A Novel Framework for Measuring and Assessing Streaming Media Quality

Amy Csizmar Dalal

Department of Mathematics and Computer Science
Carleton College

April 22, 2005

Outline

- Reminder of the problem space
- Definition of “quality”
- Defining the data of interest
- Measurement methodology
- Experiments and results
- What next?

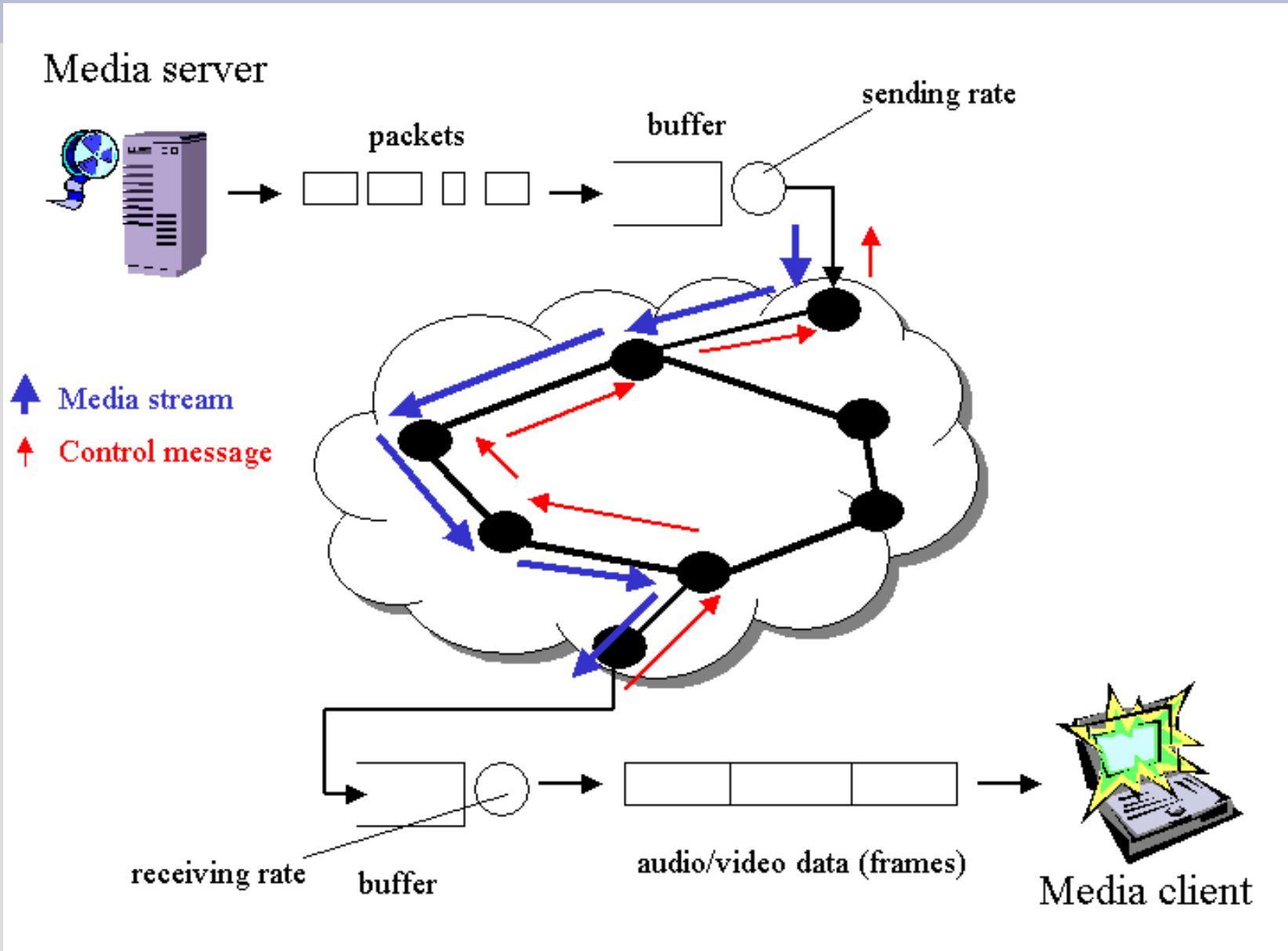
Something for everyone today

- For the computer scientists:
  - Design of the measurement tool
  - Network testbed details
  - Experimental design, execution
- For the mathematicians:
  - Statistical analysis of network, application measurements
  - Data! Numbers! Graphs!

The challenge

- We don't understand *why* streaming audio/video quality is bad
- We don't understand *when* streaming audio/video quality is bad
- We currently don't have a {good, fast, accurate, comprehensive} way to analyze the quality of audio and video streams

A typical audio/video stream



Challenges in streaming audio and video over the Internet

- Size of audio/video files
  - requires high-bandwidth links
- Duration of audio/video streams
  - longer streams are more susceptible to random loss events
- Delay intolerance
  - tight delivery deadlines = minimal *error correction*

Defining “quality”

What is “quality”?

- In this context, “quality” means the perception a user has of how good or bad an audio/video stream is

Quality categories (rankings)

- “Good”: user has few or no complaints about the video or audio
- “Acceptable”: user has some minor to moderate complaints about the video or audio, but will keep watching it anyway
- “Poor”: user has significant complaints about the video or audio, and will cease watching

Our goals

- Figure out why and when the user-perceived quality of a media stream is “bad”
  - “bad” = bad enough so that the user will go away or stop watching the stream
  - do this **without asking the user directly**!
- Use this information to predict when the quality of a media stream will deteriorate

Our approach

- Use measurements that mimic the user's experience with the stream
  - take measurements **directly from the media player**
  - combine these with our knowledge of network conditions (loss, delay) during the stream

Q: What should we measure?

Methodology

- Identify what we'd like to measure objectively
- Identify what we can measure objectively
- Software prototype
- Measurement infrastructure
- Experimental validation
- Look at harder problems

What information can we get from a media player?

- Application-layer packet information
  - # received, # lost, # retransmitted
- Frame rate
- Bandwidth (throughput, actually)
  - number of bits arriving per second
- Is the player currently “buffering”?
- Buffering percentage
- “Quality” score (time-averaged ratio of lost packets to received packets)

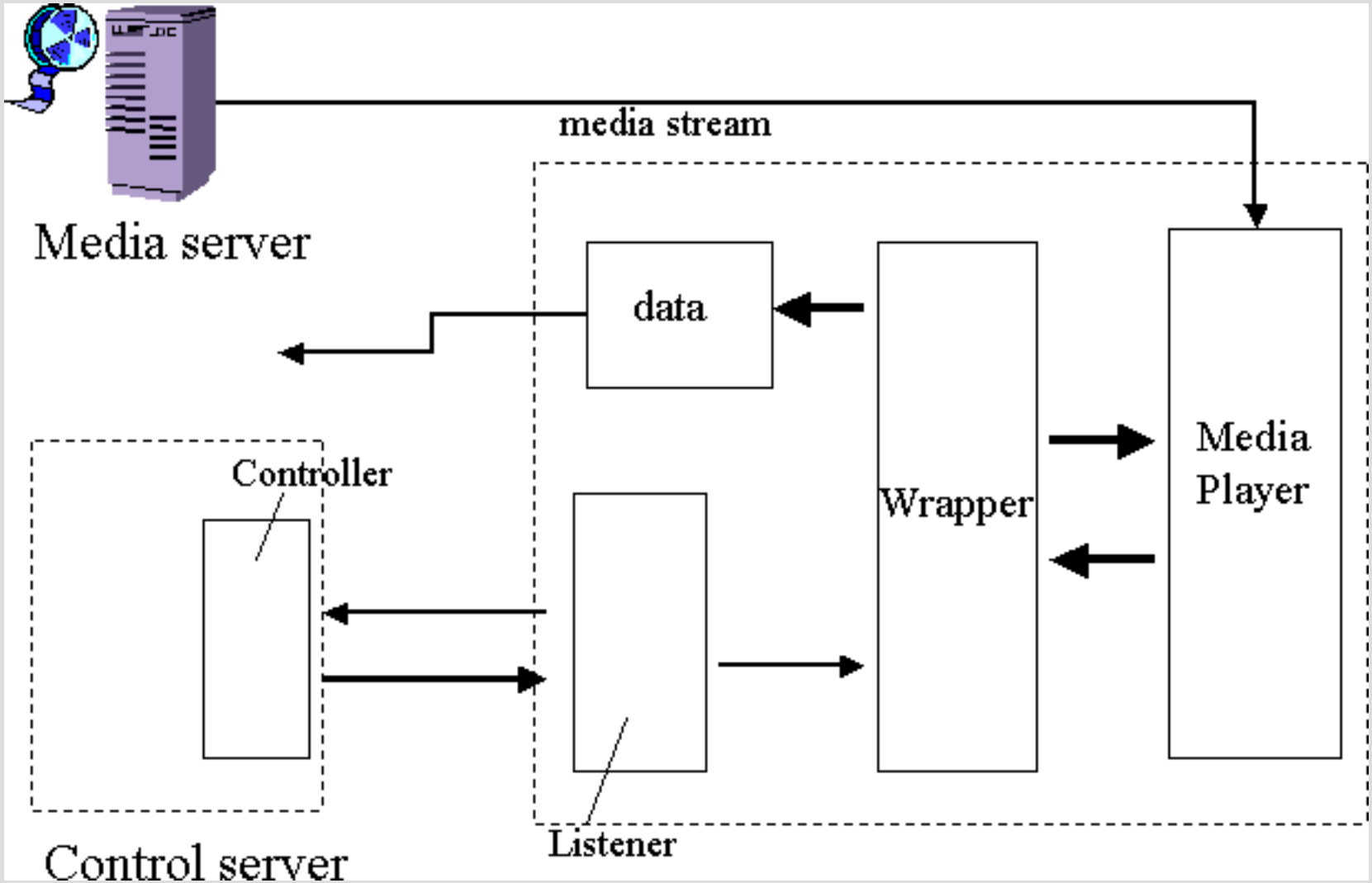
Leading vs. lagging quality indicators

- “Leading indicators” = metrics whose values are the first to change in response to network congestion
- “Lagging indicators” = metrics whose values change in response to degraded stream quality
- “Instantaneous indicators” = metrics whose values change instantaneously with degraded stream quality

Metrics of interest

- Leading indicators
  - Transmission failures
  - Retransmitted packets
  - (Startup buffering period duration)
- Instantaneous indicators
  - Buffer starvation periods
  - Lost packets

Objective measurement framework



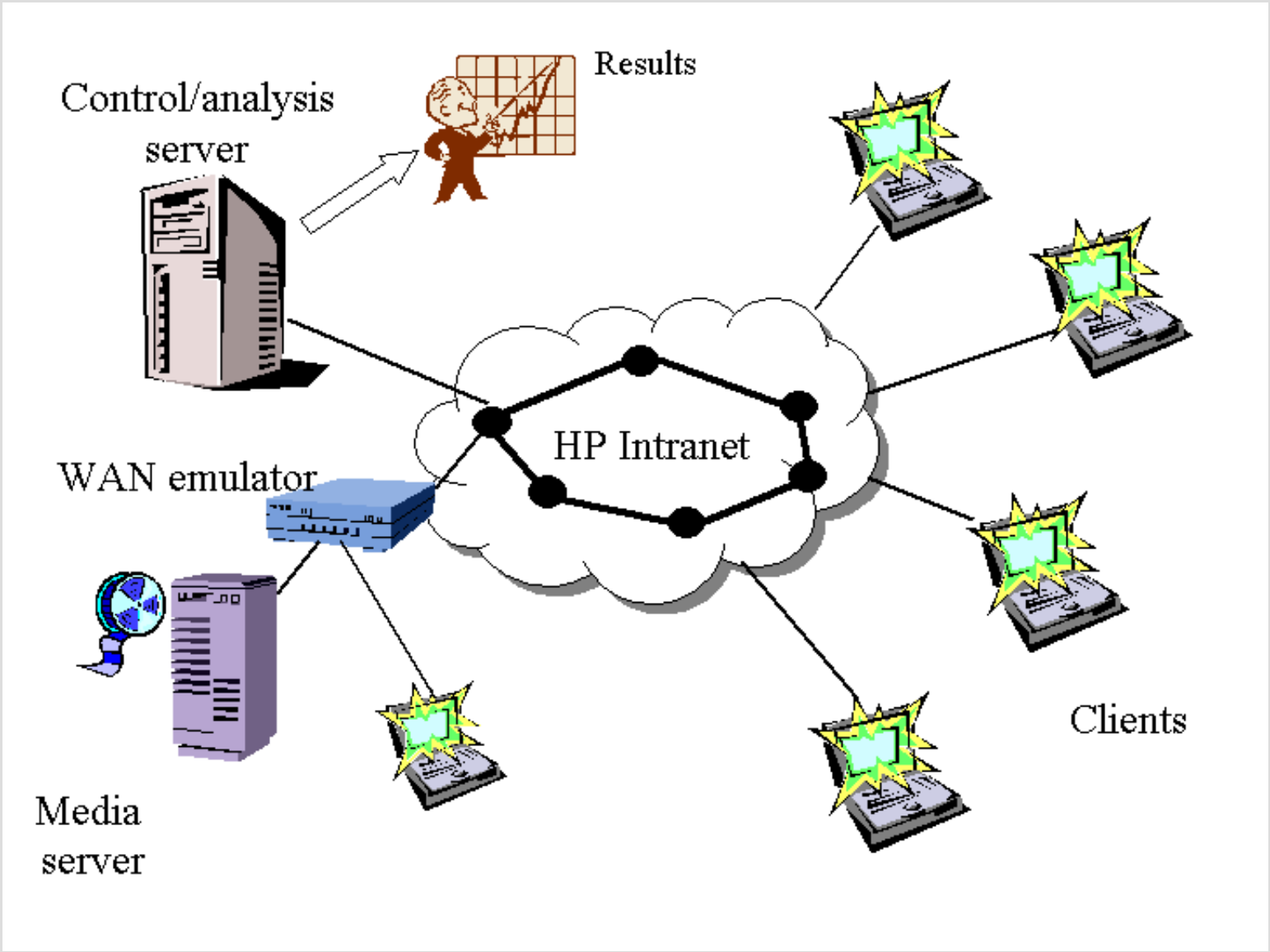
SQATool

Some experiments and (interesting) results

Validation

- First step: Analytical validation
  - send streams under controlled (known) network conditions
  - match trends in player-derived data with network “snapshots”
- Second step: Subjective validation
  - do these objective measurements accurately reflect the user's view of stream quality?

Analytical validation: network testbed

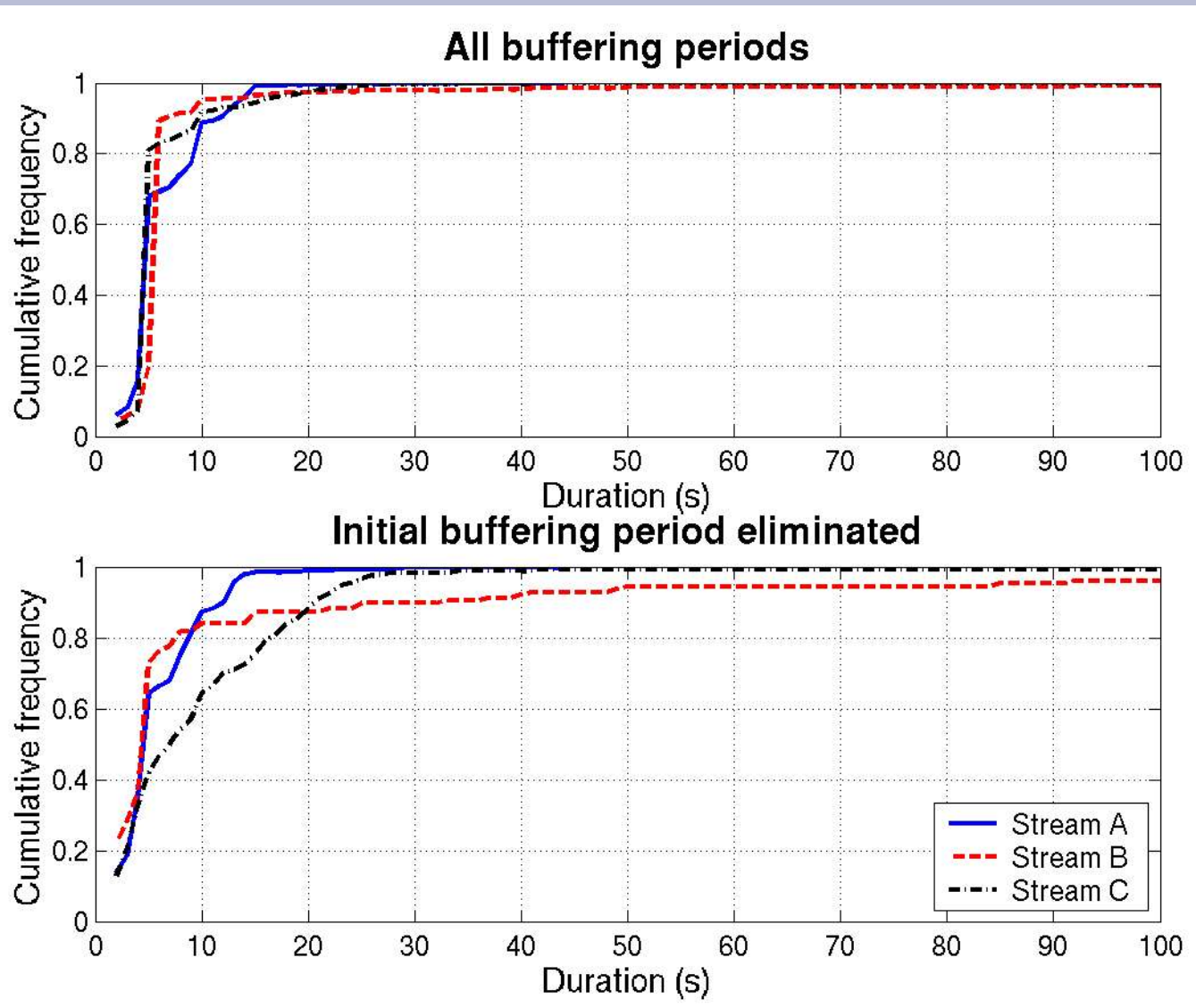


Analytical experiments (HPL)

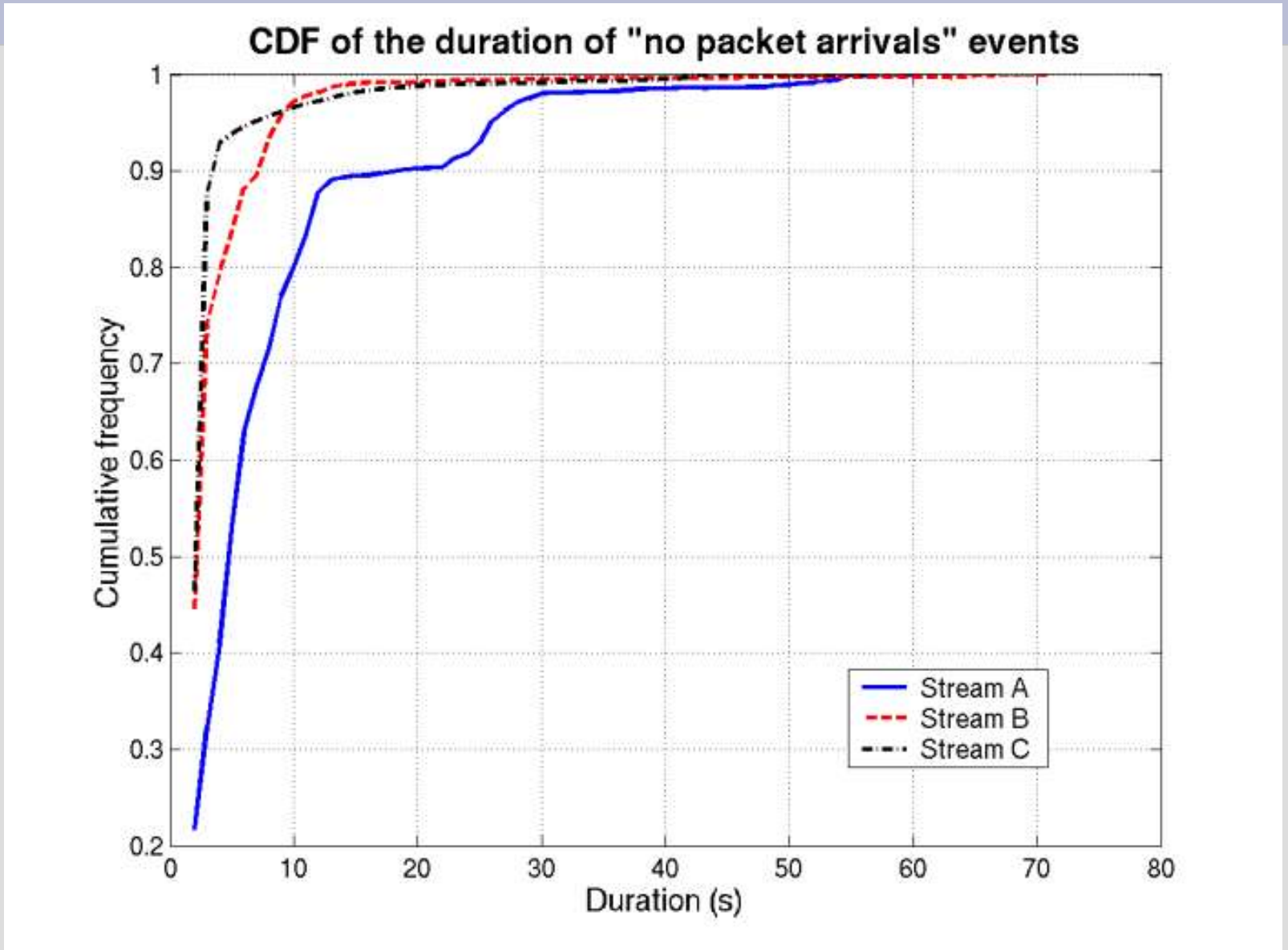
- Introduced {1, 5, 10}% packet loss on testbed network (server --> clients)
- 12 clients, geographically distributed throughout US
- September-December, 2001

Stream	Duration	Bandwidth (kbps)	Description
A	2:04	458	Animated movie trailer; high action
B	11:26	107	CEO address; low action
C	30:00	84	Technical talk; low-moderate action

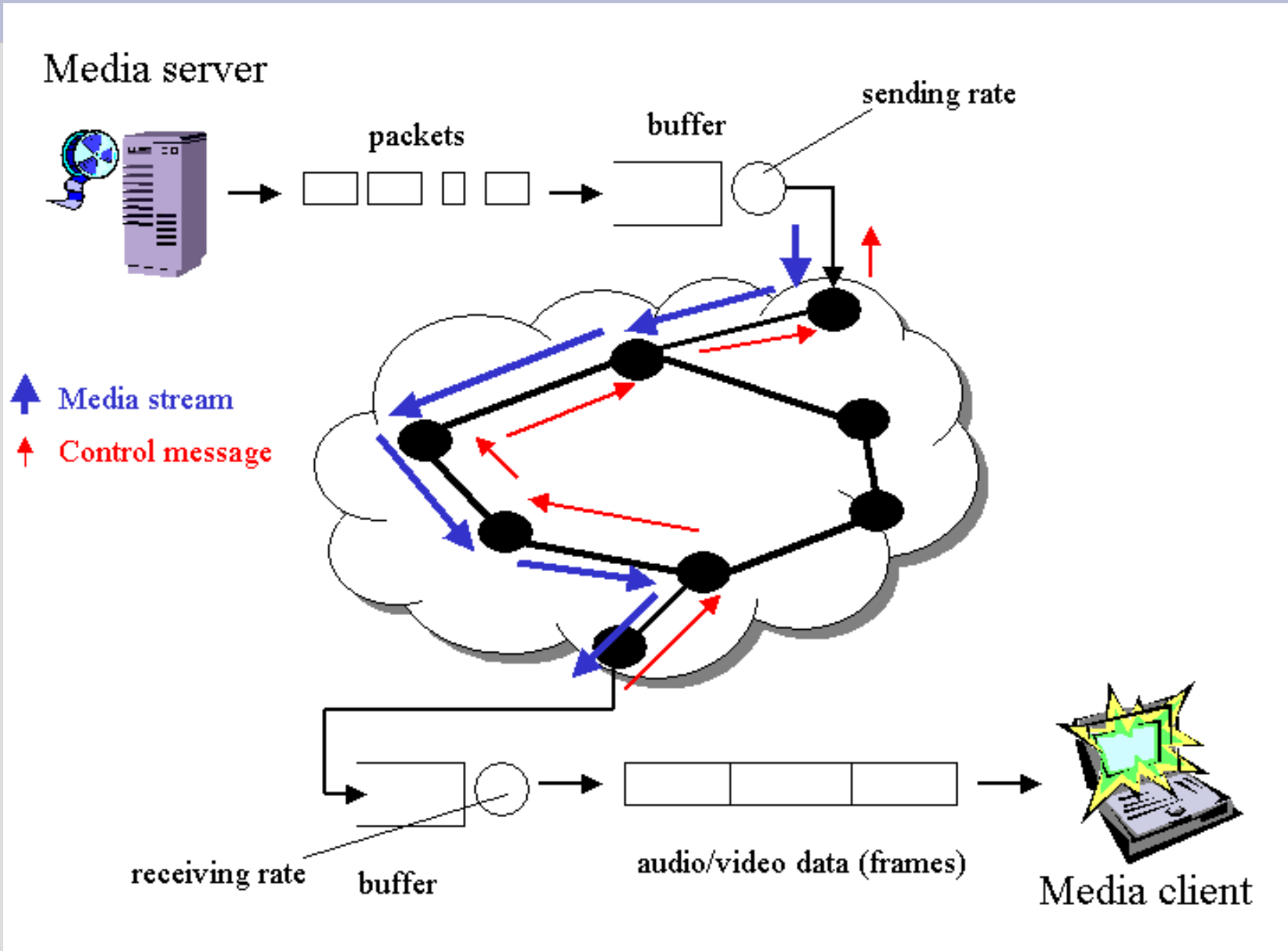
Buffer starvation periods



Transmission failures



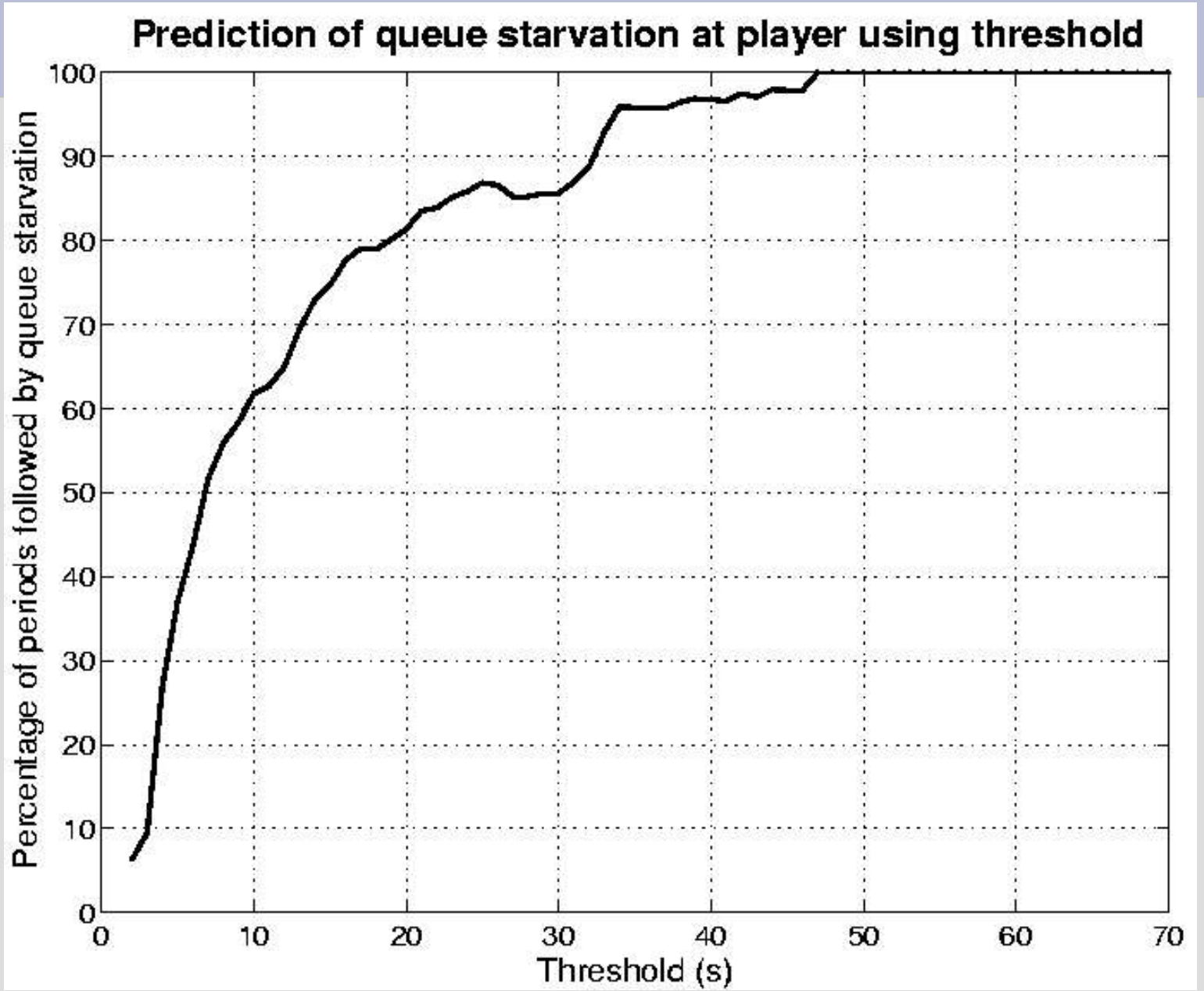
A typical audio/video stream



Prediction using purely objective metrics?

- Idea: Can we predict future periods of buffer starvation simply by observing the packet arrival process?
- Method:
  - Identify all periods of transmission failure in each stream
  - Identify all periods of buffer starvation in each stream
  - If **t(starve) – t(fail) <** eps, then we say buffer starvation **occurs as a result of** the transmission failure event

Prediction of buffer starvation



Subjective quality validation

- Users watch streams under “perfect” and “lossy” conditions, rate them
- Ratings:
  - 7-point scale (1 = poor, 7 = exceptional)
  - Additional comments (optional)
- Take measurements from media player simultaneously

User survey

http://www.mathcs.carleton.edu/faculty/adalal/research/testing/survey0.html

Challenges in experiment design

- Picking the correct loss rate for a given quality level
  - Different streams react to network loss differently
  - Windows Media Server does some “goofy” things at startup
- Ensuring consistency in quality level
  - Scripts to start/stop packet loss, start/stop measurement tool
- Survey design
  - Help from the Psychology department

Experimental info

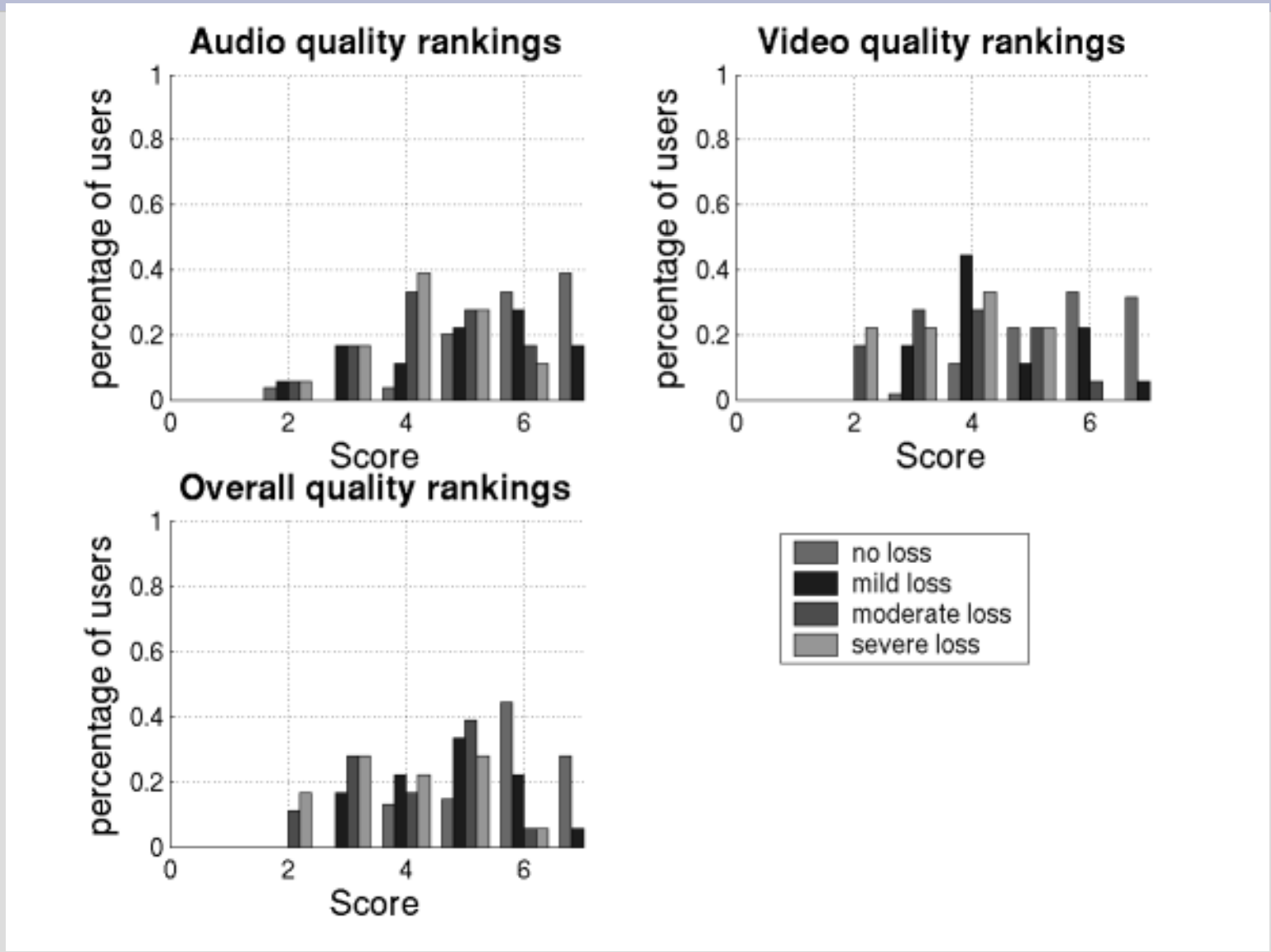
Stream information			Network congestion level (delay, loss)		
Name	**Time**	**Description**	**Mild**	**Moderate**	**Severe**
Ad	0:30	TV ad; moderate action	(50ms, 4%)	(80ms, 16%)	(175ms, 25%)
News	4:09	News clip; low-to-moderate action	(50ms, 4%)	(60ms, 11%)	(80ms, 25%)
Trailer	2:22	Animated movie trailer; high action	(100ms, 8%)	(200ms, 10%)	(200ms, 19%)

- {August, October} 2004
- 17 participants total
- All participants provided comments on each of the streams

What are we looking for?

- Relationship between MOS and % network packet loss
- Correlation between MOS and metrics of interest
  - > -0.40: “significant”

User quality rankings, all streams



Correlation coefficients, TV ad

	Correlations with MOS:		
Metric	**Audio**	**Video**	**Overall**
TX failure	-0.49	-0.70	-0.65
Lost packets	-0.44	-0.70	-0.68
Retransmissions	-0.59	-0.75	-0.70
Startup buffer period	-0.44	-0.57	-0.50

Correlation coefficients, news clip

	Correlations with MOS:		
Metric	**Audio**	**Video**	**Overall**
TX failure	-0.22	-0.41	-0.33
Lost packets	-0.30	-0.55	-0.49
Retransmissions	-0.39	-0.60	-0.56
Startup buffer period	-0.26	-0.46	-0.45

Correlation coefficients, trailer

	Correlations with MOS:		
Metric	**Audio**	**Video**	**Overall**
TX failure	-0.38	-0.36	-0.42
Lost packets	-0.57	-0.47	-0.52
Retransmissions	-0.60	-0.50	-0.53
Startup buffer period	-0.49	-0.45	-0.45

Summary

What does the future hold?

Quality prediction

- Idea: Can we **predict** the future quality of a media stream based on past measurements?
- Examples:
  - predict future buffer starvation periods based on transmission failures
  - predict future “bad” ratings based on the number of retransmission requests
  - predict future packet loss based on the number of transmission requests

What could we do if we could predict degraded stream quality?

- Play out locally-stored content (ads, previews, trailers, etc.)
- Pre-cache content and play out locally when conditions are bad
- Route around the congestion
  - serve content from another server
  - find another network route
- If all else fails, tell the user to come back later

Ongoing/future work

- Quicktime port of SQATool
  - Spring 2005
- Apply data mining techniques to the collected measurements
  - Project with Dave Musicant and 2 students, Summer 2005
- Explore the prediction angle further

Summary

- Evaluating the quality of a media stream is difficult to do efficiently and effectively
- Our approach: identify metrics that “mimic” a user's responses, taken from the media player
- Analytical results: media player measurements correspond to periods of network packet loss
- Subjective results (preliminary): strong-ish correlations between user quality rankings and media player measurements

Thank you

- Ed Perry and Sujata Banerjee at HP Labs
- Keith Purrington and Ben Sowell for designing the user testing experiments and analyzing the data
- Mike Tie and his student workers
- The Carleton students and staff members who participated in the user testing
- Mija Van der Wege for assisting in the user survey design

Questions?