CSE/EE 461 – Lecture 19

Quality of Service

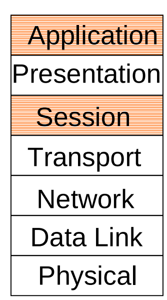
David Wetherall
djw@cs.washington.edu

Last Time …

- HTTP and the Web (but not HTML)

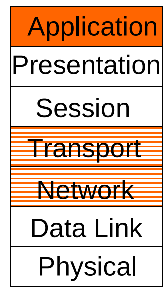
- Focus
  - How do Web transfers work?

- Topics
  - HTTP, HTTP1.1
  - Get-If-Modified
  - Caching and Consistency



This Lecture

- Introduction to Quality of Service
- Focus
  - What transports do applications need?
- Topics
  - Real-time versus Elastic applications
  - Adapting to variable delay
  - Token buckets as bandwidth descriptors

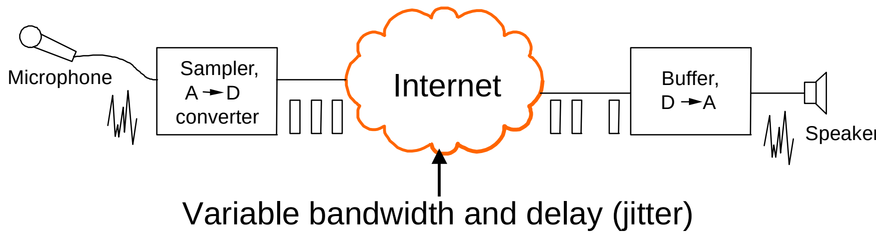


Internet "Best Effort" Service

- Our network model so far:
  - IP at routers: a shared, first come first serve (drop tail) queue
  - TCP at hosts: probes for available bandwidth, causing loss
- The mechanisms at routers and hosts determine the kind of service applications will receive from the network
  - TCP causes loss and variable delay, and Internet bandwidth varies!
- Q: What kinds of service do different applications need?
  - The Web is built on top of just the "best-effort" service
  - Want better mechanisms to support demanding applications

An Audio Example

- Playback is a real-time service in the sense that the audio must be received by a deadline to be useful

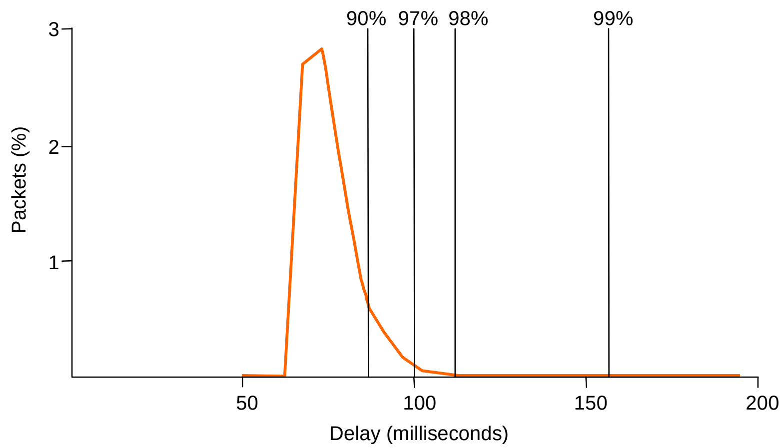


- Real-time apps need assurances from the network
- Q: What assurances does playback require?

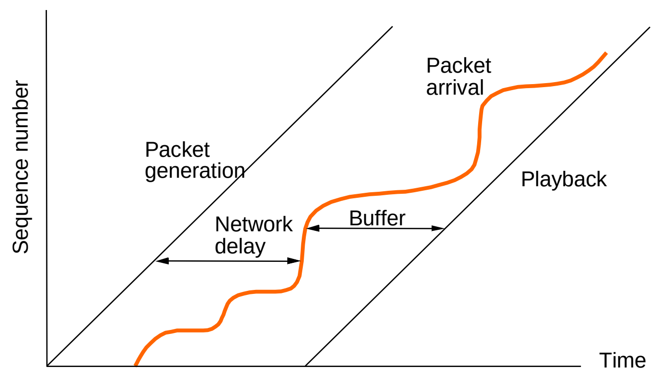
Network Support for Playback

- Bandwidth
  - There must be enough on average
  - But we can tolerate to short term fluctuations
- Delay
  - Ideally it would be fixed
  - But we can tolerate some variation (jitter)
- Loss
  - Ideally there would be none
  - But we can tolerate some losses

Example: Delay and Jitter

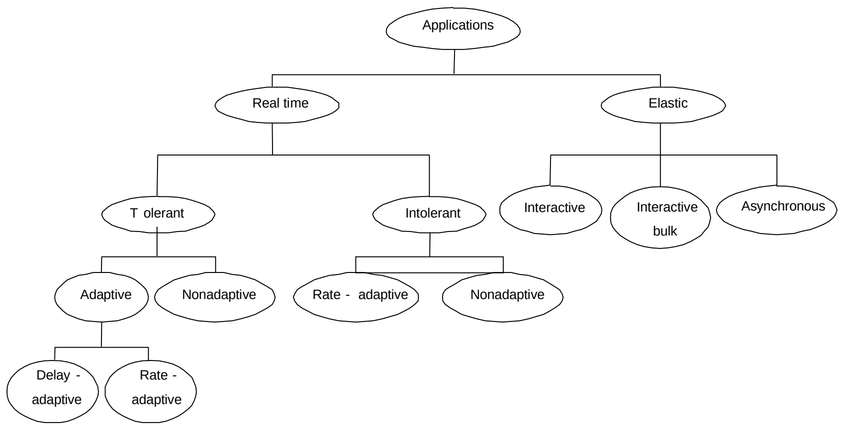


Tolerating Jitter with Buffering



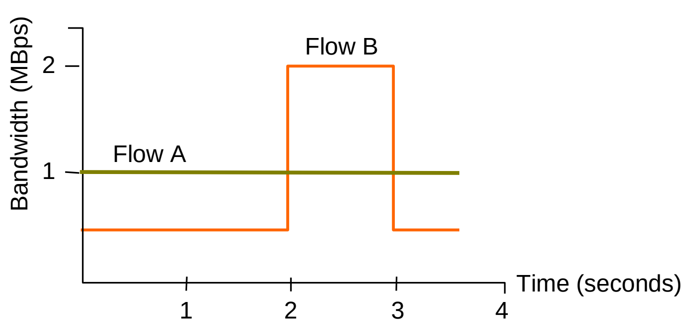
- Buffer before playout so that most late samples will have arrived

Taxonomy of Applications



Specifying Bandwidth Needs

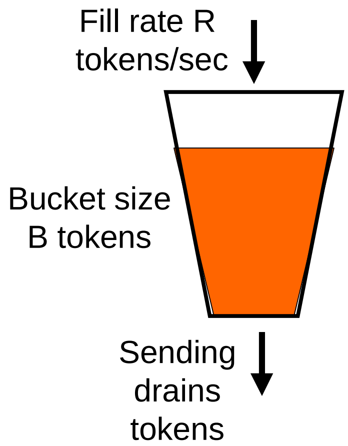
- Problem: Many applications have variable bandwidth demands



- Same average, but very different needs over time. One number. So how do we describe bandwidth to the network?

Token Buckets

- Common, simple descriptor
- Use tokens to send bits
- Average bandwidth is R bps
- Maximum burst is B bits



Key Concepts

- Different apps need different network support
  - Elastic versus real-time applications
- Adaptation is a key technique, e.g, playout buffer
- Token buckets are a simple bandwidth descriptor
- Next time: How do we build networks that provide more assurances than TCP/IP so far?