Introduction to Quality of Service

Andrea Bianco
Telecommunication Network Group
firstname.lastname@polito.it
http://www.telematica.polito.it/

Syllabus

- Applications examples (taxonomy?)
- Quality of Service
- Traffic characterization
- Network architectures characteristics in light of QoS provisioning
- Essential elements (mainly algorithms) to provide QoS

Applications

- Data
  - Generated by single users, by servers, by data centers, by enterprise networks, by P2P architectures, by computing app (e.g. Mapreduce)
  - E-mail, web, messaging, remote login, file transfer, grid computing, ….
- Voice
  - Phone calls, IP calls, …
- Audio
  - Music
- Video
- Multimedia
  - Videoconference, streaming, video distribution

Applications taxonomy

- From the bit rate requirements point of view
  - Elastic applications (opportunistic)
    - If resources are available, elastic applications try to exploit them
    - If resources become scarce, elastic applications may reduce their rate (file transfer)
  - Non-elastic applications (multimedia mostly belong to this category)
    - Require a minimum amount of resources
      - If available, the application works properly
      - If not available, the application is unable to work properly
    - May become slightly elastic if changing the coding scheme

Applications taxonomy

- From the data loss point of view
  - Loss tolerant
    - Uncompressed audio, video, voice
  - Loss intolerant
    - File transfer, e-mail, web, grid computing, compressed audio, video, voice
- From the time sensitivity point of view
  - Not sensitive
    - File transfer, e-mail, web, grid computing
  - Very sensitive (100ms)
    - Phone
  - Sensitive (few s)
    - Streaming

Multimedia real-time and streaming

- Real time applications
  - Two or more users interact in real time
  - Low delay fundamental (a delayed packet is equivalent to a lost packet)
  - Required bit-rate **may** be significant depending on whether video is involved
  - May be robust to (limited) packet losses depending on the compression level
- Streaming applications
  - No real time requirements
  - May tolerate packet delays if initial delay large (buffering)
  - Required bit-rate **may** be significant depending on whether video is involved or not
  - May be robust to (limited) packet losses depending on the compression level

Multimedia streaming

- Streaming
  - Multimedia file stored at the source
  - Sent to the receiver
    - **Buffering** at the receiver
  - File play-out starts when the file transfer is under way
  - Constraint: missing data should reach the receiver before the play-out ends
  - Alternative to file download with later playback (file transfer!)

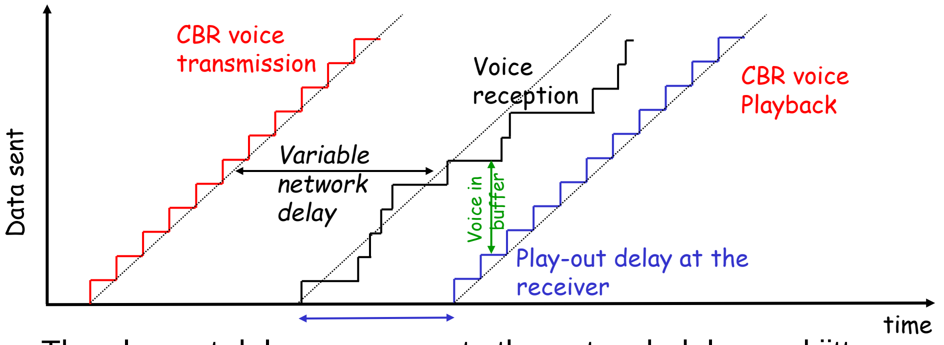
Real-time multimedia: Internet Phone

- Voice as input: sounds and silence period alternate
- Packets generated at a constant rate or when the source emitting power is above a given threshold :
  - E.g.: 20 ms of voice sample at 8kb/s
- Packets are delayed (should be compensated) and lost:
  - Network losses, due to congestion
    - max tolerable may be 10%
  - Losses due to excessive delays (IP datagram received too late for playout)
    - Max tolerable is roughly 400 ms
- Compensation techniques
  - At the transmitter (adaptive coding)
  - At the receiver (buffering)

Reaction to losses, delay and jitter

- Use of a variable bit-rate coder
  - Send small size packets when congestion is detected and the experienced delay is high
  - Send large size packets if the network is lightly loaded
- Quality of reception estimate mechanisms are needed
- The transmitter bit rate should be controlled according to:
  - Instantaneous and/or average loss rate
  - Absolute or relative delay
  - Delay jitter

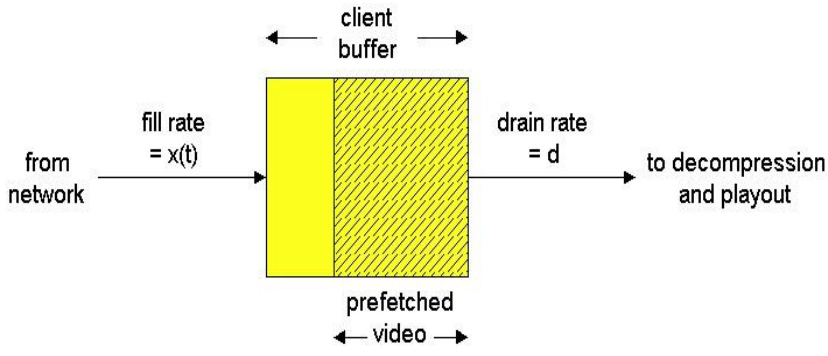
Delay and jitter compensation: buffering at the receiver



- The play-out delay compensate the network delay and jitter
  - Fixed play-out delay: easier but does not adapt
  - Adaptive play-out delay

Buffering at the client side

- Tradeoff betweeen initial delay (buffer size) and tolerance to network jitter



Fixed or adaptive playout delay

- Fixed play-out delay: easier but does not adapt
  - The receiver plays out each sample exactly q seconds after the sample generation
    - If the sample has timestamp t, it is played out at t+q
    - If the sample is received after t+q, it is considered as lost
  - The value assumed by q
    - Large q: less packets lost, higher delay, more buffering
    - Small q: improved interactivity
- Adaptive play-out delay
  - Minimize play-out delay while keeping low the loss rate
  - Estimate the network delay, to determine the play-out delay
  - Compress or extend the silence periods

In summary

- Different applications have different characteristics and different requirements
- We need to satisfy these requirements taking into account application characteristics
- Quality of service (of experience) provisioning

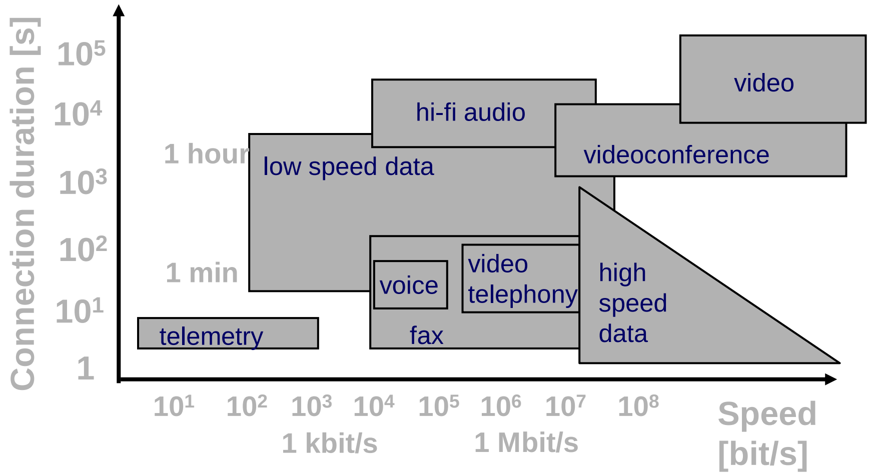
Quality of service/experience

- What is the meaning of quality of service?
- Different definitions
- We use the term mainly to describe performance seen by user traffic
  - Define indices to describe quality
  - Experience is more related to subjective measures
- Examples of indices describing quality of service:
  - Speed (in bit/s), throughput, bit rate, bandwidth
  - Delay (average, percentile, maximum, variance, jitter)
  - Loss probability
  - Error probability
  - Blocking probability
  - Fault probability or availability
  - Recovery time after a fault
  - Many others (time needed to open a connection, costs and tarifs …)

Quality of service

- Different types of traffic require attention to different indices of quality
  - Phone calls (human voice)
    - Guaranteed fixed bit rate
    - Low delays
    - Low blocking probability
  - Data traffic
    - Low or negligible loss probability
- Provide QoS in an heterogeneous environment is more difficult (traffic heterogeneity)
- Provide QoS to unpredictable traffic is more difficult (traffic characterization)

Traffic heterogenity



User traffic characterization *

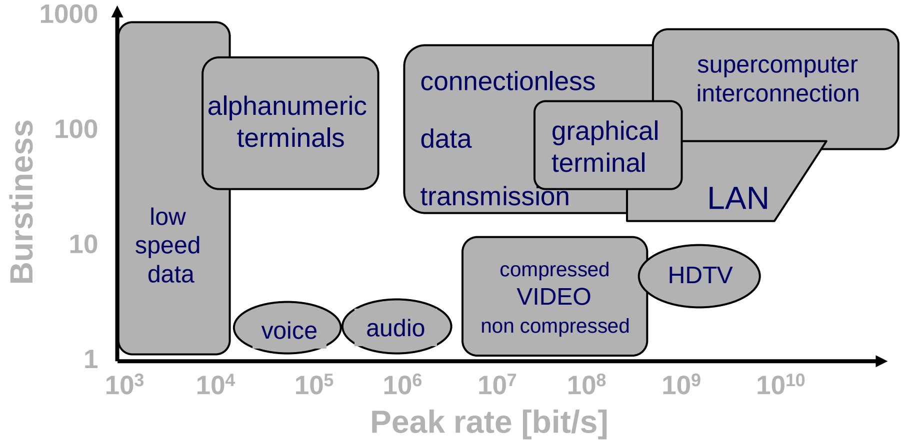
- CBR (Constant Bit Rate) sources:
  - Rate (bit/s)
    - Data size
    - “Perfectly” known
  - Call duration (s)
  - Call generation process
    - Only statistically known

User traffic characterization *

- VBR sources:
  - Average rate (bit/s)
    - Known?
    - Over which period?
  - Peak rate (bit/s) or
  - Burstiness (Peak rate/ average rate)
    - Known (worst case)
  - Burst duration
    - Known?
  - Call duration (s)
  - Call generation process
    - Only statistically known

User traffic characterization

Burstiness= Peak rate/ Average rate



Quality of service

- Networks used as examples
  - Fixed telephone network: POTS
  - Internet
  - B-ISDN
- Let’s start by describing in an informal way the quality of service provided by these networks

POTS

- Characteristics
  - CBR source completely known (generated by the network)
  - Circuit switching
    - Constant, dedicated bit rate ⇒ no congestion
    - Minimum possible delay (only propagation): order of tens of ms (real time)
    - Zero loss probability
  - Error probability smaller than few %
  - Small or negligible blocking probability
- QoS largely independent on other users (apart from blocking probability)
- Network utilization can be really low, user satisfaction very high

Internet

- Characteristics
  - Source behavior unknown
  - Packet switching with datagram service
    - Complete sharing of network resources
    - Bit rate and delay unknown
    - Possible congestion
    - Loss probability may be significant
  - Error probability negligible in wired networks
  - Zero blocking probability
- QoS largely dependent on other users
- Network utilization can be very high, user satisfaction can be very low

B-ISDN

- Intermediate situation
  - Source known (either deterministically or statistically)
  - Packet switching with virtual circuit service
    - May introduce algorithms to control network resources sharing
    - Bit rate and delay negotiable
    - Loss probability negotiable
  - Blocking probability reasonably small
  - Error probability negligible
- QoS dependent on other user behavior and on algorithms used to manage network resources
- Trade network utilization and user satisfaction

Quality of service

- Design problem
  - Given:
    - Network topology (nodes, link speed)
    - Traffic characterization
    - User behaviour
  - Jointly obtain:
    - Guaranteed QoS for each user connection
    - High network utilization
- Without the objective of high network utilization, the problem becomes trivial
  - overprovisioning (power line or water distribution network)

Design to obtain QoS *

- Different time scale (with different level of complexity)
- Network design and planning (resource deployment)
  - Possible re-design and re-planning
  - On the basis of traffic estimates and cost constraints
  - Exploits routing criteria and traffic engineering
- Network management (running a network)
  - Measurements
  - Fault management (protection and restoration)
  - May include simple re-design and re-planning
- Connection management
- Data unit transport

Our definition of QoS

- Assume that a network has been designed and is properly managed
  - Available resources are given
- Mainly study algorithms operating at the following time-scale:
  - Connection management
  - Data unit transport
- Also named **traffic control** problem
- **Must define what is meant by connection. Also named data classification problem**.
- Two different traffic control principles:
  - Preventive control : mainly executed at network ingress, with fairly tight traffic control to avoid congestion insurgence in the network
  - Reactive control: react when congestion situation occur, to reduce or eliminate congestion negative effects

Traffic control: essential elements *

- Connection oriented network
- User-network service interface
  - Traffic characterization
  - QoS negotiation
- Resource allocation (bit rate and buffer)
- Algorithms for traffic control
  - CAC (Connection Admission Control) and routing
  - Scheduling and buffer management (allocation, discard) in switching nodes
  - Conformance verification (policing or UPC: Usage Parameter Control)
  - Traffic shaping to adapt it to a given model
  - Congestion control

Traffic control: connection oriented network

- The connection oriented paradigm permits to know which are the network elements over which traffic control algorithms must be executed (path known)
  - Circuit switching
  - Packet switching with virtual circuit service
- If high utilization is a major objective:
  - Packet switching
- As such, the most suited switching technique to obtain QOS is packet switching with virtual circuit service

Traffic control: user-network service interface

- The capability to control the network increases with the knowledge of user traffic. Limiting factor is the complexity.
- Over the service interface
  - Traffic characterization
  - QoS parameters negotiation
- Can be defined on a call basis or on a contract basis
- POTS: implicit, on a contract basis
- Internet: not existing
- Frame relay: negotiable, normally on a contract basis
- B-ISDN: negotiable with traffic contract on both contract and call basis
- Internet extended to support QoS: negotiable through a SLA (Service Level Agreement) mainly on a contract basis

Traffic control: resource allocation

- Main resources:
  - Bit rate over transmission links
  - Buffer
- Resources can be allocated
  - On a contract basis (booking)
  - On a call basis
  - Packet by packet
- Allocation
  - Exclusive (dedicated resource)
  - Shared

Algorithms: CAC and routing

- Routing
  - QoS based path selection to router a connection
- CAC
  - Determine whether to accept a connection or not, depending on
    - The path chosen by the routing algorithm
    - Traffic characterization
    - QoS requests
    - Network status
- Constraints
  - It is not acceptable to destroy or even reduce the quality of service guaranteed to already accepted connections $\Rightarrow$
    - Can be relinquished
  - Connection must be refused to avoid network overload or congestion
- Preventive control (but can become reactive)

Algorithms: scheduling and buffer management

- Scheduling
  - Choice of the data unit to be transmitted among data unit stored in the switch
- Buffer management
  - Allocation (partial/total, exclusive/shared) of memories in the switch
  - Dropping policies
- Mandatory in an heterogeneous environment to support different QOS requests
  - FIFO (First In First Out) or FCFS (First Came First Served) policy with drop-tail discard is optimal in a homogeneous environment
  - Counter for less than 10 pieces at supermarket
- Preventive and reactive

Algorithms: policing e shaping

- Policing (traffic verification)
  - Network control of user behavior to guarantee conformance to traffic characterization
- Shaping (traffic conditioning)
  - User/network adaptation of data traffic to make it conformant to a given characterization
- Mandatory to control user honesty and to adapt traffic which is difficult to generate as conformant a priori
- Where algorithms must be executed?
  - Only at network edge, i.e., when user access network?
  - Multiplexing points modify traffic shape
    - Both at network access and internally to the network (UNI and NNI)
- Mainly preventive, but they can become reactive if QoS level may change over time

Algorithms: congestion control

- Congestion
  - Traffic excess over a given channel (link)
- Can occur due to
  - Short term traffic variability
  - Allocation policies that share resources to increase network utilization
- Congestion effects:
  - Buffer occupancy increase
  - Delay increase
  - Data loss
- Needed to obtain high link utilization
- Must execute at network edge, within the network or….?
- Reactive