The hand that hath made you fair hath made you good.
--- William Shakespeare

Acknowledgement: this lecture is partially based on the slides of Dr. Larry Peterson
Outline (Quality of Service)

- Real-time Applications
- QoS support
  - Integrated Services
  - Differentiated Services
Outline

- Real-time Applications
- QoS support
  - Integrated Services
  - Differentiated Services
Real-time Applications

- Require “deliver on time” assurances
  - must come from inside the network (because end-to-end retransmission may render the delay to be too huge that data becomes useless even if delivered)

- Example application (audio)
  - sample voice once every 125us
  - each sample has a playback time
  - packets experience variable delay in network
  - add constant factor to playback time: playback point
Playback Buffer

- Sequence number
- Time
- Packet generation
- Network delay
- Buffer
- Packet arrival
- Playback
Example Distribution of Delays (of an Internet connection)
Taxonomy of real-time applications

Applications

- Elastic
  - Non real-time data

Real time

  - Intolerant
    - Industrial control
  - Tolerant
    - Audio, video, ...

Nonadaptive

  - Rate adaptive
    - Video

Adaptive

  - Delay adaptive
    - Audio
Approaches to QoS support

- Objective: to provide different service to different applications

- Two broad categories
  - *Fine-grained* approaches: QoS support per individual applications/flows
    - Integrated services
  - *Coarse-grained* approaches: QoS support per classes of data or aggregated traffic
    - Differentiated services
Outline

- Real-time Applications

- QoS support
  - Integrated Services
  - Differentiated Services
Integrated Services (IntServ)

- Two service classes
  - Guaranteed: for intolerant applications
  - controlled-load: for tolerant, adaptive applications
    - To emulate a lightly loaded network, using WFQ to isolate controlled load traffic from others

- Mechanisms
  - Flowspecs
  - Admission control
  - Reservation protocol
  - Packet classifying & scheduling
Flowspec

- **Tspec**: describes flow’s traffic characteristics
  - average bandwidth + burstiness: *token bucket* filter
  - must have a token to send a byte; must have \( n \) tokens to send \( n \) bytes
  - token rate \( r \), bucket depth \( B \)
  - start with no tokens
  - accumulate tokens at rate of \( r \) per second
  - can accumulate no more than \( B \) tokens

- **Rspec**: describes service requested from network
  - guaranteed: delay target
  - controlled-load: none
Admission Control

- Decide if a new flow can be supported
- Answer depends on service class

- Not the same as *policing*
  - Admission control is applied on a per-flow basis
  - Policing is applied on a per-packet basis to make sure a flow abides by the agreement
Reservation Protocol

- Called *signaling* in ATM

- Proposed Internet standard: RSVP (resource reservation protocol)
  - Consistent with the robustness of today’s connectionless model
    - Uses soft state (refresh periodically)

- Designed to support multicast
- Receiver-oriented
  - Different from circuit switching which is sender-oriented/initiated in “resource reservation”

- Two messages: PATH and RESV
  - Source transmits PATH messages every 30 seconds
  - Destination responds with RESV message

- Merge requirements in case of multicast
RSVP Example
RSVP vs. ATM (Q.2931)

- **RSVP**
  - receiver generates reservation
  - soft state (refresh/timeout)
  - separate from route establishment
  - QoS can change dynamically
    - thanks to soft-state and thus periodic update
  - receiver heterogeneity

- **ATM**
  - sender generates connection request
  - hard state (explicit delete)
  - concurrent with route establishment
  - QoS is static for life of connection
  - uniform QoS to all receivers
    - Because it is sender-centric
Packet classifying & scheduling

- Classification: associate each packet with the appropriate reservation
  - IPv4: by <source & destination addresses, source & destination ports, protocol number>
  - IPv6: can also by *FlowLabel*

- Scheduling: manage queues so each packet receives the requested service
Outline

- Real-time Applications
- QoS support
  - Integrated Services
  - Differentiated Services
Differentiated Services (DiffServ)

- Problem with IntServ: unscalable to maintain per-flow state

- Idea: segregate packets into a small number of *classes*
  - e.g., premium vs. best-effort

- Packets marked according to class at edge of network
- Core routers implement some per-hop-behavior (PHB)
Examples of DiffServ PHB

- Expedited Forwarding (EF): real-time
  - rate-limit EF packets at the edges, so as not to exceed network capacity
  - PHB implemented with class-based priority queues or WFQ

- Assured Forwarding (AF): reliability
  - customers sign service agreements with ISPs
  - edge routers mark packets as being “in” or “out” of profile
  - core routers run RIO: RED with in/out
    - *out* packets are dropped earlier than *in* packets
Summary of congestion control & resource allocation

- Queuing Discipline
- Congestion control
  - Reacting to Congestion
  - Avoiding Congestion
- QoS support
  - Integrated Services
  - Differentiated Services
Discussion

- TCP in wireless networks?
  - Packet loss may also be due to link unreliability, in addition to queue overflow

- Congestion control and QoS in
  - wireless networks in general
  - sensor networks
  - heterogeneous networks involving both wireline and wireless communications
Further readings

- An early overview

- Scheduling

- Congestion control
  - V. Jacobson, *Congestion Avoidance and Control*, ACM SIGCOMM’88
Further readings (contd.)

- **IntServ**
Further readings (contd.)

- **Diffserv**
  - B. Davie et al., *An Expedited Forwarding PHB (per-hop behavior)*, RFC 3246, Mar. 2002
Assignments

- Exercise #5
  - Chapter 6: Exercises 6, 10, 16, 49