LOYOLA UNIVERSITY CHICAGO

Comp 346 / Comp 446

Quality of Service (QoS)

Dr. William L. Honig
Associate Professor
Department of Computer Science

Is Your Telecom Network Good?

LOYOLA UNIVERSITY CHICAGO

- Communications networks and reliability
  - Telephone
    - System is very reliable
    - Some chance a single call will fail
  - Internet
    - System is less reliable; protocols adapt for problems
    - Big chance a single packet will fail
- Common ground in interest on "Quality" of Service
  - reliability
  - capacity
  - delay
- An area of much work and likely evolution

28.2 Measures of Performance

Measure	Description
Latency (delay)	The time required to transfer data across a network
Throughput (capacity)	The amount of data that can be transferred per unit time
Jitter (variability)	The changes in delay that occur and the duration of the changes

Figure 28.1 Key measures of data network performance.

28.3 Latency or Delay

- Latency
  - specifies how long it takes for data to travel across a network from one computer to another
  - it is measured in fractions of seconds
- Delays
  - across the Internet depend on the underlying infrastructure as well as the location of the specific pair of computers that communicate
- Users care about the total delay of a network
- Engineers usually report both the maximum and average delays
- Also it's custom to divide a delay into several constituent parts
- Figure 28.2 lists the various types of delay

28.3 Latency or Delay

Type	Explanation
Propagation Delay	The time required for a signal to travel across a transmission medium
Access Delay	The time needed to obtain access to a transmission medium (e.g., a cable)
Switching Delay	The time required to forward a packet
Queuing Delay	The time a packet spends in the memory of a switch or router waiting to be selected for transmission
Server Delay	The time required for a server to respond to a request and send a response

Figure 28.2 Various types of delay and an explanation of each.

28.4 Throughput, Capacity, and Goodput

- Capacity
  - often expressed as the maximum throughput
- Throughput
  - a measure of the rate at which data can be sent through the network
  - specified in bits per second(bps)
- Throughput can be measured several ways
- We should specify exactly what has been measured
- There are several possibilities:
  - Capacity of a single channel
  - Aggregate capacity of all channels
  - Theoretical capacity of the underlying hardware
  - Effective data rate achieved by an application (goodput)

28.4 Throughput, Capacity, and Goodput

- Hardware capacity
  - often cited as an approximation of the potential throughput
- The capacity gives an upper bound on performance
  - it is impossible for a user to send data faster than the rate at which the hardware can transfer bits
- Users typically assess the effective data rate
  - that an application achieves by measuring the amount of data transferred per unit time
  - the term goodput is sometimes used to describe the measure
- The goodput rate is less than the capacity of the hardware
  - because protocols impose overhead

28.5 Understanding Throughput and Delay

- The terminology that professionals use to describe network throughput or network capacity can be confusing
  - often they use the terms bandwidth and speed as synonyms for throughput
- Throughput is a measure of capacity
  - not speed
- As an analogy
  - Imagine a network to be a road between two locations and packets traveling across the network to be cars traveling down the road
    - throughput determines how many cars can enter the road each second
    - and the propagation delay determines how long it takes a single car to travel the road from one town to another

28.5 Understanding Throughput and Delay

- Networking professionals have an interesting aphorism:
  - You can always buy more throughput
  - But you cannot buy lower delay
- The analogy to a road helps explain:
  - adding more lanes to a road will increase the number of cars that can enter the road per unit of time
    - but will not decrease the total time required to traverse the road
- Networks follow the same pattern:
  - adding more parallel transmission paths
    - will increase the throughput of the network
    - but the propagation delay will not decrease

28.6 Jitter

- Another measure of networks is used for the transmission of real-time voice and video
  - It is known as a network's jitter
  - It assesses the variance in delay
- Two networks can have the same average delay
  - but different values of jitter
- If all packets that traverse a given network have exactly the same delay, D
  - the network has no jitter
- If packets alternate between a delay of D+ξ and D-ξ
  - the network has the same average delay, but has a nonzero jitter

28.6 Jitter

- Why is jitter important?
  - consider sending voice over a network
- If the network has zero jitter (i.e., each packet takes exactly the same time to transit the network)
  - the audio output will exactly match the original input
- Otherwise, if the network has non-zero jitter
  - the output will be flawed
- There are two general approaches to handling jitter:
  - Design an isochronous network with no jitter
  - Use a protocol that compensates for jitter

28.6 Jitter

- A traditional telephone system uses the first approach:
  - The phone system implements an isochronous network
- Telephone system guarantees the delay along all paths is the same
  - Thus, if digitized data from a phone call is transmitted over two paths
    - the hardware is configured so that both paths have exactly the same delay
- Transmission of voice or video over the Internet takes the second approach
- In Internet, the underlying network may have substantial jitter
  - voice and video applications rely on real-time protocols to compensate for jitter
- Using real-time protocols is much less expensive than building an isochronous network

28.7 The Relationship Between Delay and Throughput

28.7.2 Delay-Throughput Product

- Once a network's delay and throughput are known
  - it is possible to compute the delay-throughput product
  - the delay-throughput product is often called delay-bandwidth product
- You can again think of it as road analogy
- In terms of networks, the number of bits traveling through a network at any time is given by:

$$\textit{Bits present in a network} = D \times T$$

where D is the delay measured in seconds

T is the throughput measured in bits per second

- The product of delay and throughput measures the volume of data that can be present on the network

28.10 Quality of Service (QoS)

- The counterpart of network measurement is network provisioning:
  - designing a network to provide a specific level of service
- This topic is known as Quality of Service (QoS)
- What is QoS?
  - consider the contract between a service provider and a customer
- The simplest contracts define a service
  - by specifying the data rate that the provider guarantees
- More complex contracts define tiered services
  - where the level of service received depends on the amount paid
- Large corporate customers often demand more stringent service guarantees
  - The financial industry typically creates service contracts
    - with bounds on the delay between specific locations

28.11 Fine-Grain and Coarse-Grain QoS

- How can a provider specify QoS guarantees?
- What technologies does a provider use to enforce QoS?
- Figure 28.3 (below) lists the two general approaches that have been proposed for service specification
  - As the figure indicates, the approaches differ in their granularity and whether a provider or a customer selects parameters

Approach	Description
Fine-Grain	A provider allows a customer to state specific QoS requirements for a given instance of communication; a customer makes a request each time a flow is created (e.g., for each TCP connection)
Coarse-Grain	A provider specifies a few broad classes of service that are each suitable for one type of traffic; a customer must fit all traffic into the classes

28.11 Fine-Grain and Coarse-Grain QoS

28.11.1 Fine-Grain QoS and Flows

- The designers assumed a connection-oriented data network modeled after the telephone system:
  - when a customer needed to communicate with a remote site, the customers would create a connection
- Designers assumed a customer would issue QoS requirements for each connection
  - and a provider would compute a charge according to the distance spanned and QoS used
- The phone companies incorporated many QoS features in the design of Asynchronous Transmission Mode (ATM)
  - ATM did not survive and providers do not generally charge for each connection
    - But some of the terminology that ATM created for fine-grain QoS still persists with minor modifications

28.11 Fine-Grain and Coarse-Grain QoS

28.11.1 Fine-Grain QoS and Flows

- Instead of specifying the QoS on a connection
  - the term flow is used
  - A flow generally refers to transport-layer communication
- Figure 28.4 (below) lists four main categories of service
  - that were present in ATM, and explains how they relate to flows

Abbreviation	Expansion	Meaning
CBR	Constant Bit Rate	Data enters the flow at a fixed rate, such as data from a digitized voice call entering at exactly 64 Kbps
VBR	Variable Bit Rate	Data enters the flow at a variable rate within specified statistical bounds
ABR	Available Bit Rate	The flow agrees to use whatever data rate is available at a given time
UBR	Unspecified Bit Rate	No bit rate is specified for the flow; the application is satisfied with best-effort service

28.11 Fine-Grain and Coarse-Grain QoS

28.11.1 Fine-Grain QoS and Flows

- VBR asks users to specify:
  - Sustained Bit Rate (SBR)
  - Peak Bit Rate (PBR)
  - Sustained Burst Size (SBS)
  - Peak Burst Size (PBS)
- ABR service implies sharing:
  - a customer is willing to pay for any amount of service that is available
- When Internet QoS was first considered
  - telephone companies argued that fine-grain services would be needed before the quality of voice telephone calls over a packet network would be acceptable
    - consequently, in addition to the work on ATM, the research community began to explore fine-grain QoS on the Internet
- The research was known as Integrated Services (IntServ)

28.12 Implementation of QoS

- Figure 28.5 (below) illustrates the steps a switch or router uses to implement QoS

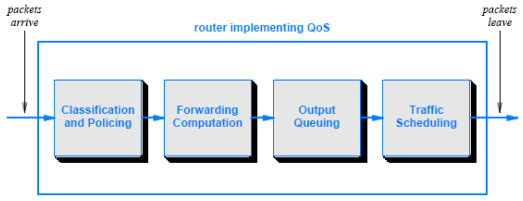

Figure 28.5 The four key steps used to implement QoS.