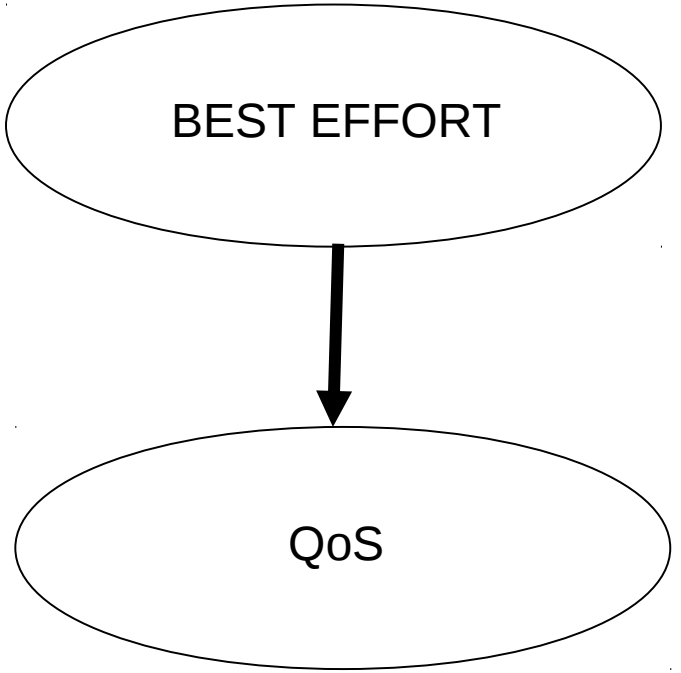
Quality of service in IP networks

- Currently, most Internet services are based on the classic Best-Effort paradigm
- The Best-Effort service does not provide any guarantee on bandwidth, end-to-end delay, packet loss
- The Best-Effort service is very simple and its simplicity has been the most important factor which has determined its worldwide success
- However, the demand of quality associated to the transport of data and media is growing and we are witnessing, in standardizing bodies, in the industry and in the academia, to a significant effort towards a QoS-enabled IP network



Quality of service (QoS)

- Traditionally, QoS is defined in terms of availability, i.e., the percentage time in which the reference system is available and working
- Service Level Agreements are defined as a function of percentage availability, for example 99.975% and of the time required to identify and repair a fault in the connection/equipment
- Usually, a Service Level Agreement is implemented as follows:
  - The system's perimeter to which the SLA applies is defined
  - The availability levels are defined
  - The procedure to measure availability is also defined
  - When the system is in production, its availability is measured
  - If the measured availability is below the minimum threshold, penalties can be applied (they must also be defined in the contract)

Advanced Service Level Agreements

- Real time applications need more advanced SLAs then the simple availability measure
- For example, telephony has strict requirements on delay (related to packet delay and codec delays in the case of VoIP services) and signal quality, related also to packet loss in VoIP services
- More in general, multimedia applications are sensitive to both delay and packet loss
- Advanced SLAs should include packet delay and loss in addition to availability

Real-time and elastic applications

- The quality of some types of applications depends on delay and/or packet loss
- Other applications are robust against delay and/or loss
- For *real-time* applications a packet arriving after a given delay threshold is useless (see for example VoIP)
- Moreover, for real-time applications the retransmission of lost packet is as well useless, as the retransmitted packet is likely to reach its destination after the maximum allowed delay
- Other applications, such as FTP and more in general (even if not always) data transfer, are more robust against delay and they are referred to as *elastic*
- Elastic applications are more robust than real-time applications against packet loss, as they admit retransmission, usually accomplished in an end-to-end fashion through the TCP mechanisms

Real time applications: playback applications (I)

- An important class of real-time applications are *playback* applications (see IETF RFC 1633)
- In a playback application, the source takes some signal, packetizes it, and then transmits the packets over the network
- The network inevitably introduces some variation in the delay of the delivered packets
- The receiver depacketizes the data and then attempts to faithfully play back the signal
- This is done by buffering the incoming data and then replaying the signal at some fixed offset delay from the original departure time
- The term playback point refers to the point in time which is offset from the original departure time by this fixed delay
- Any data that arrives before its associated playback point can be used to reconstruct the signal; data arriving after the playback point is essentially useless in reconstructing the real-time signal

Real time applications: playback applications (II)

- In order to choose a reasonable value for the offset delay, an application needs some "a priori" characterization of the maximum delay its packets will experience
- This "a priori" characterization could either be provided by the network in a quantitative service commitment to a delay bound, or through the observation of the delays experienced by the previously arrived packets
- The performance of a playback application is measured along two dimensions:
  - latency and
  - Fidelity
- Some playback applications, in particular those that involve interaction between the two ends of a connection such as a phone call, are rather sensitive to the latency; other playback applications, such as transmitting a movie or lecture, are not
- Similarly, applications exhibit a wide range of sensitivity to loss of fidelity

Real time applications: playback applications (III)

- There exist two dichotomous classes:
  - *intolerant applications*, which require an absolutely faithful playback (see for example circuit emulation),
  - and *tolerant applications*, which can tolerate some loss of fidelity
- The vast bulk of audio and video applications will be tolerant, but there will be other applications, such as circuit emulation, that are intolerant
- Delay can affect the performance of playback applications in two ways
  - First, the value of the offset delay, which is determined by predictions about the future packet delays, determines the latency of the application
  - Second, the delays of individual packets can decrease the fidelity of the playback by exceeding the offset delay
    - the application then can either change the offset delay in order to play back late packets (which introduces distortion) or
    - merely discard late packets (which creates an incomplete signal)
  - The two different ways of coping with late packets offer a choice between an incomplete signal and a distorted one, and the optimal choice will depend on the details of the application, but the important point is that late packets necessarily decrease fidelity