University of North Carolina at Chapel Hill

Adaptive, Best-Effort Congestion Control Mechanisms for Real-Time Communications on the Internet

Kevin Jeffay

F. Donelson Smith
Gunner Danneels (Intel)

Mark Parris
Michele Clark
Peter Nee

http://www.cs.unc.edu/Research/dirt

Multimedia Networking Research at UNC

- What are we doing?
  - Trying to understand how “broken” the Internet is today
  - Trying to understand how to design real-time multimedia applications for the Internet
- Why are we doing this?
  - We want to understand if we should spend our efforts building a better Internet, making “smarter” applications, or both

Multimedia Networking Research at UNC

- How are we doing this?
  - Developing real-time communications and computation middle-ware
  - Building real-time applications with experimental communications software
  - Evaluating their performance on controlled and production networks
  - Running long-term performance studies on the Internet

Adaptive, best-effort congestion control for real-time communications

Outline

- Our driving problem — realizing distributed, immersive, virtual laboratories
  - The UNC *nanoManipulator* system
- The continuous media congestion control problem
- 2-Dimensional media scaling techniques
- Experimental results for Internet videoconferencing

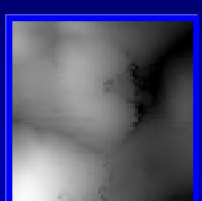
Distributed, Immersive, Virtual Laboratories

- Advanced scientific instruments have computer-based or computer-enhanced interfaces
- Treating these systems as distributed systems enables...
  - Better user interfaces
  - Remote operation of instruments
  - Multi-user and collaborative operation
  - Sharing of instruments and specialized computing equipment

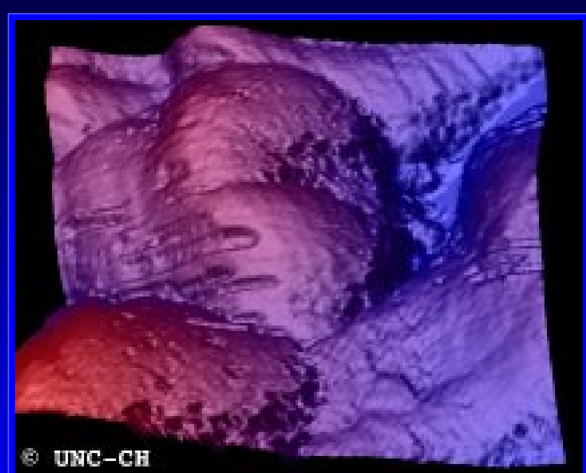
Distributed, Immersive, Virtual Laboratories

- Example — Atomic Force Microscopes

Distributed, Immersive, Virtual Laboratories



CCD Image

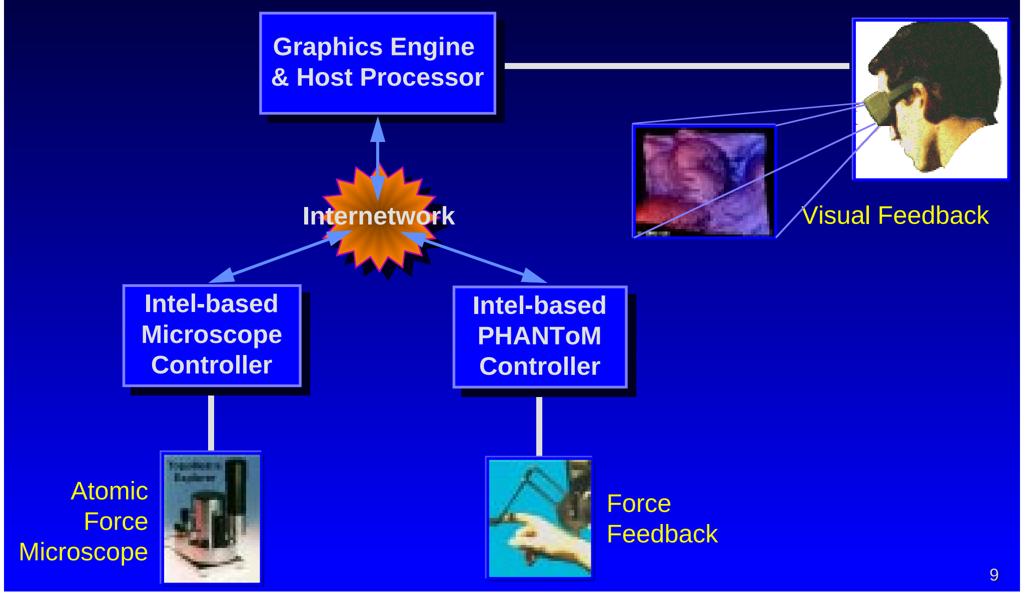


Computer Enhanced Image

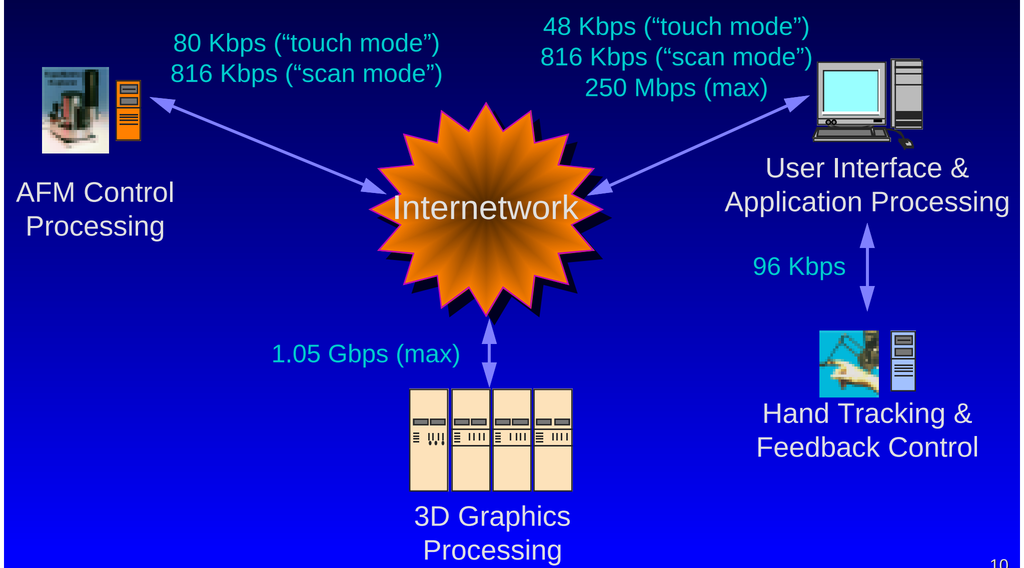
Distributed, Immersive, Virtual Laboratories

- The UNC *nanoManipulator* system
  - A virtual environment interface to a scanning-probe microscope
  - Provides *telepresence* on sample surfaces scaled 1,000,000:1

Distributed Virtual Laboratories
Networking challenges

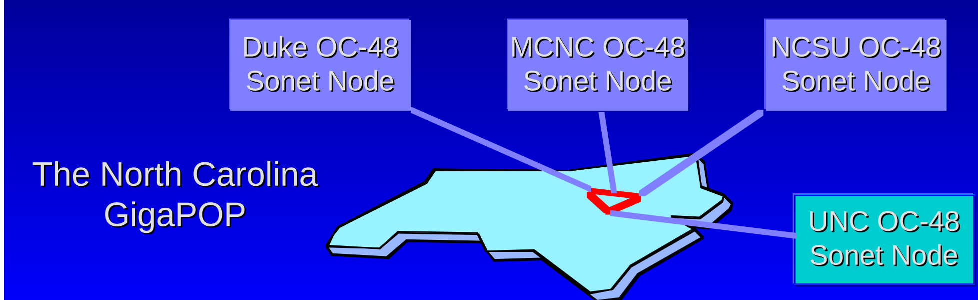


Distributed Virtual Laboratories
Networking challenges



Distributed Virtual Laboratories
nM distribution scenarios

- Scientific collaboration over Integrated Services networks
  - Equal distribution of graphics, tracking, and microscope hardware



Distributed Virtual Laboratories
nM distribution scenarios

- Scientific collaboration over Integrated Services networks
- Scientific collaboration over high-speed, best effort internetworks
  - Co-located graphics & microscope hardware, remote tracking & user interface
  - Co-located microscope hardware, tracking & user interface, remote graphics engine

Distributed Virtual Laboratories
nM distribution scenarios

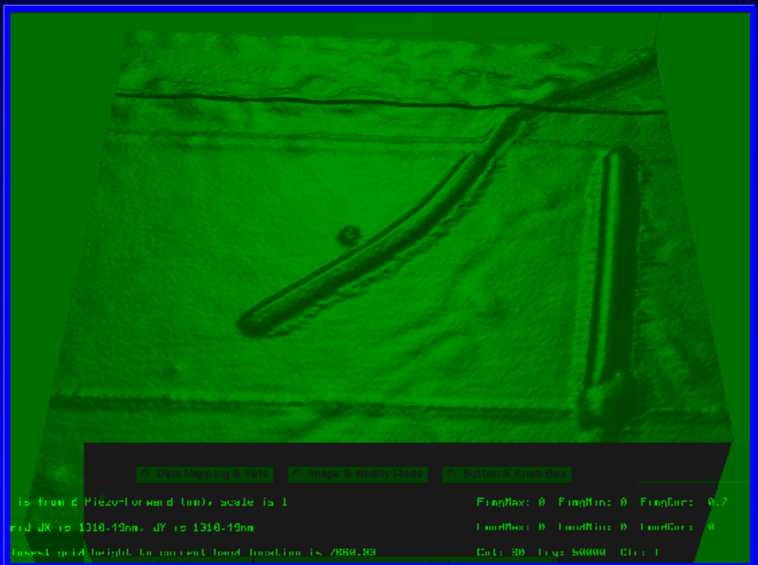
- Scientific collaboration over Integrated Services networks
- Scientific collaboration over high-speed, best effort internetworks
- Educational outreach over the Internet
  - Co-located graphics & microscope hardware, remote tracking & user interface

Adaptive, best-effort congestion control for real-time communications

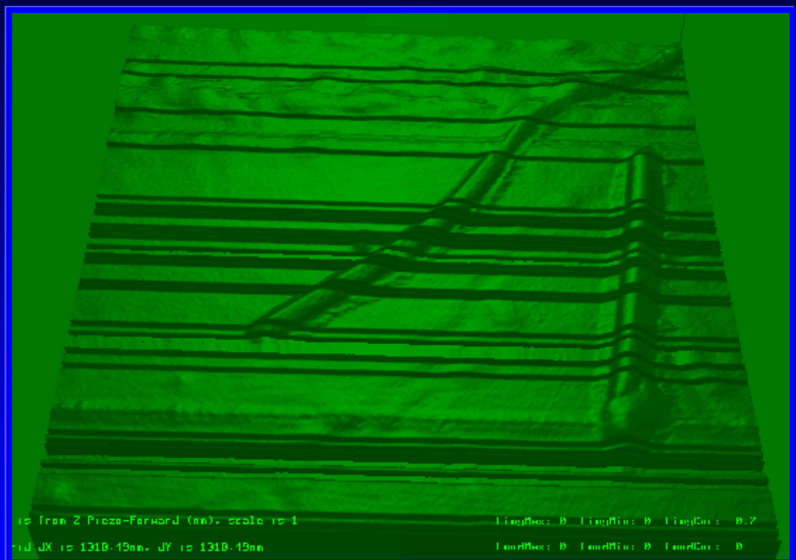
Outline

- Our driving problem — realizing distributed, immersive, virtual laboratories
  - The UNC *nanoManipulator* system
- The continuous media congestion control problem
- 2-Dimensional media scaling techniques
- Experimental results for Internet videoconferencing

Cont. Media Congestion Control
Effect of packet loss on UNC campus



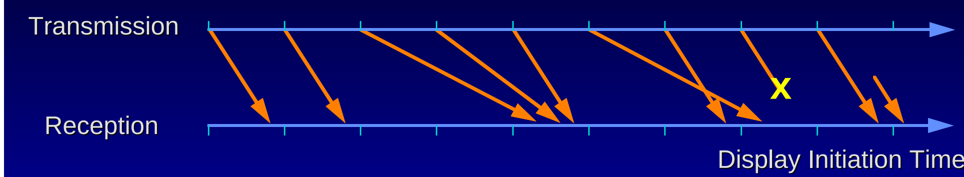
Cont. Media Congestion Control
Effect of packet loss across 16 hops



Cont. Media Congestion Control
The UNC approach

- Operating principle:
  - Network elements that cannot reserve, or support real-time allocation of resources, will persist for the foreseeable future.
- Focus on adaptive, best-effort transmission...
  - Treat the network as a black box – Assume only that sufficient bandwidth exists for some useful execution of the system
- ... with real-time media control at the endpoints

Cont. Media Congestion Control
The UNC approach



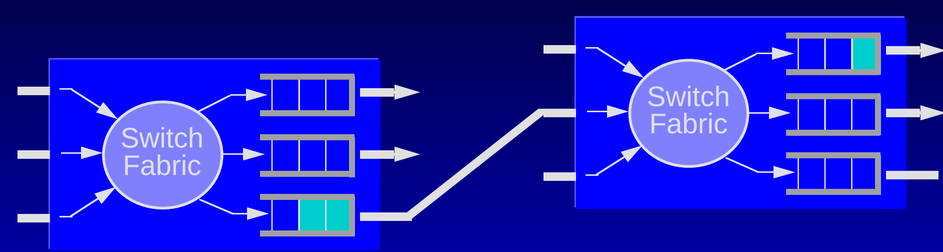
- Congestion in the small: *delay-jitter*
  - *Elastic queueing* to manage the trade-off between low latency playout and gap-rate
- Congestion in the large: *packet loss*
  - *Adaptive media scaling and packaging* to decrease network queueing (latency) and minimize packet loss

Adaptive, best-effort congestion control for real-time communications

Outline

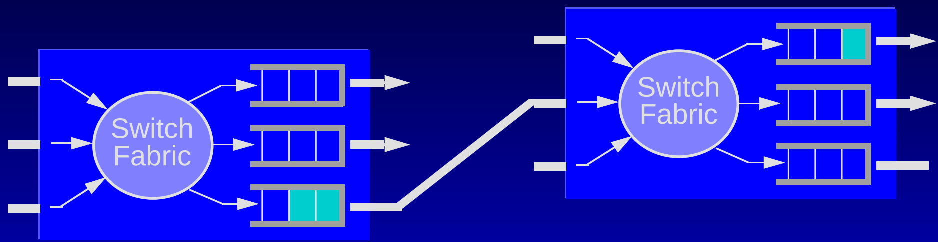
- Our driving problem — realizing distributed, immersive, virtual laboratories
  - The UNC *nanoManipulator* system
- The continuous media congestion control problem
- 2-Dimensional media scaling techniques
- Experimental results for Internet videoconferencing

Congestion Control
The nature of congestion



- What causes congestion?
  - Did our multimedia stream(s) cause the network to be congested?
  - Are there simply too many connections competing for too little bandwidth?

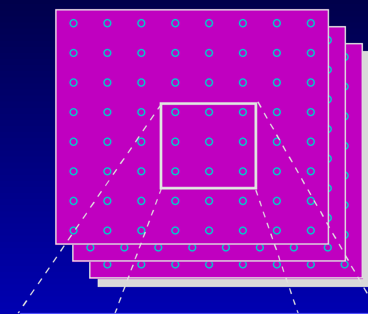
Congestion Control
The nature of congestion



- How can we make the best use of the (time varying) bandwidth that is available to our streams?
  - How can we determine what this bandwidth is?
  - How can we track how it changes over time?
  - How can we match our application's output to the available bandwidth?

Canonical Adaptive Congestion Control
Video bit-rate scaling

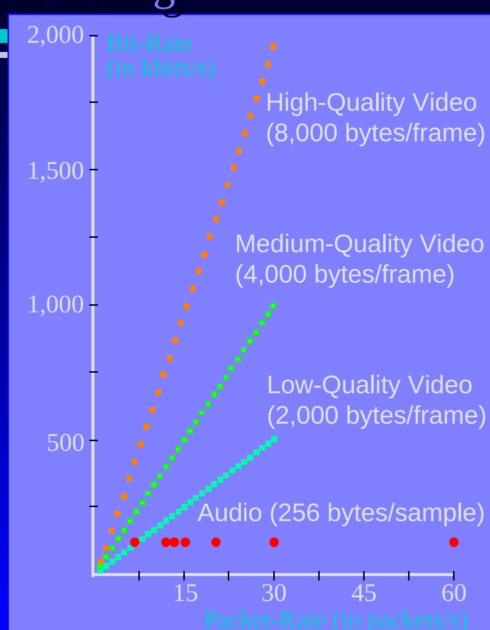
- Temporal scaling
  - Reduce the resolution of the stream by reducing the frame rate
- Spatial scaling
  - Reduce number of pixels in an image
- Frequency scaling
  - Reduce the number of DCT coefficients used in compression
- Amplitude scaling
  - Reduce the color depth of each pixel in the image
- Color space scaling
  - Reduce the number of colors available for displaying the image



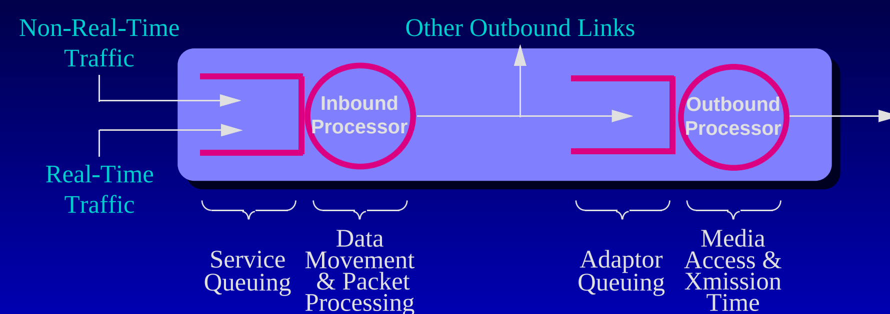
172	-18	15	-8	23	-9	-14	19
21	-34	24	-8	-10	11	14	7
-9	-8	-4	6	-5	4	3	-1
-10	6	-5	4	-4	4	2	1
-8	-2	-3	5	-3	3	4	6
4	-2	-4	6	-4	4	2	-1
4	-3	-4	5	6	3	1	1
0	-8	-4	3	2	1	4	0

UNC Adaptive Congestion Control
2-Dimensional media scaling

- Canonical approach to congestion
  - Reduce (video) bit-rate
- Alternate approach
  - View congestion control as a search of a 2-dimensional *bit-rate* x *packet-rate* space
  - Scale bit- and packet-rates simultaneously to find a sustainable *operating point*

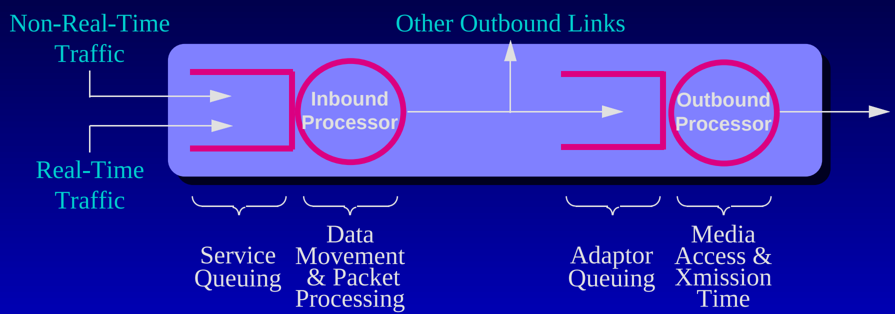


Bit- and Packet-Rate Scaling
An analytic model of media scaling



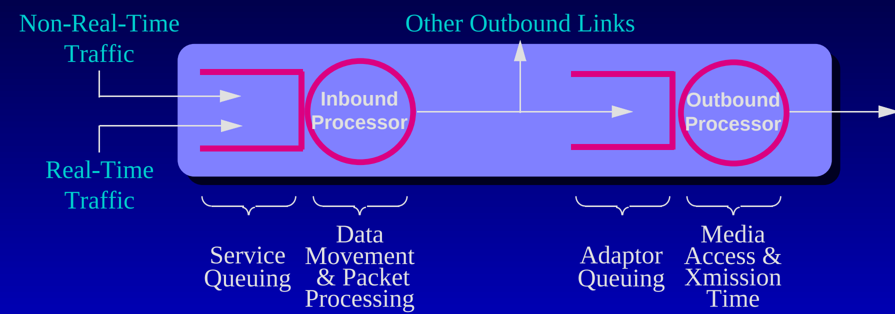
- Capacity constraints
  - the network is incapable of supporting the desired bit rate in any form
- Access constraints
  - the network can not support the desired bit rate with the current packaging scheme

Bit- and Packet-Rate Scaling
An analytic model of media scaling



- Reduce the packet-rate to adapt to an access constraint
  - Change the packaging or send fewer video frames
  - Primary Trade-off: higher latency (potentially)

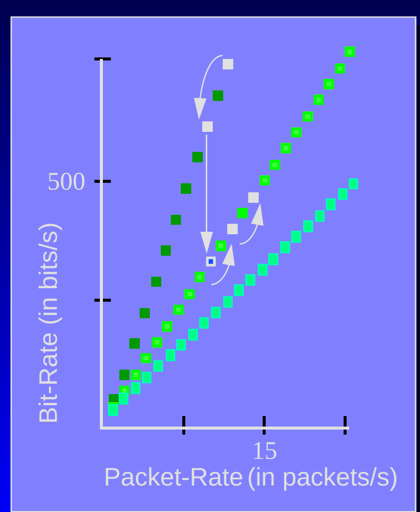
Bit- and Packet-Rate Scaling
An analytic model of media scaling



- Reduce the packet-rate to adapt to an access constraint
- Reduce the bit-rate to adapt to a capacity constraint
  - Send fewer video frames or fewer bits per video frame
  - Primary Trade-off: lower fidelity

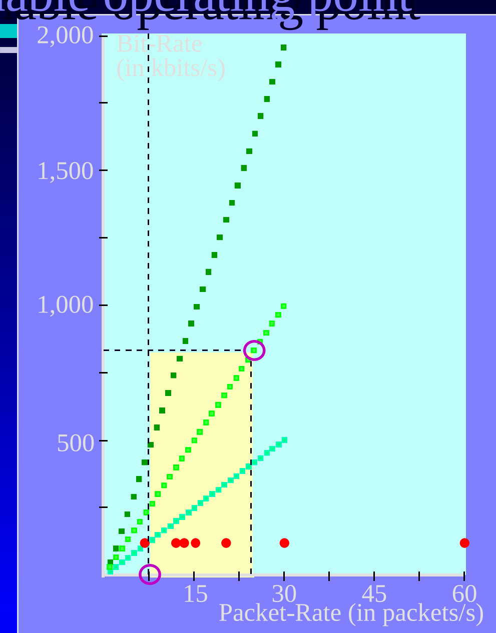
2-Dimensional Scaling Example
The "Recent Success" heuristic

- Initial operating point:
  (*high quality*, 12 *fps*)
- First adaptation:
  (*high quality*, 10 *fps*)
  - congestion persists
- Second adaptation:
  (*medium quality*, 10 *fps*)
  - congestion relieved
- First probe:
  (*medium quality*, 12 *fps*)
- Second probe:
  (*medium quality*, 14 *fps*)



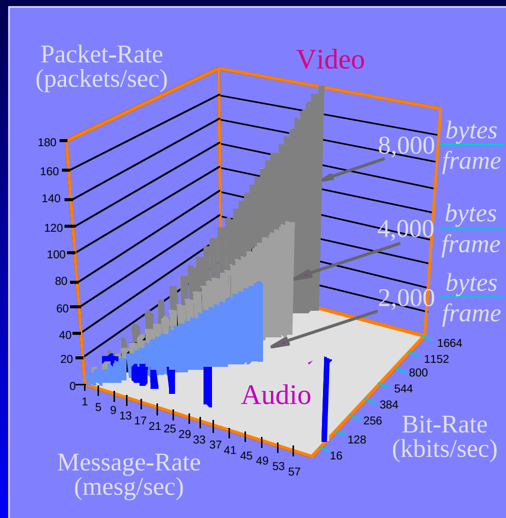
2-Dimensional Media Scaling
Finding a sustainable operating point

- The search space can be pruned by eliminating
  - points that lead to inherently high latency
  - points that lead to high latency given the state of the network

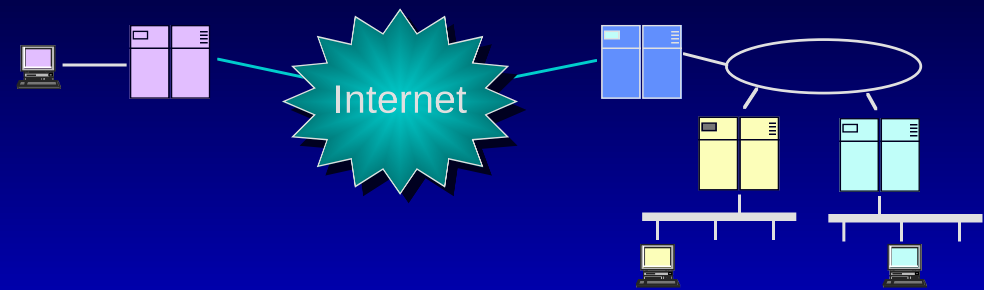


2-Dimensional Media Scaling
Dealing with effects of fragmentation

- The problem
  - A sender can only (directly) effect the *message rate*, not the *packet rate*
- Does fragmentation render message-rate scaling obsolete?

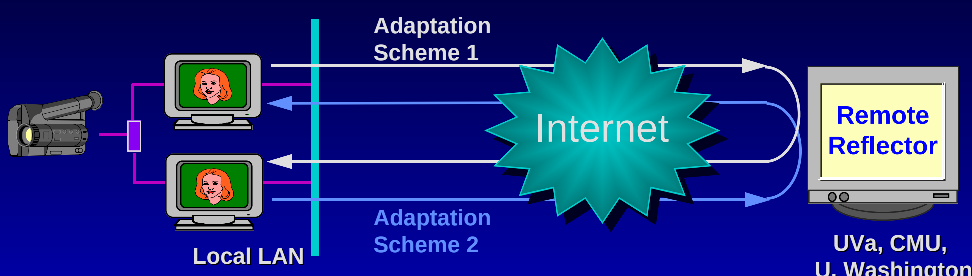


2-Dimensional Media Scaling
Does it work?



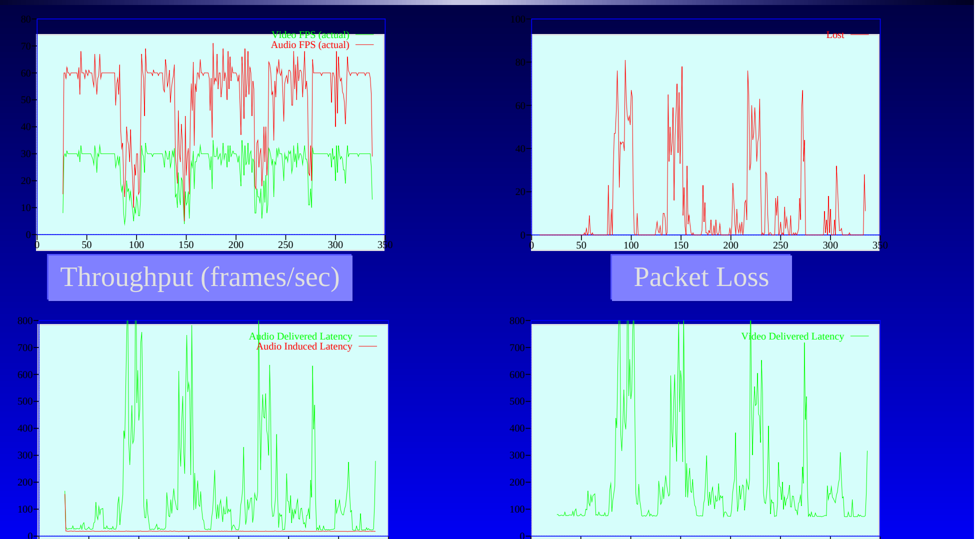
- Campus-sized internets?
- The Internet?

2-Dimensional Media Scaling
Does it work?



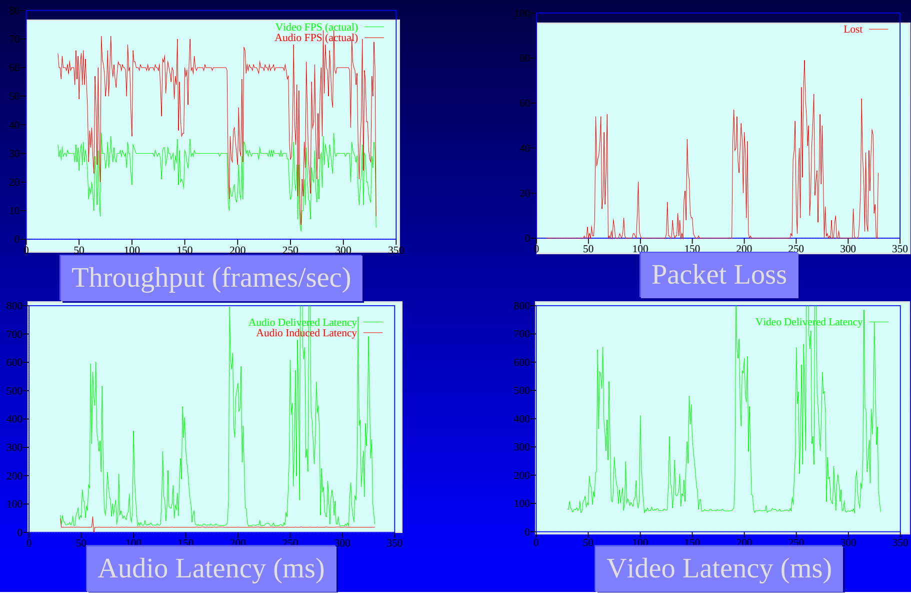
- Experiments
  - Baseline – UDP transmission, no adaptations
  - 1-Dimensional media scaling (video bit-rate scaling)
  - Audio and video media scaling & packaging
- Metrics
  - Delivered media frame rate (throughput)
  - Packet loss
  - Media stream latency
  - Adaptations performed over time

2-D Scaling on the UNC Campus
Performance with no media scaling

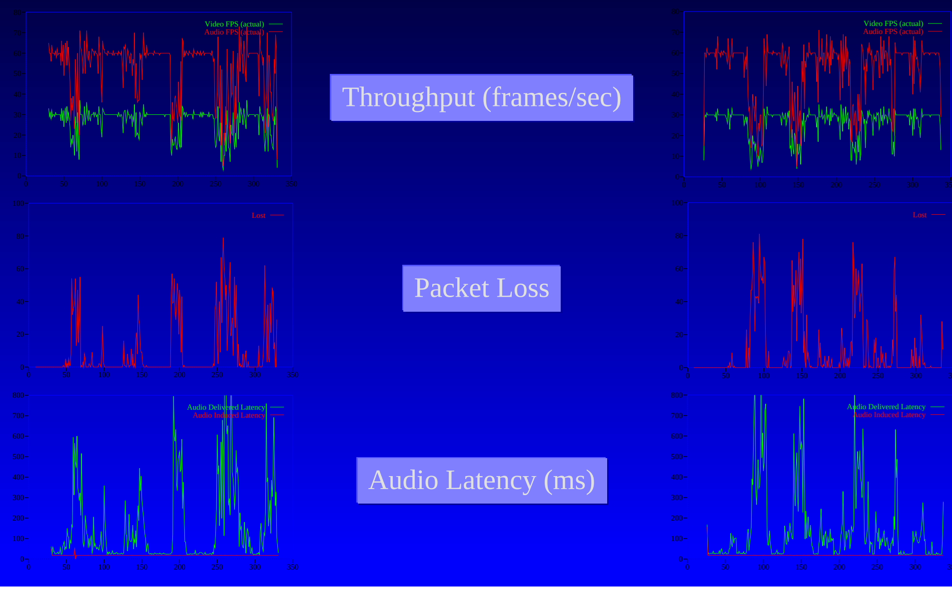


Throughput (frames/sec) | Packet Loss | Audio Latency (ms) | Video Latency (ms)

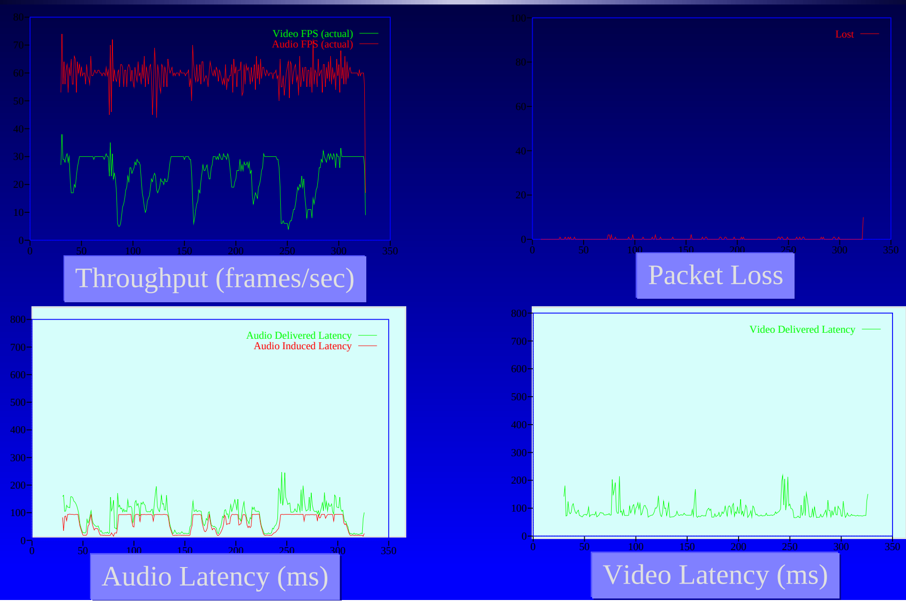
2-D Scaling on the UNC Campus
Performance with video scaling only



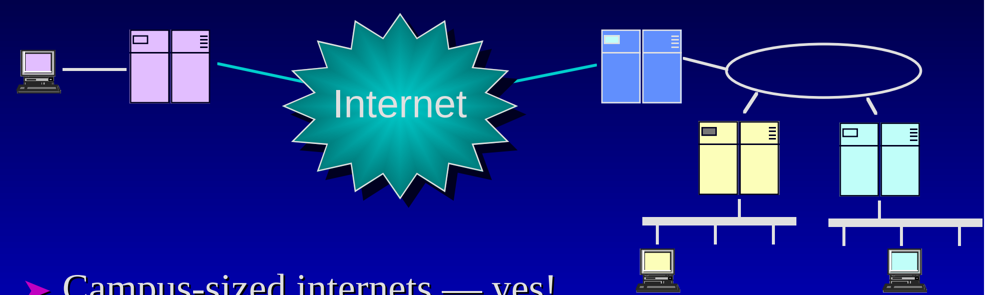
2-D Scaling on the UNC Campus
Video scaling *v.* no adaptation



2-D Scaling on the UNC Campus
Performance with 2-dimensional scaling



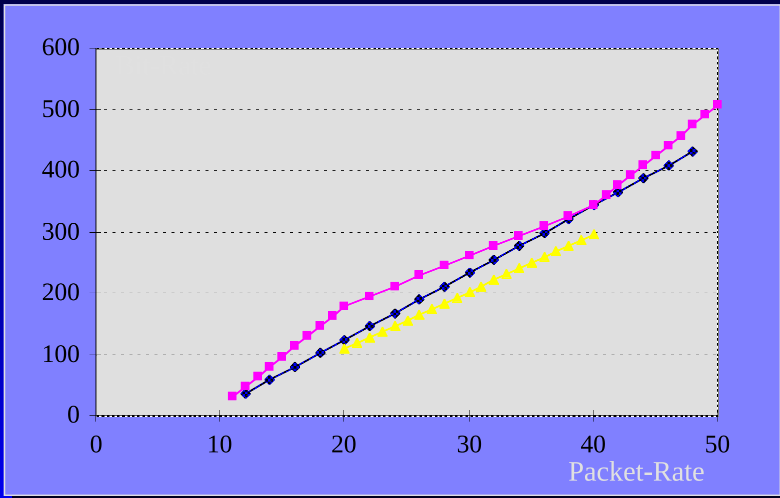
2-Dimensional Media Scaling
Does it work?



- Campus-sized internets — yes!
  - It “solves” the first-mile/last-mile problem
- The Internet? — *well...*
  - Does our necessary condition for success hold?
  - Does it hold often enough to be useful?
  - How much “room” is there for 2-D scaling in most codecs?

2-D scaling evaluation on the Internet

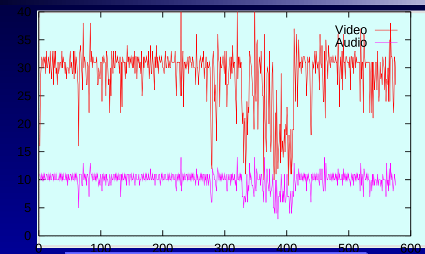
Media scaling in Intel’s ProShare™ codec



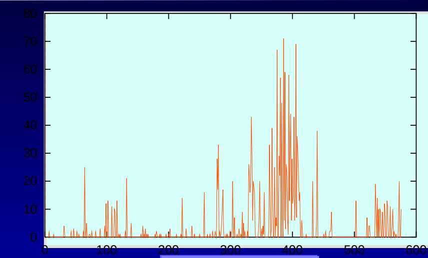
ProShare operating points

2-D scaling evaluation on the Internet

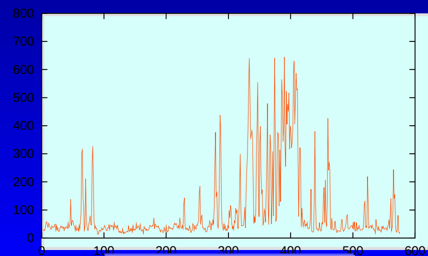
ProShare with no media scaling



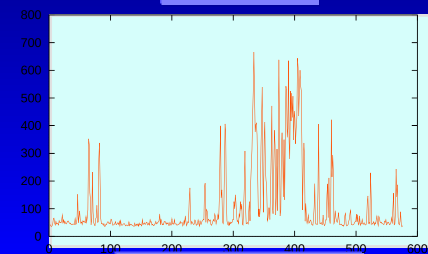
Throughput (frames/sec)



Packet Loss



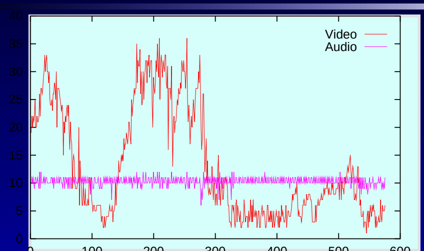
Audio Latency (ms)



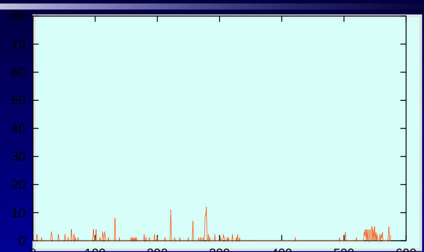
Video Latency (ms)

2-D scaling evaluation on the Internet

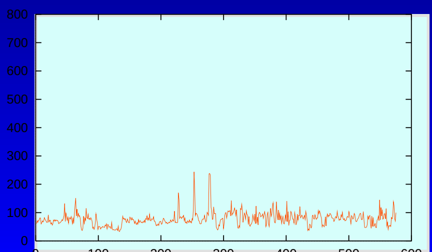
ProShare with 2-dimensional media scaling



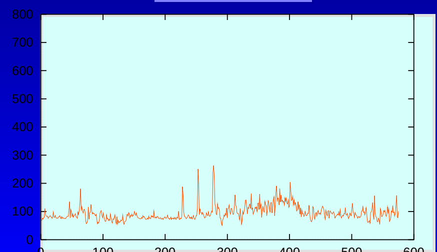
Throughput (frames/sec)



Packet Loss



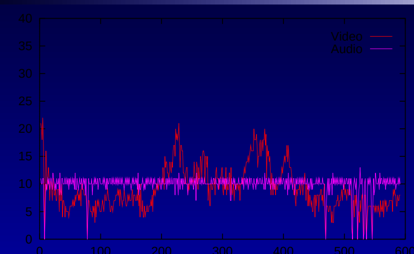
Audio Latency (ms)



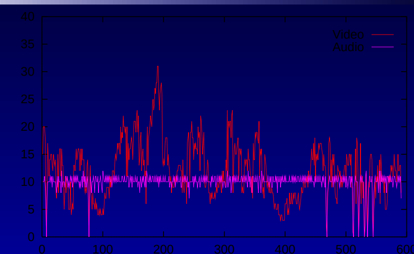
Video Latency (ms)

2-D scaling evaluation on the Internet

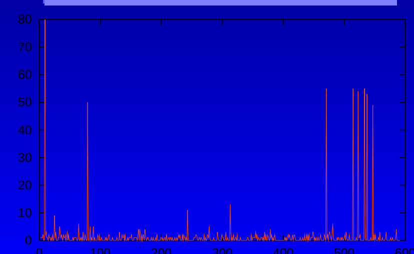
2-dimensional *v.* 1-dimensional media scaling



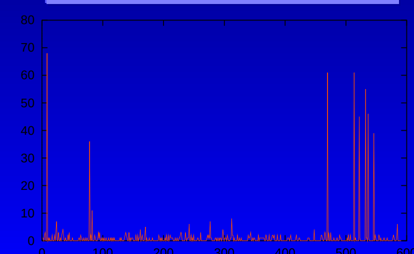
1-D Throughput (frames/sec)



2-D Throughput (frames/sec)



1-D Packet Loss



2-D Packet Loss

Sustainability Results
Adaptive methods on the Internet

- Results of an Internet performance study from UNC to UVa
  - Repeated trials from 10 am to 7 PM weekdays
  - Trials separated by at least two hours
  - Scattered over three months

Time Slot	Sustainable	Not Sustainable
10:00-12:00	67%	33%
12:00-14:00	50%	50%
14:00-16:00	8%	92%
16:00-18:00	25%	75%
18:00-20:00	44%	56%
Percentage	39%	61%

Adaptive, best-effort congestion control for real-time communications

Outline

- Our driving problem — realizing distributed, immersive, virtual laboratories
  - The UNC *nanoManipulator* system
- The continuous media congestion control problem
- 2-Dimensional media scaling techniques
- Experimental results for Internet videoconferencing

Adaptive, best-effort congestion control for real-time communications

Summary

- Real-time applications must be adaptive to be effective on the Internet
- Simple middleware adaptations are sufficient for accommodating most Internet pathologies within "the intranet"
  - Biasing how a bit-stream is partitioned into packets is more effective than reducing the bit-stream

Will best-effort techniques scale?
Router-based congestion control

- Recursively apply endpoint media adaptations in the network
  - Delay-jitter management adaptations
  - Congestion/flow control adaptations
- Compare performance against CBQ gateways
  - RED packet discard for TCP
  - "Delete Oldest & Advance" discard for multimedia