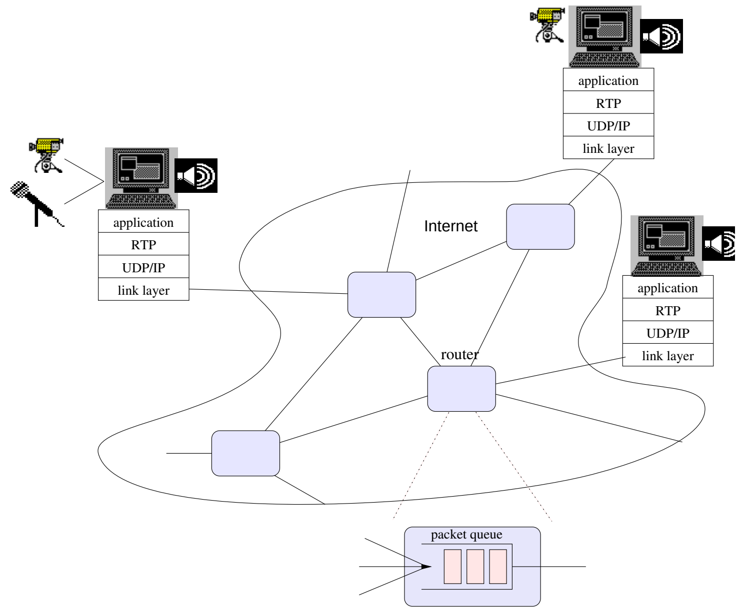
Real-Time Multimedia Services in the Internet

Henning Schulzrinne
GMD Fokus, Berlin
schulzrinne@fokus.gmd.de

March/April 1996

Overview



- Internet multimedia services:
  - packet scheduling for real-time services
  - real-time transport protocol and adaptive applications
  - personal mobility for teleconferences
- future work

Internet real-time services

- applications:
  - teleconferencing
  - teleteaching, seminars
  - media-on-demand
- characteristics:
  - unicast or multicast ($2 \ldots > 1000$)
  - delay sensitive: late packet = lost packet
  - loss tolerant
  - adapt to varying network delays & network bandwidths
- need to work in large-scale networks:
  - with and without resource reservation
  - sharing resources with other multimedia streams

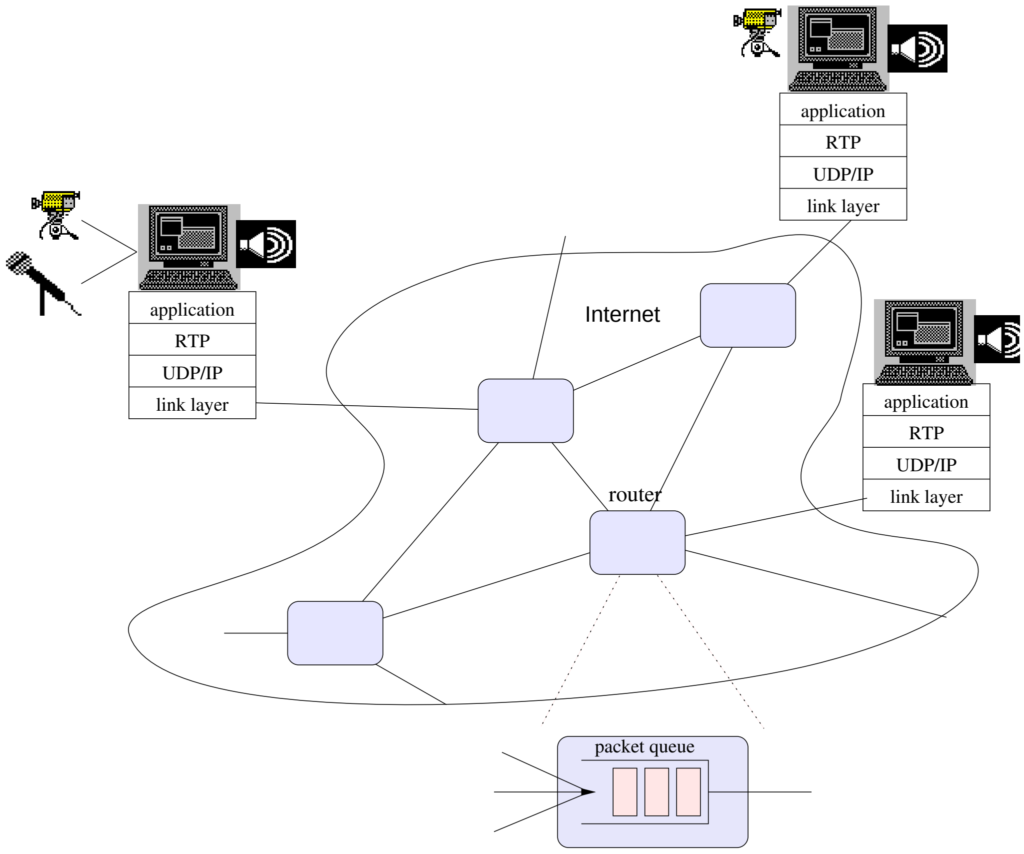
Challenges in Internet multimedia

- Internet = first (almost) integrated-service network
- large scale – user population, geography
- bandwidth (2.4 kb/s to 155 Mb/s) and end system diversity
- time-varying network parameters
- less than reliable
- multitude of operators

Contributions of my research

- analysed & implemented algorithms for reducing packet loss in real-time networks
  ➠ packet scheduling & drop policies
- analysed packet loss behavior
  ➠ queueing theoretic analysis
- designed transport protocol for real-time media
  ➠ RTP Internet standard
- experiments with delay & bandwidth adaptive applications
- designed and implemented protocol for advanced Internet telephony services

Packet Scheduling for Real-Time Services



Real-time network services

- real time = packets must reach destination within fixed delay, otherwise lost
- applications: interactive audio/video, sensor data, remote control, ...
- some real-time traffic *loss tolerant* (audio: 1 – 10% tolerable)
- performance metrics:
  - loss probability (instead of average delay)
  - loss correlation
- complementary to feedback control: faster, "safety valve"
- large number of flows ➠ limit state information

Role of scheduling

Goal: reduce end-to-end loss due to excessive delay by

- dropping late packets early on their path
- scheduling according to delay sensitivity

Research methodology:

- queueing theoretic models
- simulation
- implementation in Unix kernel, DARTnet experiments & measurements

Real-time packet scheduling policies

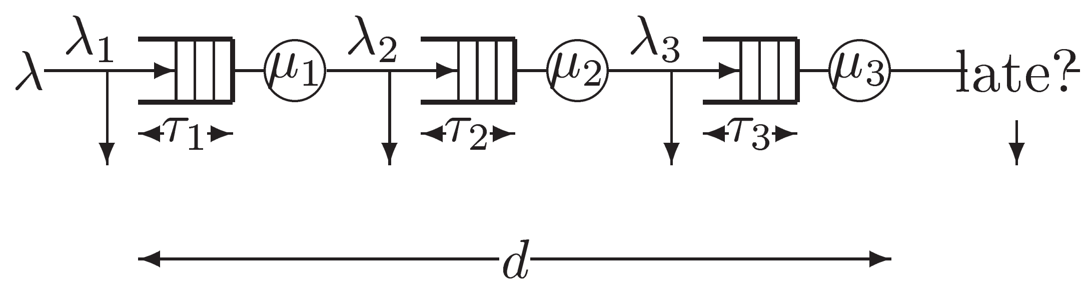
1. jitter-controlled: bounded delay variation
2. no lower delay bound
   - bandwidth guarantee (possibly delay bounded)
   - need-based (laxity), no guarantees
   - need-blind, best-effort

jitter-controlled, bandwidth guarantee may cause

- higher delay
- lower utilization

Selective dropping: goals and policies

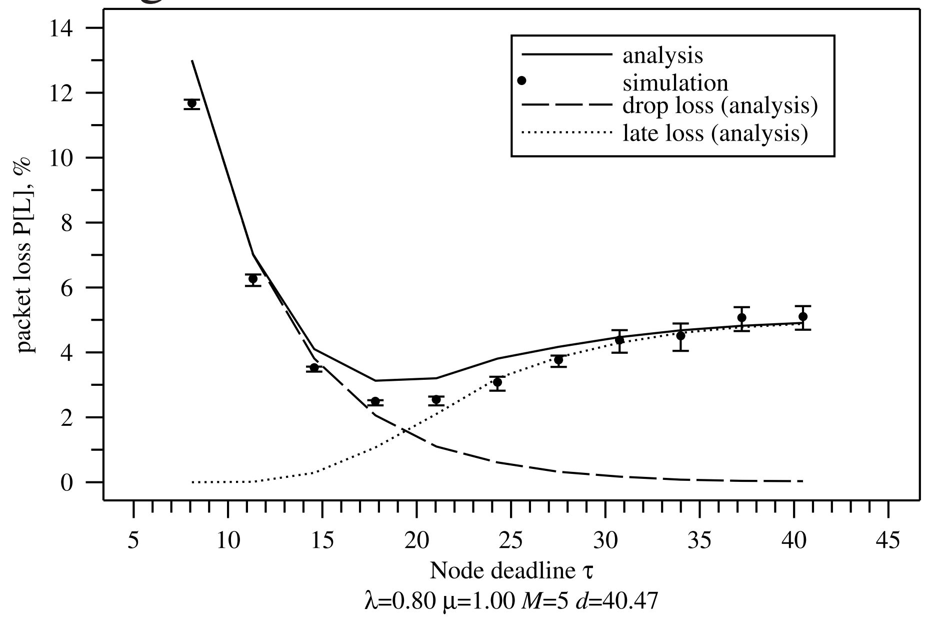
- goal: limit excessive delays and buffer overflows; graceful degradation
- discard packets that have missed or are *likely* to miss deadline
- $\boxed{\text{Impose } \textit{local} \text{ deadlines } \tau \text{ based on } \textit{global} \text{ deadline } d}$



- queueing analysis for queues with bounded waiting time:
  - local continuous/discrete time queues in isolation
  - end-to-end results

Selective dropping: sample results

Fix end-to-end deadline d to achieve nominal loss, then tune local deadline τ_i. For homogeneous traffic and deadlines:



- significant reduction in losses
- independence assumption reasonable
- analysis, simulation: same optimal τ

Scheduling: "need-based" policies

FIFO first-in, first-out

FIFO+ (Clark et al.) node estimates avg. delay for class

$$\text{priority} = \text{packet delay} - \text{avg. delay estimate}$$

HL (our approach) hop laxity:

$$\text{priority} = \frac{\text{time left (laxity)}}{\text{hops left}}$$

➠ give late packets chance to "catch up"

TP transit priority

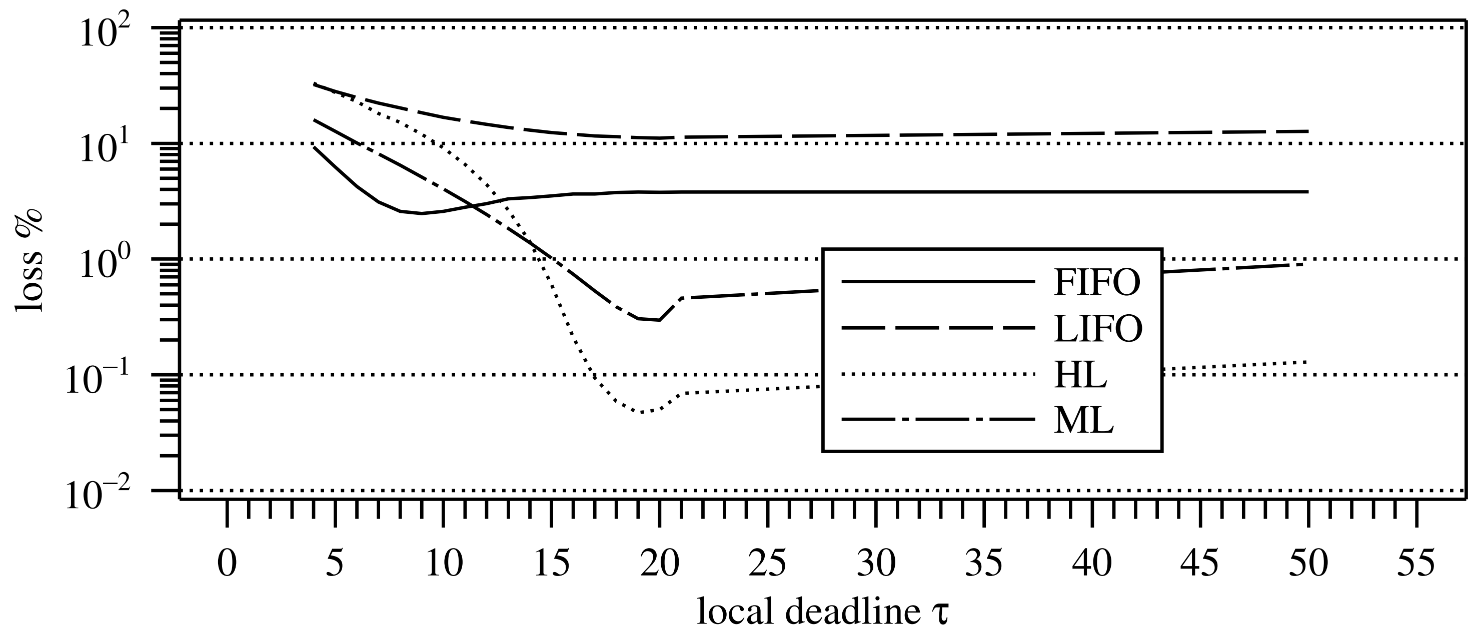
ML minimum laxity

Scheduling: fixed deadline simulation experiments

metric: fraction of packets exceeding e-e deadline ($d = 20$)

network: symmetrically loaded 5-hop network

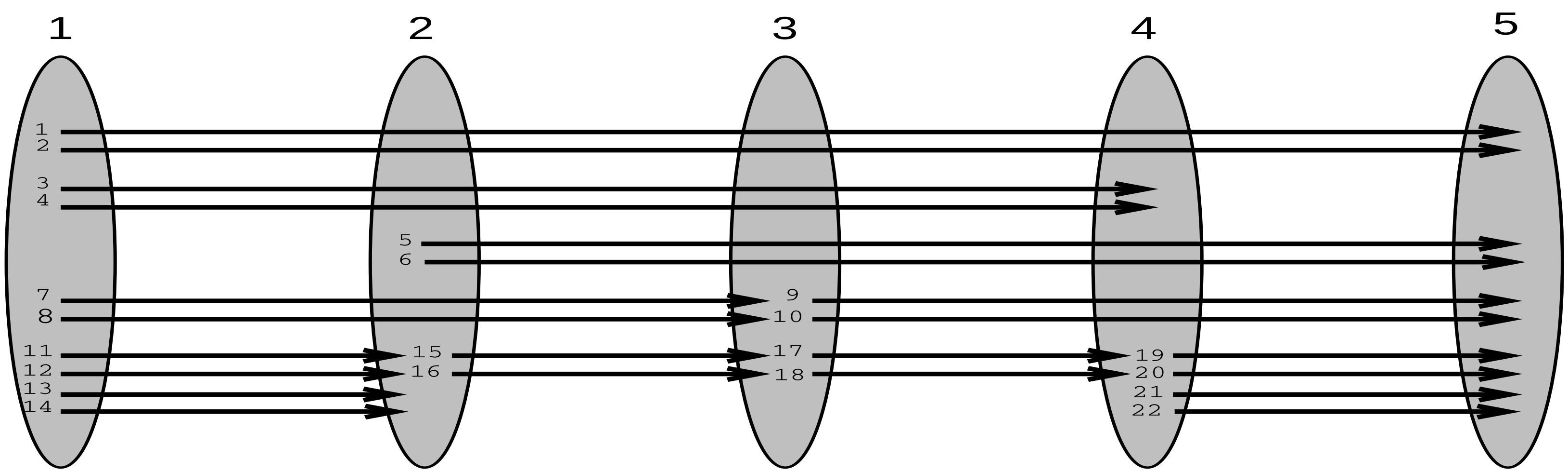
traffic model: discrete-time (batch arrivals)



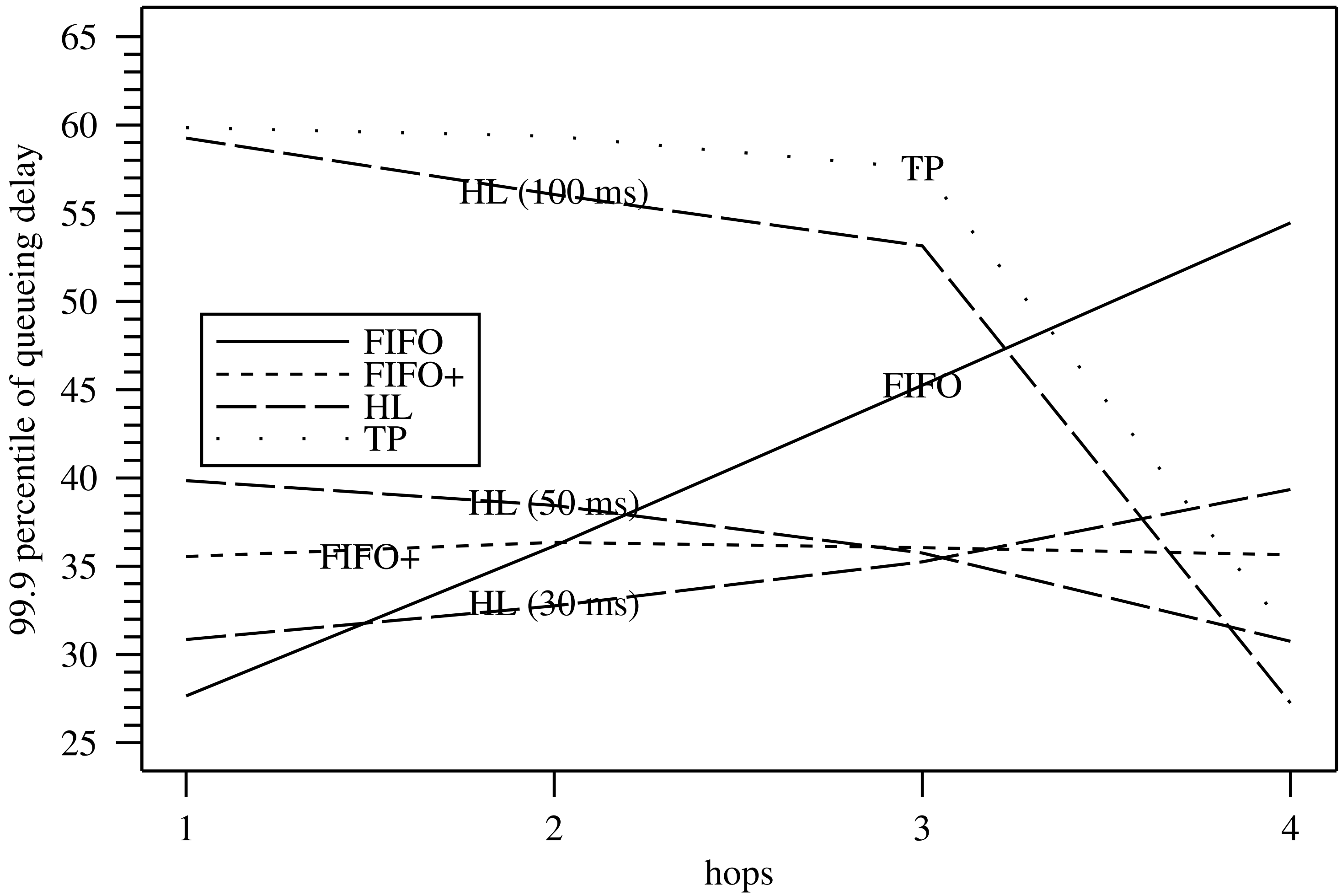
Scheduling: adaptive deadline services

metric: services can (slowly) adjust deadline ➠ mean delay and delay percentile

traffic models: two-state Markovian, conversational voice



Scheduling: adaptive deadline services



Result: HL provides controllable, distance independent service

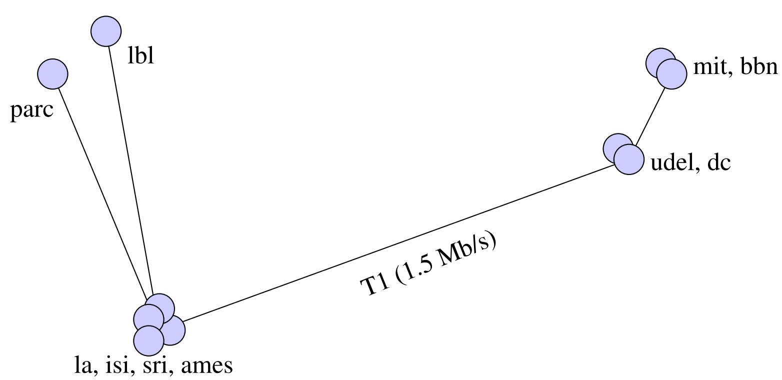
Scheduling: experimental evaluation

- router kernel (SunOS) implementation of HL scheduling: ➠ device driver packet scheduler

  IP connectionless (datagram) protocol

  ST-II connection-oriented network layer, supports resource reservation

- DARTnet test network:



Hop-laxity results

- end-to-end performance with interfering traffic
- single 5 hop path, laxity = 500 ms
- traffic: recorded conversational voice streams

		percentile (ms)		
	mean (ms)	0.90	0.99	0.999
FIFO	60.5	80.1	162.5	243.0
HL	55.5	64.8	99.9	133.5

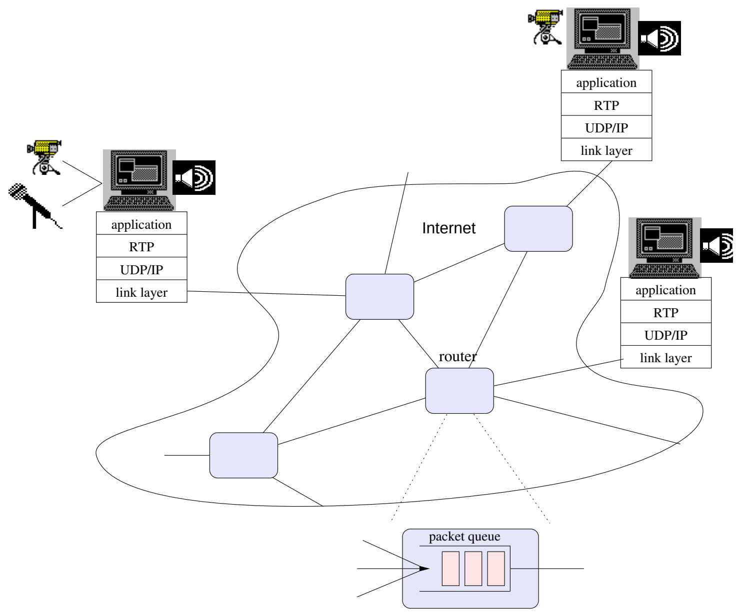
➠ HL reduces tail of delay distribution significantly

- significant improvement if several nodes are highly loaded
- observation: doesn't work well for single bottleneck node
- laxity parameter strongly affects results

Scheduling: summary

- selective dropping ➠ overload control
- laxity-based scheduling to improve distance fairness
  ... but: requires revised header, per-packet scheduling overhead
- real-time equivalent to early random drop
- best suited for slower, congested networks
- related work: statistical characterization of packet loss process via queueing analysis

The Real-Time Transport Protocol and Adaptive Applications



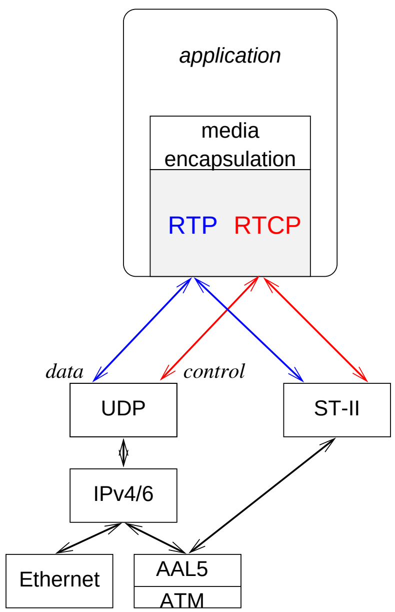
Challenges for a RT transport protocol

- usable for wide range of audio/video encodings + other RT services
- support intra/intermedia synchronization
- support layered encodings
- support forward error correction
- provide scalable quality-of-service (QOS) feedback from network & receivers
- work with variety of conferencing styles
- support delay & bandwidth adaptive applications
- end-to-end ➠ relies on lower-layer scheduling/reservation

➠ general framework + profiles

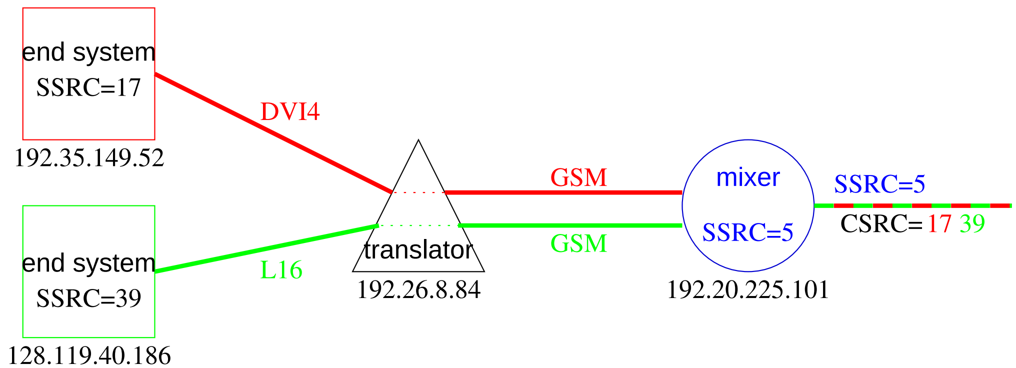
RTP

- real-time transport protocol (proposed standard RFC 1889/1890) (joint work with S. Casner, V. Jacobson, R. Frederick)
- widely used in multimedia tools ($\geq$ 10: nv, vic, NeVoT, IP/TV, . . .)
- part of ITU H.323, Netscape LiveMedia framework
- separation of data and control ➠ RTP and RTCP



RTP features

- separate data and control
- supports mixers and translators



- provides QOS feedback mechanism ➠ periodic multicast reports
  - allows maintenance of member list
  - membership $\uparrow \rightarrow$ report frequency $\downarrow$ ➠ $\approx$ const. control bw
  - data senders get more control bandwidth
  - monitor loss, jitter, delay to sender
  - correlate media time and real time ➠ intermedia synchronization
- conference control ➠ information to link different sessions

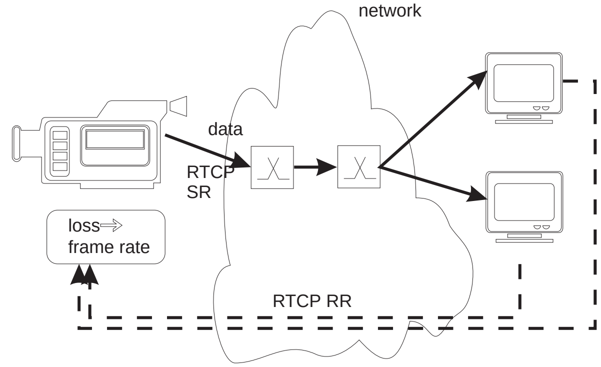
Application use of RTP: bandwidth adaptation

Many multimedia applications . . .

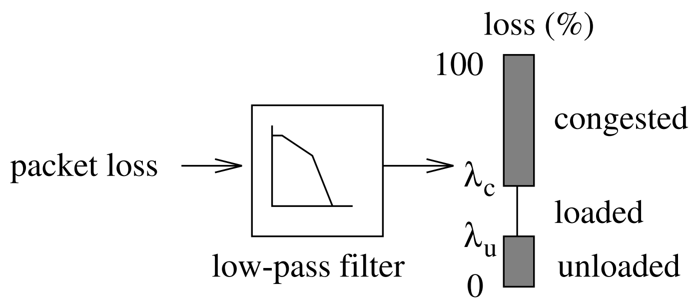
- are long-lived ➠ hard to allocate bandwidth
- are able to adjust their data rates (above minimum)

Also: current networks do not guarantee bandwidth
➠ control bandwidth based on RTP QOS feedback

Use of RTP as a feedback mechanism

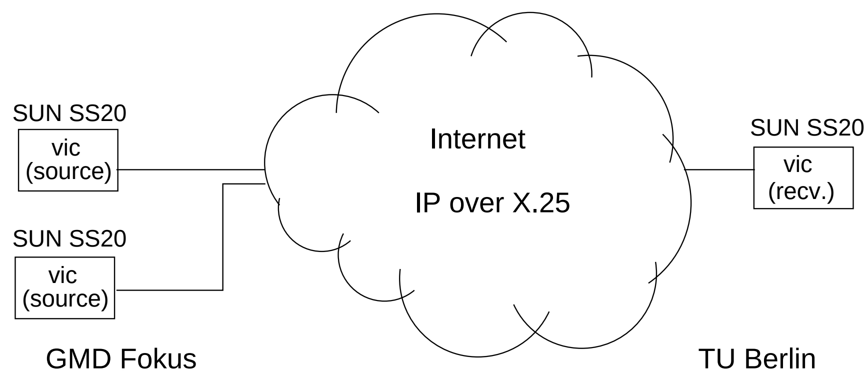


- video application bandwidth is based on RTP-based network feedback:
- RTP provides framework for using feedback, e.g.,
- low losses ➠ slow bandwidth increase ➠ higher framerate
- high losses ➠ bandwidth decrease ➠ lower framerate

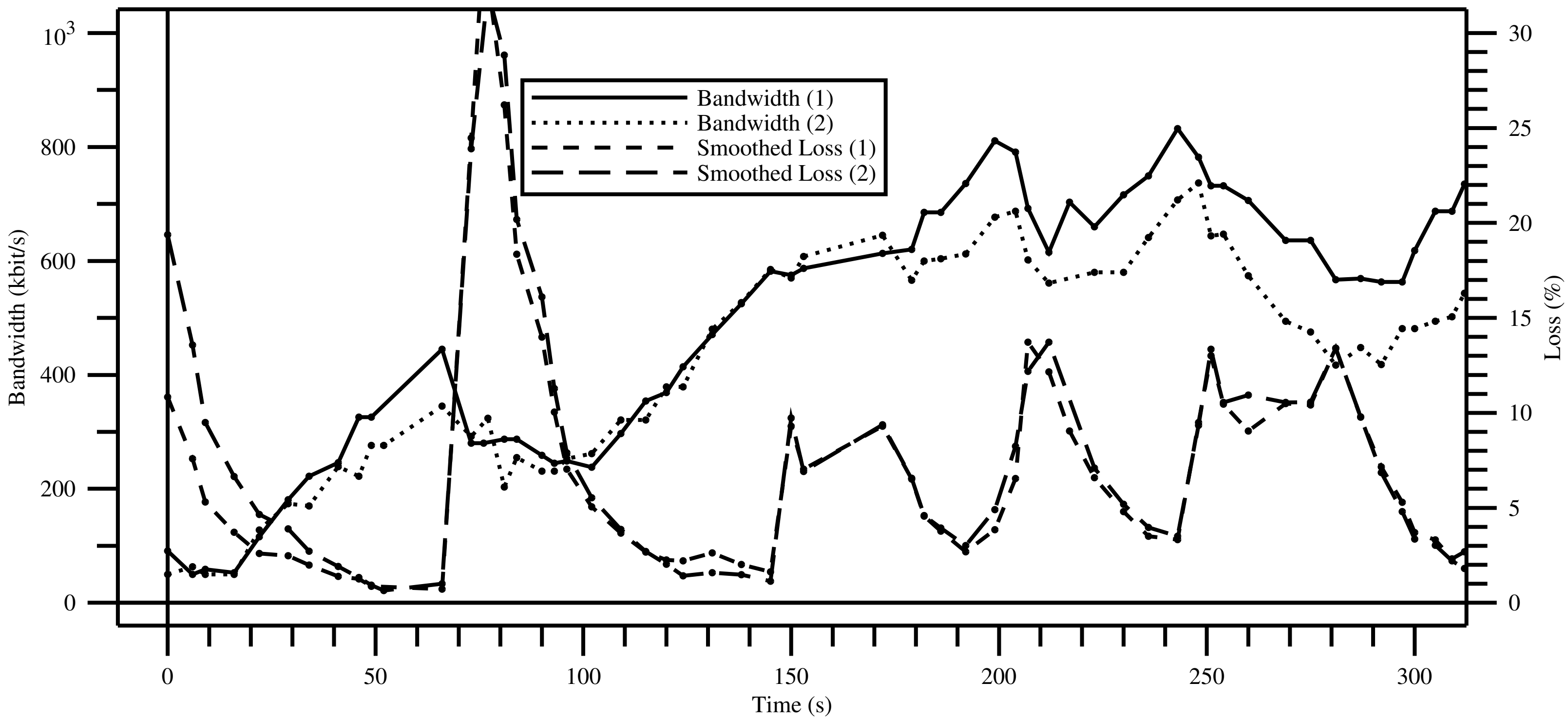


Internet Scenario

- Measurements made on the 2 Mbit/s X.25 link between GMD Fokus and TU Berlin (5 hops distance)
- linear regulator with deadzone between 5% and 10%



Internet measurement

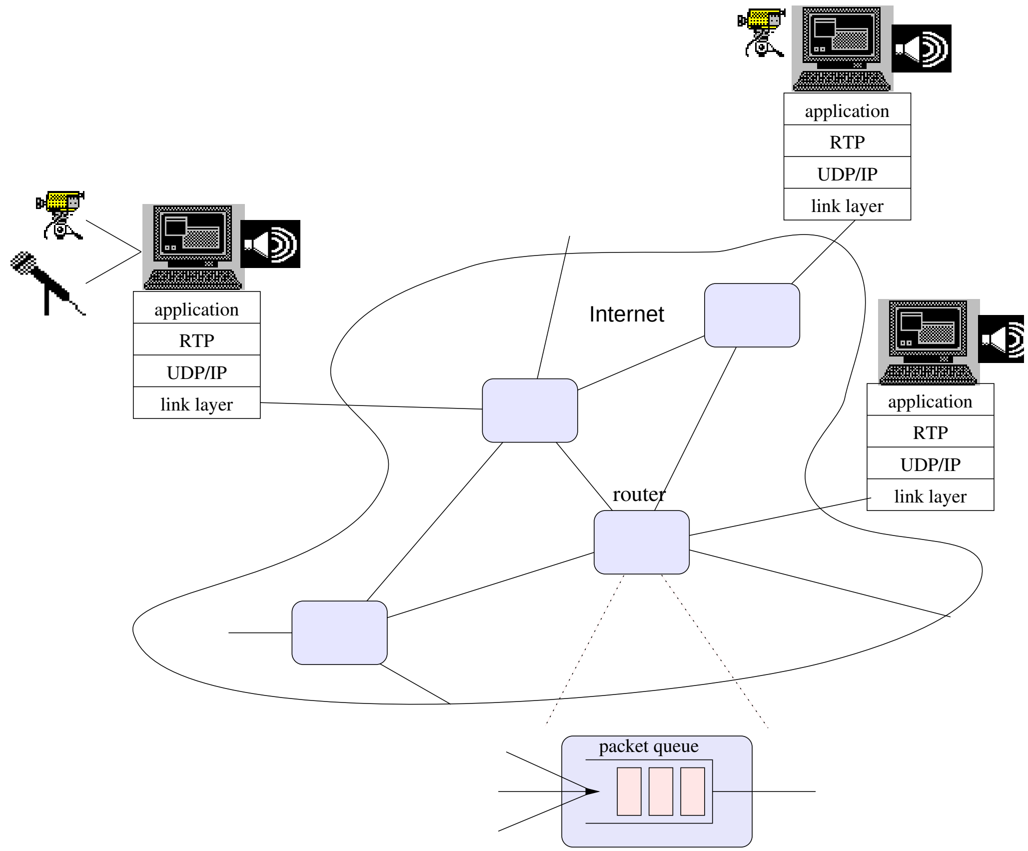


➠ divides bandwidth reasonably fairly, within loss region

RTP and adaptive applications: summary

- RTP $\rightarrow$ common protocol for real-time multimedia
- a protocol with the functionality to support:
  - QOS feedback
  - mixers and translators
  - a variety of conferencing styles
- application control allows fair sharing of available bandwidth
- very bandwidth diverse internetworks ➠ RTP support for
  - adaptive bandwidth
  - layered encodings
  - filters and translators
  - general framework ➠ extensions for reliable multicast

Personal Mobility for teleconferences



Internet multimedia and telephony

- "integrated services": same network & protocols for data + media-on-demand + interactive multimedia
- same tools, media formats for media-on-demand, teleteaching, conferencing, telephony
- but: each with different control requirements
- current control:
  - multicast directories for pre-arranged meetings
  - arrange Internet phone call by phone or email...
  - need to know IP number (may change!)

The telecom approach: UPT

UPT = universal personal telecommunications

- (mostly) dumb end systems
- permanent telephone *number*
- mapped to current landline number or mobile number
- separation: service provision $\leftrightarrow$ mapping (“number portability”)?
- year 2010?

Internet personal mobility

- Internet “telephone number”: email address
- reuse directory infrastructure, mailing lists, . . .
- maintain email address across moves
- separate provision of “pipe” and control
- SMTP for locating users
- HTTP-like protocol for mapping email address to current location
- implementation integrated with multimedia building blocks

Conference control: future work

- freephone (1-800) facilities (ecash, reimburse provider, …)
- how to reach from pay phone?
- full PBX functionality: transfer in mid-call, switch between calls, …
- interoperation with ISDN, POTS, H.320, T.120, …
- forwarding to answering machine?
- transition from two-party to multicast?
- only usable with permanently connected system (intranet, CATV Internet appliance)

Conclusion

- Internet multimedia: more than just bandwidth reservation
- real-time scheduling and selective dropping ➠ improved application behavior
- loss correlation properties ➠ loss recovery
- real-time transport protocol ➠ adaptive applications
- conference control ➠ "universal communication identifier"

Future work

Supporting Internet multimedia on a large scale ➠ research issues throughout the stack

- ATM ABR services for bandwidth-adaptive services?
- bandwidth adaptation for large multicast groups
- fair sharing of bandwidth between data and real-time services
- integration of telephones and Internet conferencing/telephony