CPE348: Introduction to Computer Networks

Lecture #21: Chapter 6.2

Jianqing Liu
Assistant Professor of Electrical and Computer Engineering, University of Alabama in Huntsville

jianqing.liu@uah.edu
http://jianqingliu.net

Quality of Service

- For many years, packet-switched networks have offered the promise of supporting multimedia applications, that is, those that combine audio, video, and data.
  - Once digitized, audio and video information become like any other form of data—a stream of bits to be transmitted.
  - One obstacle to the fulfillment of this promise has been the need for higher-bandwidth links.
- There is more to transmitting audio and video over a network than just providing sufficient bandwidth, however.

Quality of Service

- Participants in a telephone conversation, for example, expect to be able to converse in such a way that one person can respond to something said by the other and be heard almost immediately.

- Thus, the timeliness of delivery can be very important. We refer to applications that are sensitive to the timeliness of data as *real-time applications*.

Quality of Service

- Voice and video applications tend to be the general rule examples, but there are others such as industrial control—you would like a command sent to a robot arm to reach it before the arm crashes into something.

- Even file transfer applications can have timeliness constraints, such as a requirement that a database update complete overnight before the business that needs the data resumes on the next day.

Quality of Service

- The distinguishing characteristic of real-time applications is that they need some sort of assurance *from the network that data is likely to arrive on time (for some definition* of "on time").
- Whereas a non-real-time application can use an end-to-end retransmission strategy to make sure that data arrives *correctly, such a strategy cannot provide* timeliness.
- This implies that the network will treat some packets differently from others—something that is not done in the best-effort model.
- A network that can provide these different levels of service is often said to support quality of service (QoS).

Quality of Service

- Taxonomy(classification) of Real-Time Applications
  - The first characteristic by which we can categorize applications is their tolerance of loss of data,
    - One lost audio sample can be interpolated from the surrounding samples with relatively little effect on the perceived audio quality.
    - A robot control program is likely to be an example of a real-time application that cannot tolerate loss—losing the packet that contains the command instructing the robot arm to stop is unacceptable.
  - real-time applications are categorized as *tolerant or intolerant* depending on whether they can tolerate occasional loss

Quality of Service

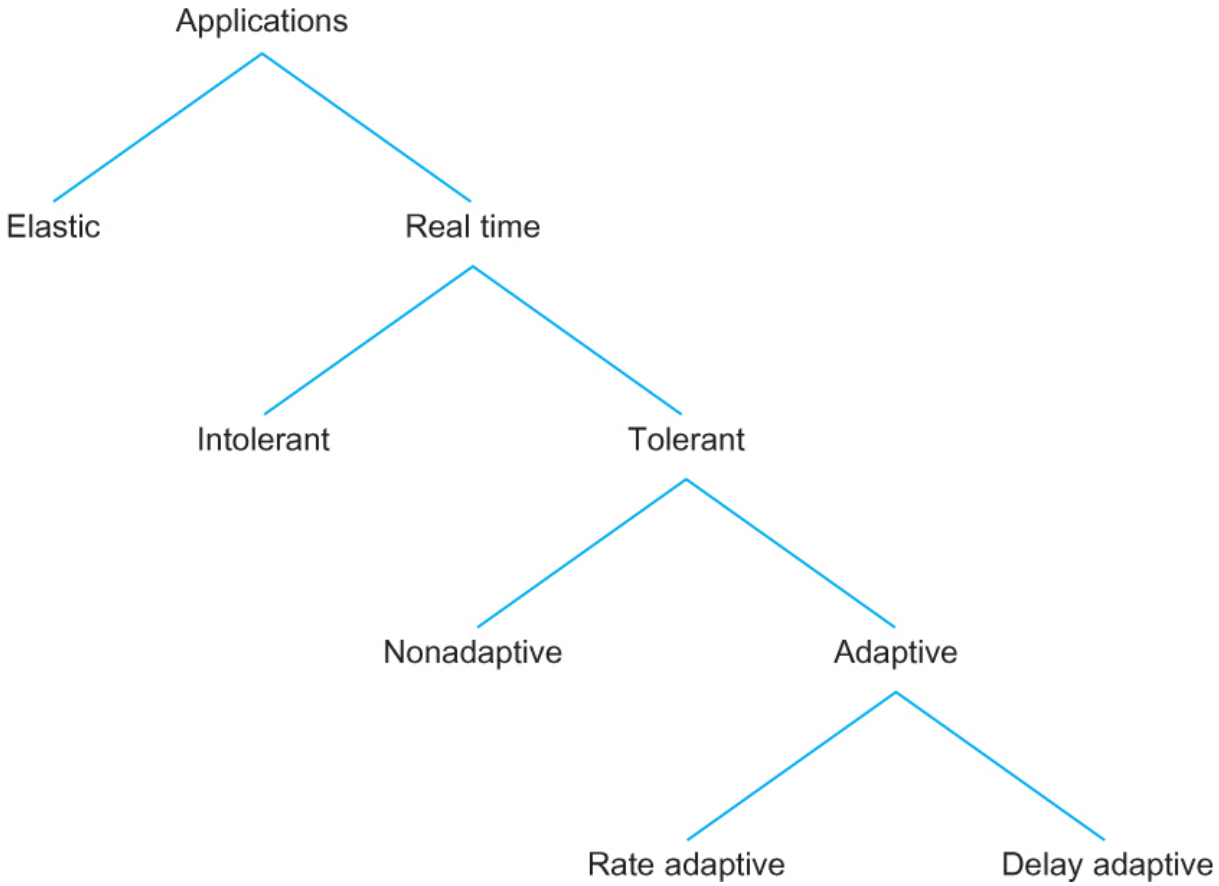
- Taxonomy of Real-Time Applications
  - A second way to characterize real-time applications is by their adaptability.
    - For example, an audio application might be able to adapt to the amount of delay that packets experience as they traverse the network.
  - We call applications that can adjust their playback point *delay-adaptive applications.*

Quality of Service

- Taxonomy of Real-Time Applications
  - Another class of adaptive applications are *rate adaptive.*
    - *For example, many* video coding algorithms can trade off bit rate versus quality.
    - Thus, if we find that the network can support a certain bandwidth, we can set our coding parameters accordingly.
    - If more bandwidth becomes available later, we can change parameters to increase the quality.

Quality of Service

- Taxonomy of Real-Time Applications



Quality of Service

- Approaches to QoS Support
  - *fine-grained approaches, which provide QoS to individual applications or flows*
    - In this category we find “Integrated Services,” a QoS architecture developed in the IETF and often associated with RSVP (Resource Reservation Protocol).
  - *coarse-grained approaches, which provide QoS to large classes of data or aggregated* traffic
    - In this category lies “Differentiated Services,” which is probably the most widely deployed QoS mechanism.

Quality of Service

- Integrated Services (RSVP)
  - Service Classes
    - Guaranteed Service
      - The network should guarantee that the maximum delay that any packet will experience has some specified value
    - Controlled Load Service
      - The aim of the controlled load service is to emulate a lightly loaded network for those applications that request the service, even though the network as a whole may in fact be heavily loaded

Quality of Service

- Integrated Services (RSVP)
  - Overview of Mechanisms
    - Flowspec
      - With a best-effort service we can just tell the network where we want our packets to go and leave it at that, a real-time service involves telling the network something more about the type of service we require
      - The set of information that we provide to the network is referred to as a *flowspec.*
    - Admission Control
      - When we ask the network to provide us with a particular service, the network needs to decide if it can in fact provide that service. The process of deciding when to say no is called *admission control*.

Quality of Service

- Integrated Services (RSVP)
  - Overview of Mechanisms
    - Resource Reservation
      - Mechanism by which the users of the network and the components of the network itself exchange information such as requests for service, flowspecs, and admission control decisions. We refer to this process as *resource reservation*
    - Packet Scheduling
      - Finally, when flows and their requirements have been described, the network switches and routers need to meet the requirements of the flows.
      - A key part of meeting these requirements is managing the way packets are queued and scheduled for transmission in the switches and routers.
      - This last mechanism is *packet scheduling.*

Quality of Service

- Differentiated Services
  - Whereas the Integrated Services architecture allocates resources to individual flows,
  - The Differentiated Services model (DiffServ for short) allocates resources to a small number of classes of traffic.
  - In fact, some proposed approaches to DiffServ simply divide traffic into two classes.

Quality of Service

- Differentiated Services - Approach
  - Suppose that we have decided to enhance the best-effort service model by adding just one new class, which we'll call "premium."
  - Clearly we will need some way to figure out which packets are premium and which are regular old best effort.
  - Use the packets to identify themselves to the router when they arrive. This could obviously be done by using a bit in the packet header—if that bit is a 1, the packet is a premium packet; if it's a 0, the packet is best effort

Quality of Service

- Differentiated Services - Approach
  - If an identifying bit is used, there are two questions we need to address:
    - Who sets the premium bit, and under what circumstances?
    - What does a router do differently when it sees a packet with the bit set?

Quality of Service

- Differentiated Services
  - There are many possible answers to the first question, but a common approach is to set the bit at an administrative boundary.
  - For example, the router at the edge of an Internet service provider's network might set the bit for packets arriving on an interface that connects to a particular company's network.
  - The Internet service provider might do this because that company has paid for a higher level of service than best effort.

Quality of Service

- Differentiated Services
  - Assuming that packets have been marked in some way, what do the routers that encounter marked packets do with them?
  - Here again there are many answers. In fact, the IETF standardized a set of router behaviors to be applied to marked packets. These are called “per-hop behaviors” (PHBs), a term that indicates that they define the behavior of individual routers rather than end-to-end services

Quality of Service

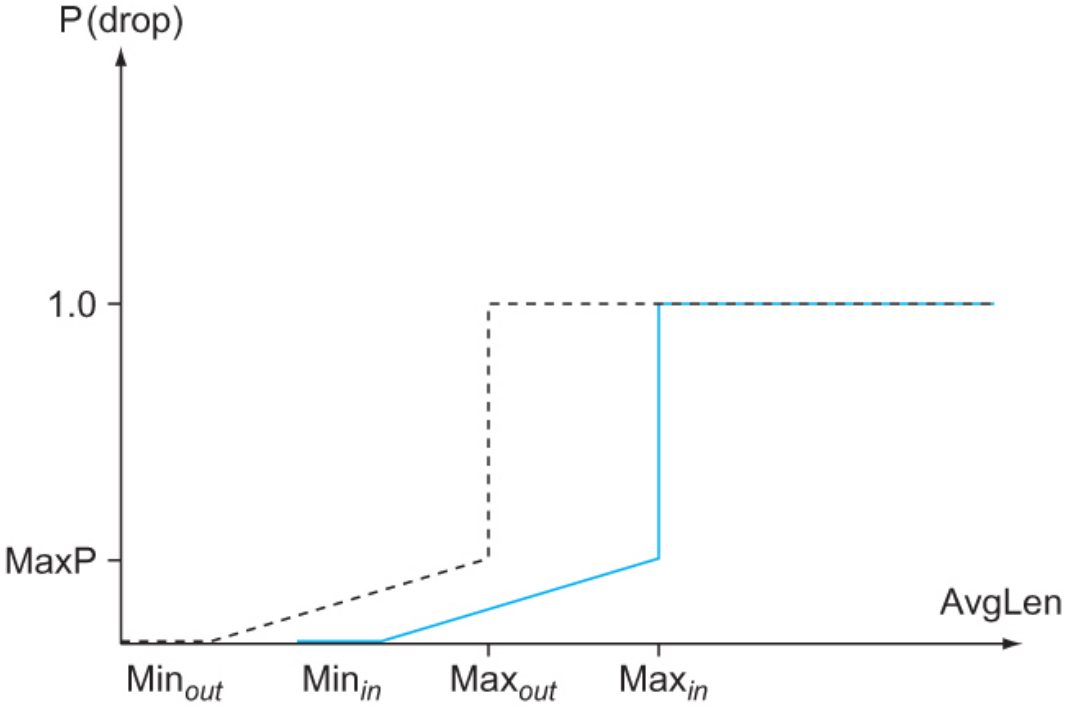
- Differentiated Services
  - The Expedited Forwarding (EF) PHB
    - One of the simplest PHBs to explain is known as “expedited forwarding” (EF). Packets marked for EF treatment should be forwarded by the router with minimal delay and loss.
    - The only way that a router can guarantee this to all EF packets is if the arrival rate of EF packets at the router is strictly limited to be less than the rate at which the router can forward EF packets.

Quality of Service

- Differentiated Services
  - The Assured Forwarding (AF) PHB
    - The “assured forwarding” (AF) PHB has its roots in an approach known as “RED with In and Out” (RIO) or “Weighted RED,” both of which are enhancements to the basic RED algorithm.
    - For our two classes of traffic, we have two separate drop probability curves. RIO calls the two classes “in” and “out”
    - The “out” curve has a lower MinThreshold than the “in” curve
      - Under low levels of congestion, only packets marked “out” will be discarded by the RED algorithm.
      - If the congestion becomes more serious, a higher percentage of “out” packets are dropped
      - If the average queue length exceeds Min_{in}, RED starts to drop “in” packets as well.

Quality of Service

- Differentiated Services
  - The Assured Forwarding (AF) PHB



RED with In and Out drop probabilities