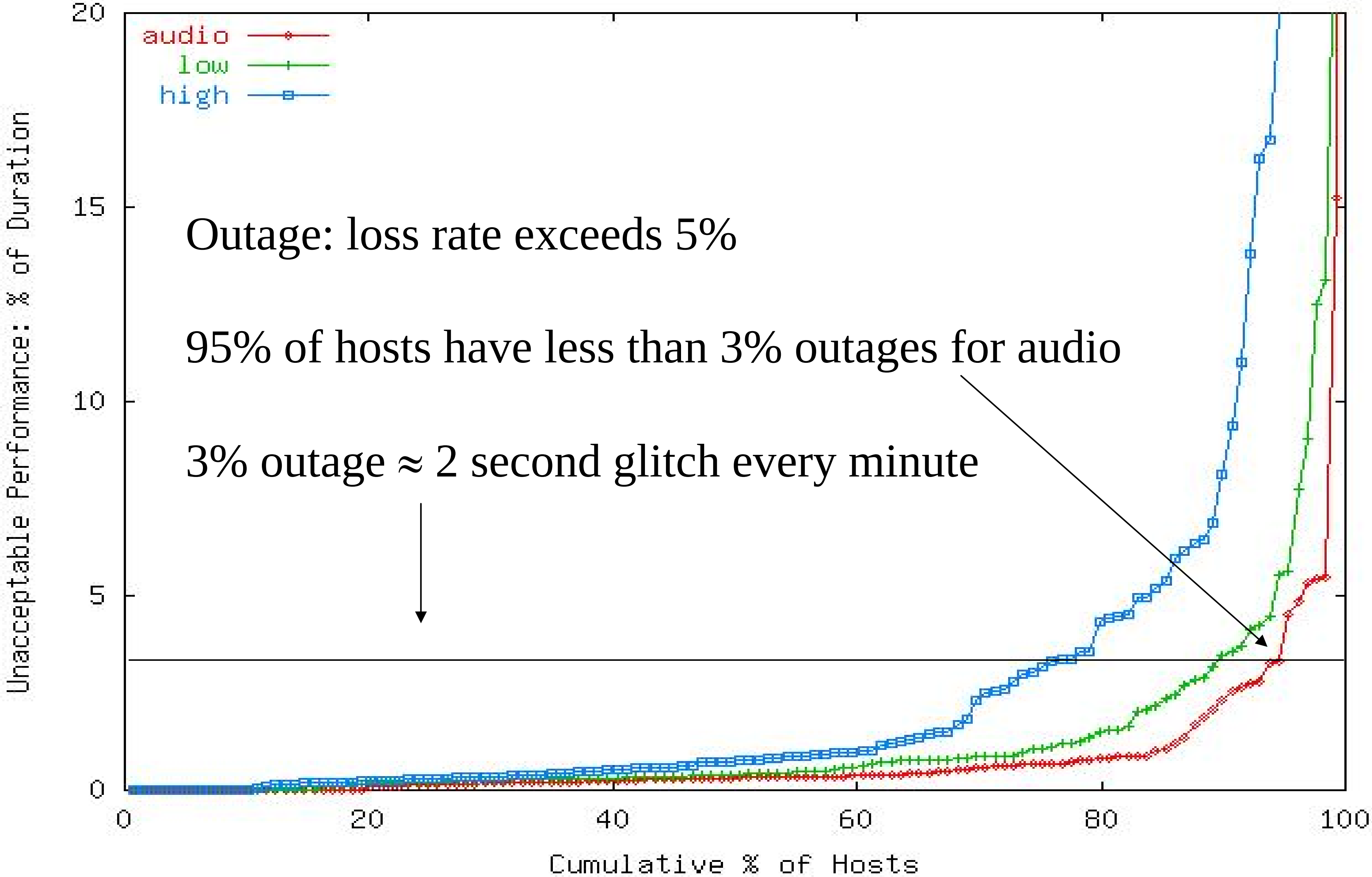


Receiver Bandwidth



u.edu/

Transient Performance: Outages

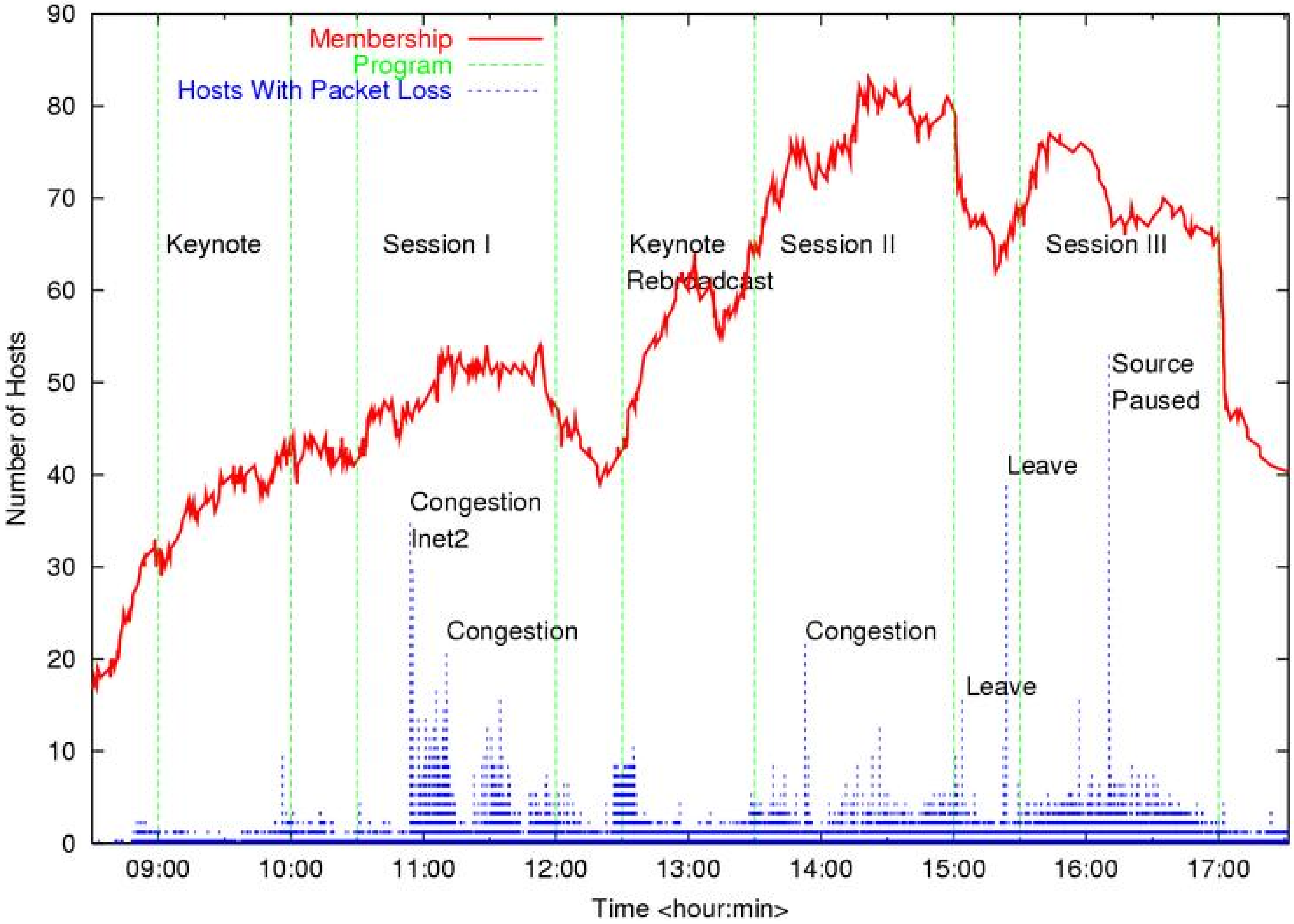


Outage: loss rate exceeds 5%

95% of hosts have less than 3% outages for audio

3% outage ≈ 2 second glitch every minute

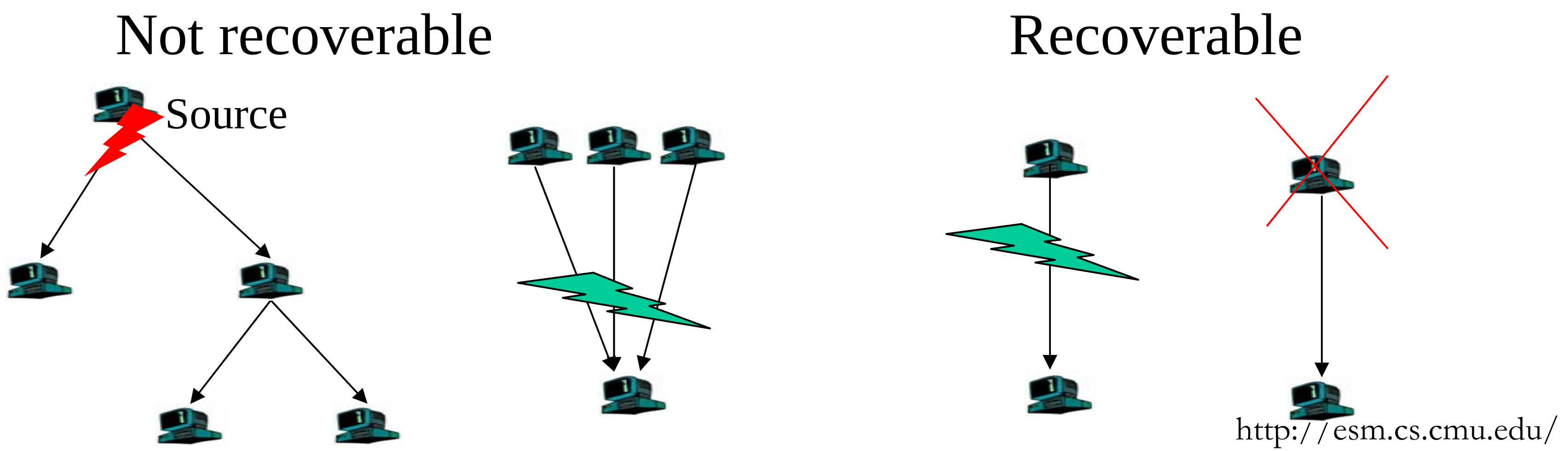
System Dynamics



ıu.edu/

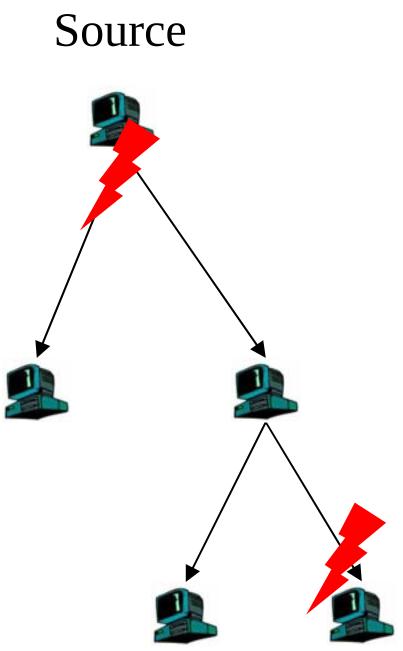
Loss Diagnosis

- **Not all losses recoverable**
  - Congestion near source
  - Constrained host, or congestion near host
- **51% of loss events : not recoverable**
  - Explains the tail



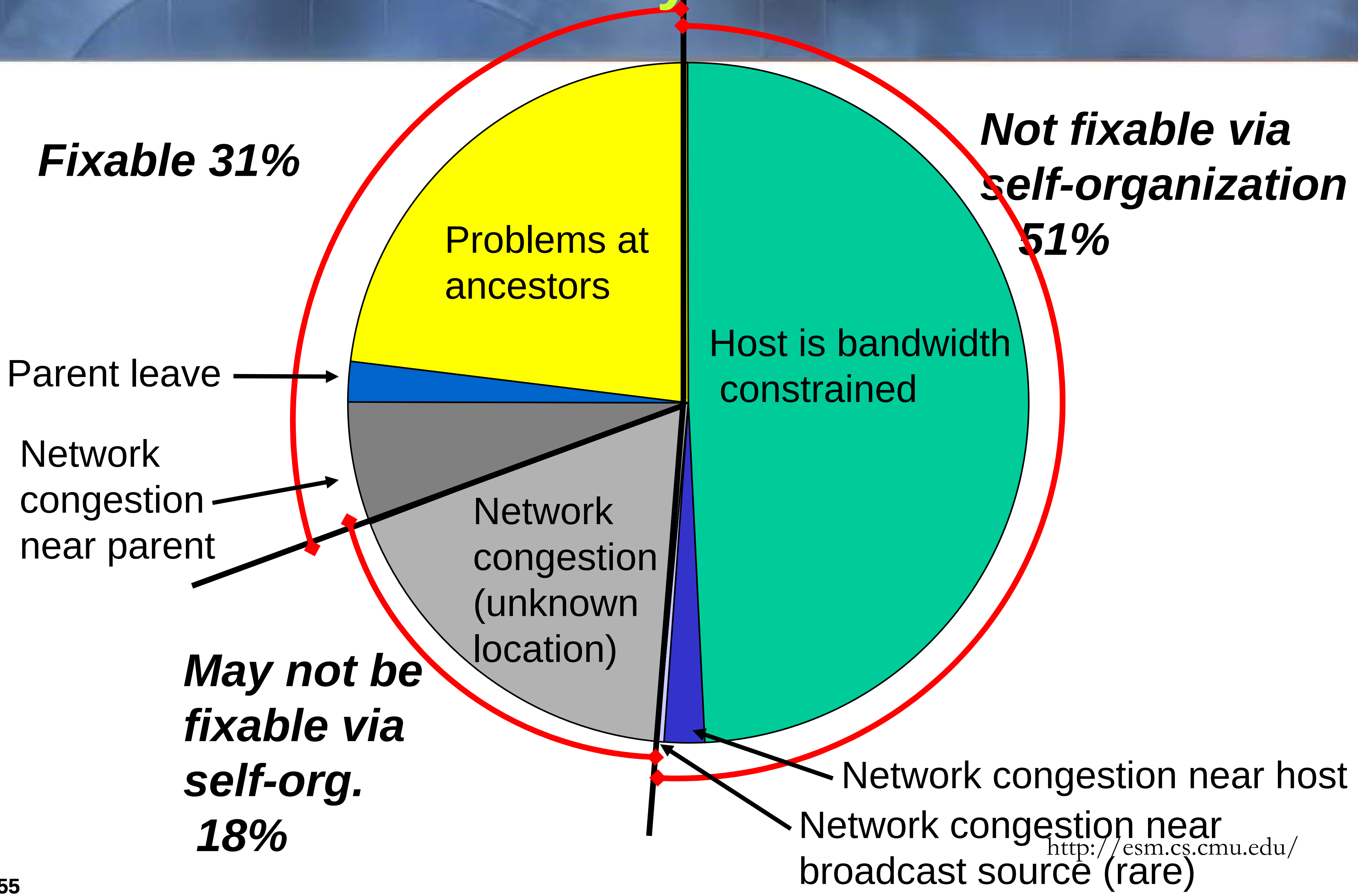
Loss Diagnosis

- **Loss event: any packet loss in 5-second interval**
- **Loss "not recoverable" (51%)**
  - Constrained hosts (49%)
  - Local congestion (2%)
  - Congestion near source (rare)
- **Loss potentially recoverable (31%)**
  - Loss at parent / ancestor
  - Congestion near parent
  - Parent leave
- **Loss not categorized (18%)**



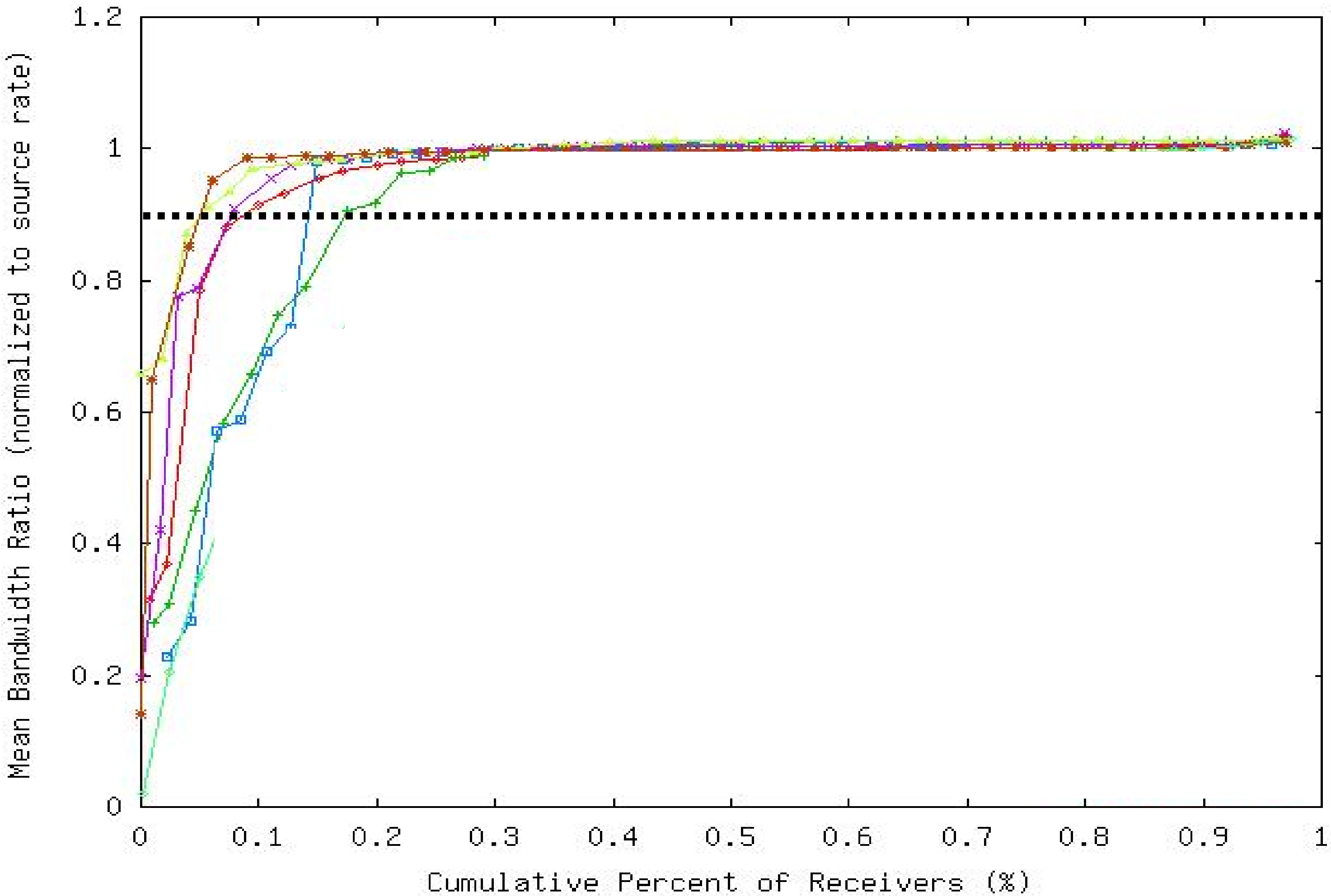
http://esm.cs.cmu.edu/

Loss Analysis Result



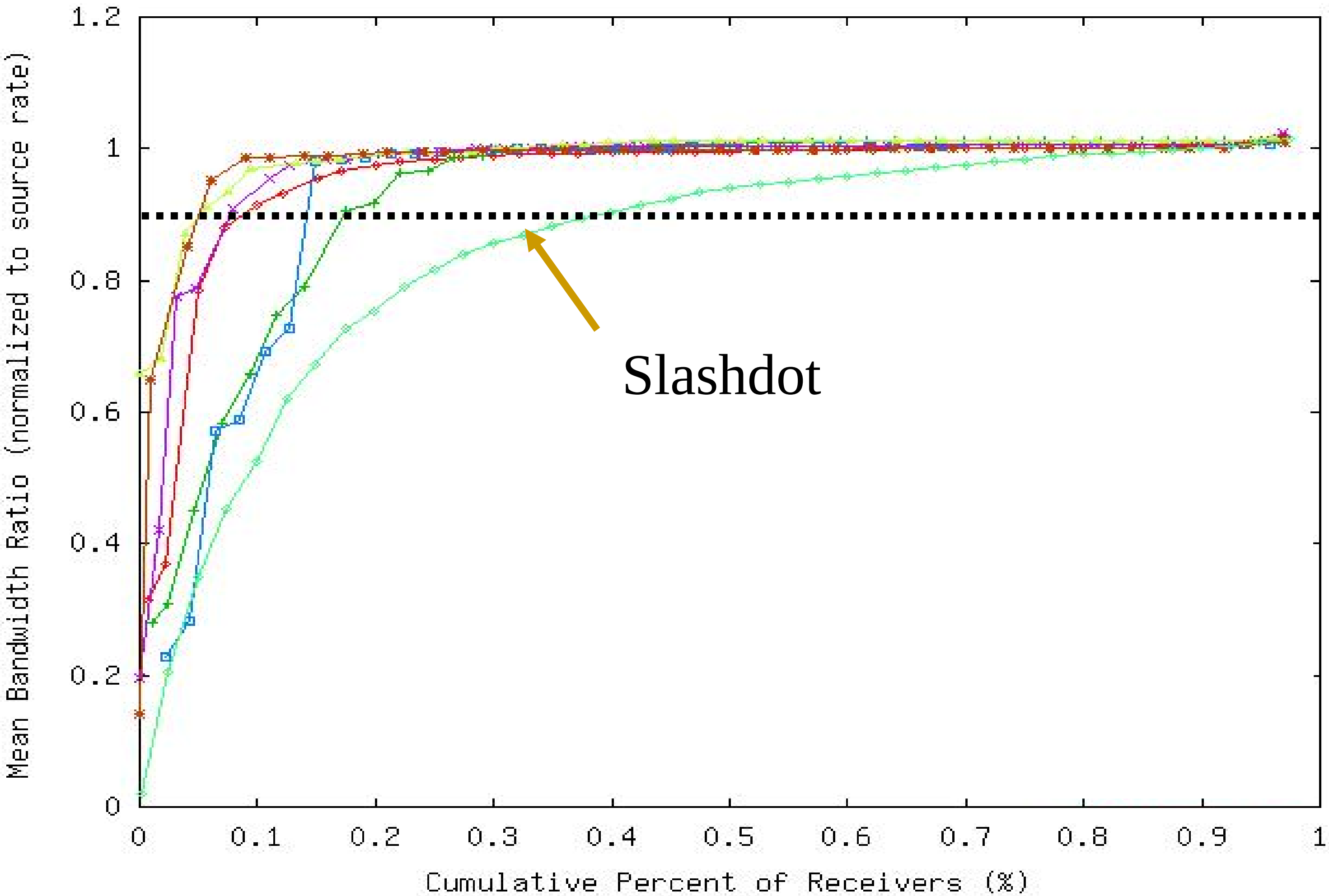
http://esm.cs.cmu.edu/

Performance: Multiple Broadcasts



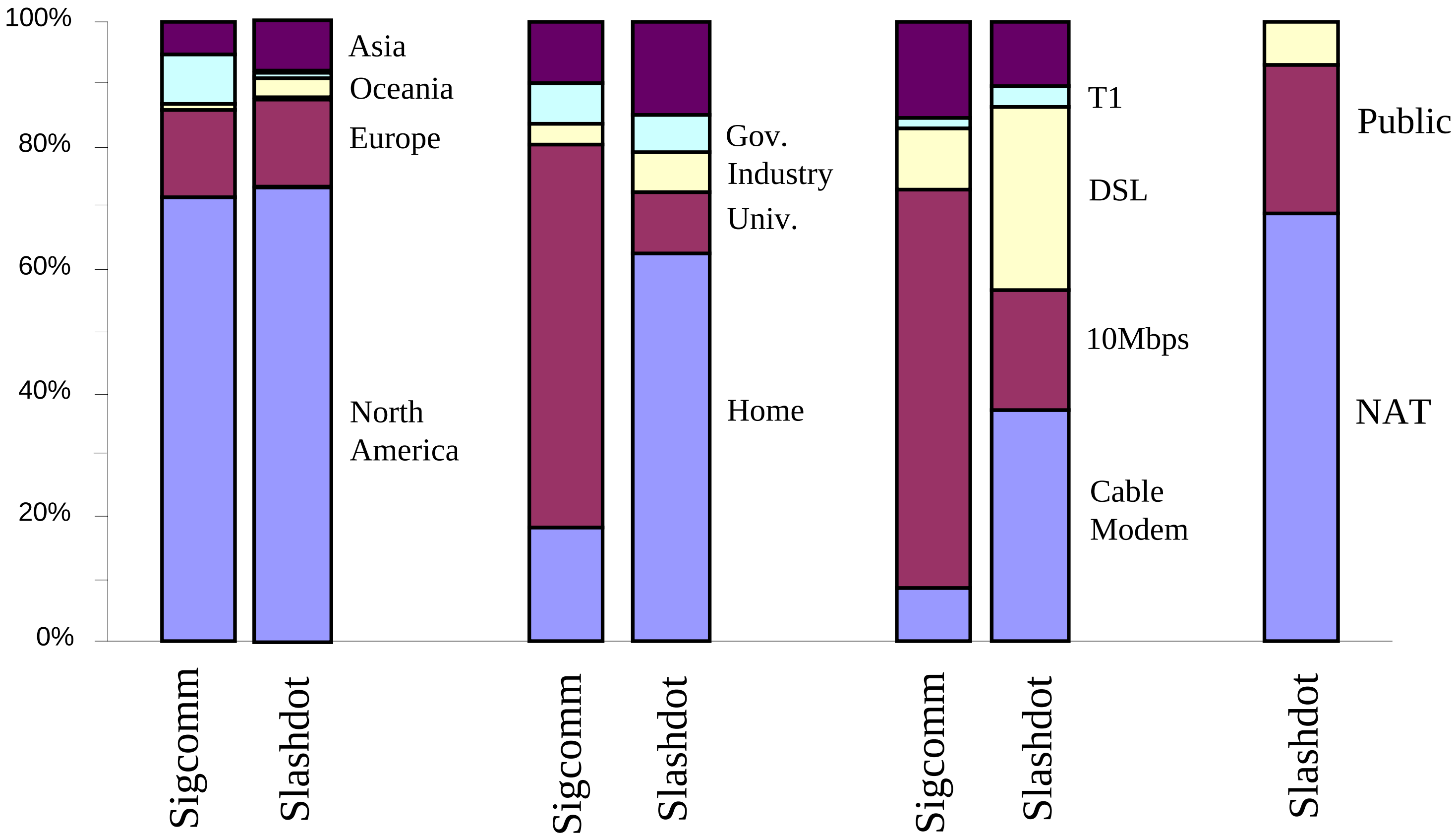
cmu.edu/

Performance: Multiple Broadcasts



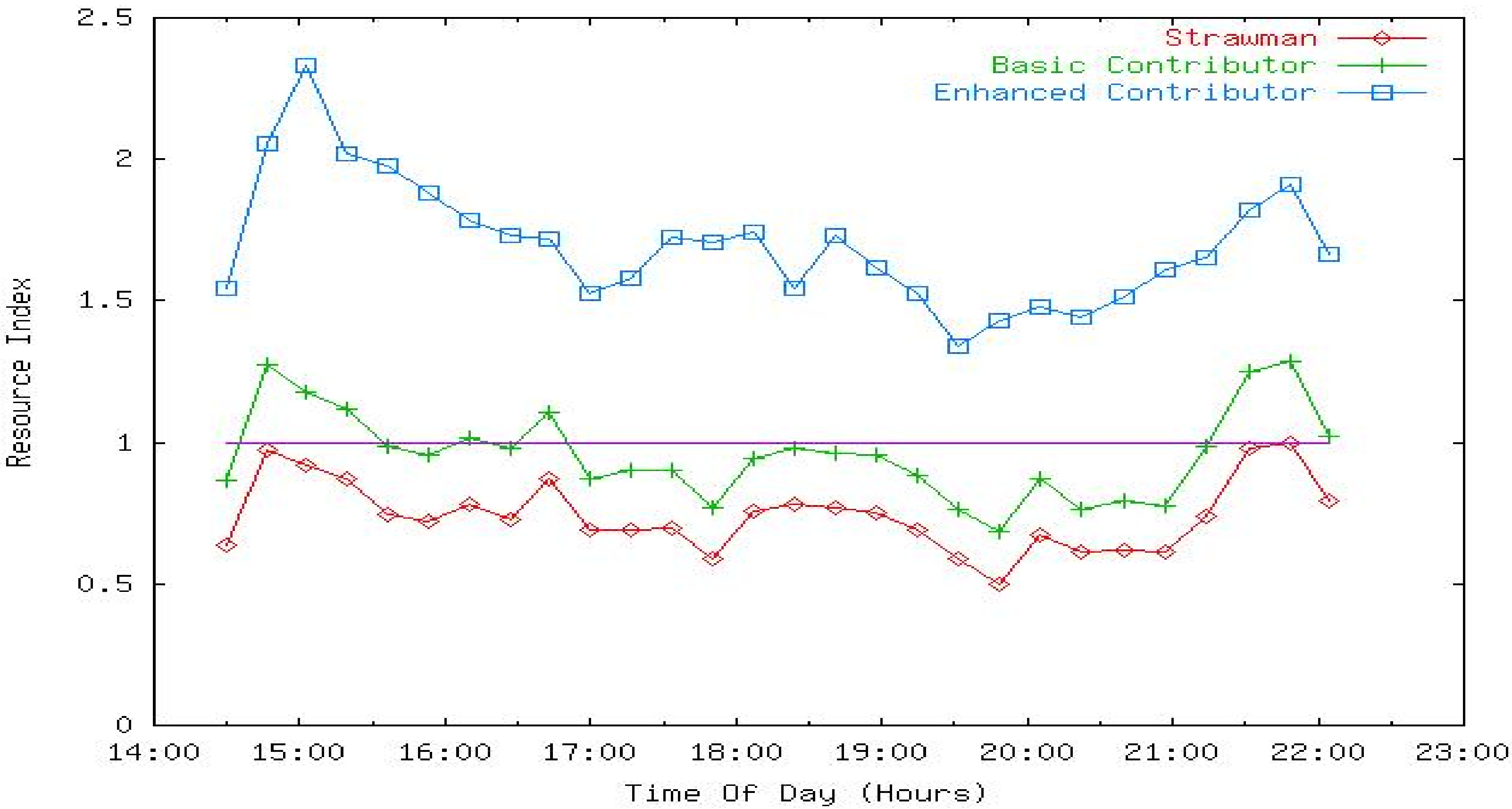
mu.edu/

Diversity of Hosts



http://esm.cs.cmu.edu/

Coping with NATS



http://esm.cs.cmu.edu/

Where We Stand

- **ESM deployment**
  - Extremely easy to deploy
  - Zero effort Internet broadcast achievable

Ongoing Research: Scalability

- **What about large groups?**
  - Same or different problems as small-scale?
  - Chicken and egg problem
- **Issues with large scale**
  - Enough forwarding resource?
  - Rapid joins/leaves?
- **Approach**
  - Trace-driven simulation with Akamai data
  - Evaluate intrinsic resource availability and stability
  - Initial results promising

Ongoing Research: Incentive

- **Why would a host contribute more than it receives?**
  - Bit-by-bit scheme will leave up to 80% hosts unserved
  - Needs to create incentive for resource-rich hosts to contribute
- **Key observations**
  - Asymmetry role of publisher and subscribers
  - Publisher has incentive to maximize social welfare
  - Publisher leverage multiple video quality levels to create incentives for subscribers
  - Apply the theory of taxation

Ongoing Research: On-Line Community

- **We observe that some people like Internet broadcast better than lecture hall**
- **Can we make Internet participation a unique experience?**
  - More than just a sub-optimal imitation of the physical experience
  - Leverage on the strength of virtual presence offered by the Internet

Related Work

- **Yoid: architecture contribution, independently conceived**
- **Follow-up overlay multicast protocols**
  - Reducing group management overhead for larger group size
    - NICE, Overcast, HMTP, CAN,Bayuex,Delaunay,Scribe …
  - Redundancy in data delivery
    - Coopnet, Splitstream, Bullet
- **ESM Contributions**
  - First to argue for architectural alternative
  - Evaluation framework: RDP, stress
  - Systems approach
  - “Father” of P2P Streaming

http://esm.cs.cmu.edu/

Other Overlay Systems

- **MBONE, RON, Planetlab**
  - Infrastructure
  - Mainly used by network researchers
- **ESM**
  - Infrastructure-less
  - Instantaneously deployable
  - Application that targets common Internet users

http://esm.cs.cmu.edu/

Other Broadcasting Systems

- **Mbone/IP Multicast Based**
  - Vic/Vat
- **Infrastructure-Centric**
  - Akamai/Real Broadcasting
- **Recent commercial peer-to-peer systems:**
  - Allcast, Chaincast, Streamer, Peercast

http://esm.cs.cmu.edu/

Summary

- **Division of functionalities between end system and network**
  - One of the most important network architecture decision
- **IP Multicast is the wrong path**
  - Intractable technical challenges remain
  - Wrong direction to channel energy on
- **End System Multicast supports all multicast related functionalities in end system**
  - Scalable, deployable, easy to support higher level functionalities
  - Can be designed to be efficient also
- **Application centric approach achieves multiple goals**
  - Validate internet-scale systems with real users/workload
  - Valuable tool for ordinary users
  - Valuable tool for researchers

http://esm.cs.cmu.edu/