Services and Applications
(VITMM131)

Service Management and QoS

Attila Vidács
Dept. of Telecommunications and Media Informatics
I.B.228, T:19-25, vidacs@tmit.bme.hu

Outline – 09/03/23

- Service Management
- Service Level Agreement (SLA)
- Quality of Service
  - QoS attributes
  - Human factors, and QoS zones
  - QoS handling in abnormal situations

Service Management (cont'd)



- **Service Management** = functions required to ensure a telecommunications service can be maintained.
- **Tasks** of Service Management:
  - ***Service assignment*** (szolgáltatás hozzárendelés)
    - Establishing information necessary *to assign a service to a customer* (e.g., billing address, location(s), type of service provided).
  - ***Connection management*** (kapcsolat menedzsment)
    - The making or changing of *connections in the network* required as part of a service.
  - ***Fault management*** (hiba menedzsment)
    - The detection, diagnosis and correction of hardware or software failures, which could interfere with the delivery of the service.
  - ***Accounting*** (számlázás)
    - The ability to *monitor service availability* and *usage* in order to ensure appropriate *revenue is collected* from the customer.

Service Management



- **Tasks** of Service Management (cont'd):
  - ***Performance monitoring*** (teljesítmény figyelés)
    - Usage statistics on services, on quality metrics, and on critical resources required to ensure commitments are being kept and to forecast required network growth or equipment upgrades.
  - *Security* (biztonság)
    - Technology and procedures required to ensure *customer privacy* as well as the *safety and integrity of the network*.
- It is desirable to automate as many of these tasks as possible in order *to minimize costs* and *improve the delivery* of services.

Service Level Agreements



- Note: *The deregulation of the industry and the evolution of services have also affected the definition of „what a service is".*
  - „Earlier" the requirements that a service (e.g., basic telephony) had to meet *were prescribed by regulation*, as were the tariffs.
  - Services today may not be strictly defined and so *customers* and *service providers* use contracts to spell out the *requirements* and *what will be paid*.
- An **SLA** (Service Level Agreement – szolgáltatás szintű megállapodás) is a contract between…
  - a service provider and a customer, or
  - between two service providers.

Service Level Agreements (cont'd)



- SLAs are used to specify…
  - service constraints,
  - what **QoS** (Quality of Service – szolgáltatásminőség) will be provided,
  - the cost of a service.
- E.g., SLAs are used to specify…
  - costs;
  - type of connections (e.g., voice, video, protocols);
  - size (e.g., number of channels or bit/second);
  - data reliability (e.g., bit error rate tolerable);
  - responsiveness (e.g., connection set-up time, server response time);
  - availability (e.g., 24 hours seven days a week with no more than *x* seconds of downtime in a year)

Service Level Agreements (cont'd)

- Having an **SLA between service providers** can simplify the Service Management task by partitioning the problem into separate regions.
  - This removes the need to share Service Management information between service providers, but…
  - at the expense of having to monitor the connection points between service providers to ensure conformance to the SLA.

Quality of Service

- Remark: *QoS refers to the **end-to-end QoS** as measured by customers on the outside of the network*.
  - (QoS could be specified for intermediate points in a network or at a juncture between two providers' networks (e.g., in an SLA).)
- Def: ***QoS** can be defined as the quantitative (számszerű) and qualitative (minőségi) characteristics that are necessary to achieve a level of functionality and end-user satisfaction with a service*.

QoS attributes



- **QoS attributes** (QoS attribútumok) tend to fit into two categories: *quality* and *timing*.
- Key ***quality*** *attributes*:
  - ***Fidelity*** *(hűség)*: how faithfully the source content is reproduced.
    - This is usually a function of bandwidth available, sampling granularity, and encoding schemes.
  - ***Loss*** *(vesztés)*: missing packets in a digital stream resulting in missing portions of the source content.
  - ***Corruption*** *(rontás)*: having bits or packets changed resulting in incorrect or modified source content.
  - ***Security*** *(biztonság)*: ensuring the source content is protected from being received by unintended recipients.

QoS attributes (cont'd)



- Key ***timing*** *attributes*:
  - ***Delay*** *(késleltetés)*: also known as latency, is the average amount of time elapsed from the time source material is sent until it is presented at the receiving end.
  - ***Jitter*** *(késleltetés ingadozás)*: also known as delay variability, is the extent to which actual delays deviate from the average.
    - I.e., jitter represents a measure of how much the minimum and maximum delays differ for a single media stream.
  - ***Synchronization*** *(szinkronizáció)*: the difference in delay between more than one media stream, which need to be sent and received together (e.g., sound and video).
  - ***Set-up time*** *(felépülési idő)*: how long it takes to establish access to the service.
    - This is also known as start-up time or start-up delay. (E.g., dial tone in POTS.)
  - ***Tear-down time*** *(lebontási idő)*: how long it takes to discontinue access to the service and free resources to allow another set-up to be initiated.

Typical QoS trade-offs



- Improving one or more QoS attributes often has *consequences for another* (assuming no extra bandwidth and processing power).
  - E.g., Reducing loss is achieved by using protocols that retransmit required packets, which usually increases delay.
  - E.g., jitter and synchronization can be reduced by buffering the incoming data streams, which usually increases delay.
- In general, *delay* seems to be the favorite QoS attribute to suffer.
  - Especially in IP-based packet networks.

Human factors and QoS



- It is often stated that „QoS targets will always be tightening up because users always want it better or faster"…
- BUT human factors are well understood and are unchanging, unlike technology.
  - E.g., the number of frames per second (24 to 30) has been chosen to take advantage of „flicker fusion". This standard has been in place for nearly a hundred years!

Service types as a function of QoS



- There are **four categories of service** in terms of delay, when considering QoS from a *human factors point of view*.
  - The model was used to create the ITU-T Recommendation G.1010 (End-User Multimedia QoS Categories).
- **Perceptual** (észlelés)
  - Limits based on the perceptual limits of the human sensory systems (e.g., auditory or visual).
  - These limits are the shortest in terms of delay, typically within 200 milliseconds.
- **Cognitive** (~megismerés/megértés)
  - Limits based on limits such as short-term memory and natural attention span which range from 0.25 to 3 seconds.

Service types as a function of QoS (cont'd)



- **Four categories of service** (cont'd)
- **Social** (társadalmi)
  - Limits based on social expectations of what a reasonable response time is when a question or request is posed.
  - The user's understanding of how complicated the original request was can temper these limits.
  - Typically these are delays of up to 10 seconds.
- **Postal** („postai")
  - Limits based on expectations of delivery to another person of things like mail or fax.
  - Expectations range from tens of seconds to several minutes and, in some cases, hours.
  - In contrast with the Social category the response for these service types is generally perceived by a person other than the sender, which is one of the reasons for more relaxed timing needs.

Service types as a function of QoS (cont'd)



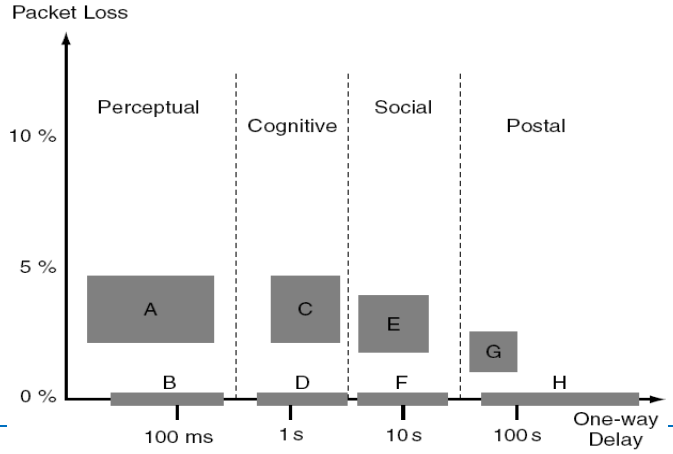
- Both the *Perceptual* and *Cognitive* categories are *neurologically based* and are therefore *valid across all demographics*.
- The *Social* and *Postal* categories may be subject to cultural and experience-related variation.

- In human terms, the reproduction of the source material must be either **precise** (precíz) (i.e., error free) or can be **forgiving** (elnéző) (i.e., a low number of errors may be of no consequence).
  - I.e., *precise* usually corresponds to *digital* source (e.g., bank account number). *Forgiving* usually corresponds to *analogue* source (e.g., favourite colour).

Service types as a function of QoS (cont'd)



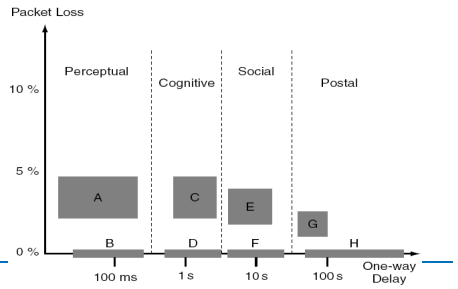
- The *four QoS delay* categories along with the *two types of reproduction* (i.e. precise and forgiving) yield **eight target zones** in terms of *delay* and *loss*:



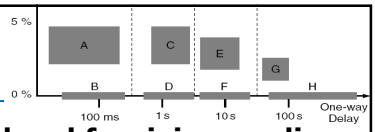
QoS target zones



- **Target areas A, C, E,** and **G** are for source media that can tolerate **some loss** (e.g., *analogue*) and where some *trade-off of loss versus delay* may be possible.
- **Targets B, D, F,** and **H** on the horizontal axis are for source media requiring **0 % loss** (e.g., *digital*) and where there can be *no compromise on loss versus delay*. Delay is the only QoS attribute that can be allowed to vary.



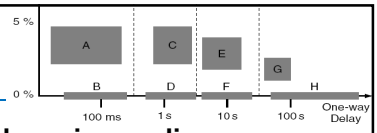
QoS target zones (cont'd)





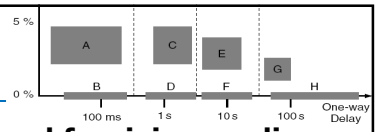
- **Zone A: QoS for perceptual and forgiving media**
  - **Two-way conversational voice and/or video** are the typical service types.
  - The source content is *analogue* in nature and provides a *continuous stream of information*.
  - **Some lost information is tolerated** because *human hearing and visual systems can compensate* for some noise.
  - The delays must typically be **within 200 milliseconds**.
  - Delays in excess of the target introduce noticeable pauses and/or synchronization skew in the conversation (e.g., satellite communications). These effects interfere with *emotional cues*, lead to *user frustration*, and *undermine trust*.
  - Remark: *In some cases users can mitigate some of the effects by adopting a „Ham Radio" protocol in their conversation.*

QoS target zones (cont'd)



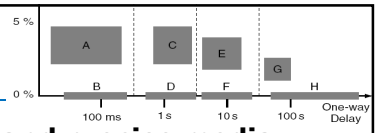
- **Zone B: QoS for perceptual and precise media**
  - Services based on digital media such as *telnet sessions* and *interactive* or immersive *computer games*.
  - The source material content is *digital*. Loss is not acceptable and hence **0 % is the target**.
  - The delays must typically be **within 200 milliseconds**.
  - Delays outside the target reduce the usability of the service by not enabling the user to remain in step with the other end.
  - There is nothing that can be done to mitigate delay in excess of the target.

QoS target zones (cont'd)



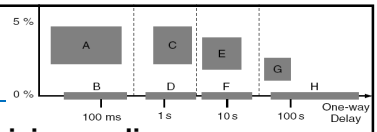
- **Zone C: QoS for cognitive and forgiving media**
  - This is for *one-way analogue services* such as *voice messaging* systems.
  - It is similar to Zone A with respect to loss but the one-way distinction means that the source content *can be delayed more* from the source to the destination without the user noticing.
  - The delay is only apparent at the beginning of the stream.
  - Delays can be in the range of **one second or so**.
  - If the delay in receiving the stream is greater than the target, it can be mitigated by giving the user some feedback within a second to ensure they know the material is coming (e.g., a message or tone confirming the message about to be played back).

QoS target zones (cont'd)



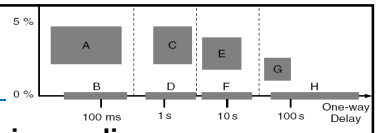
- **Zone D: QoS for cognitive and precise media**
  - *Interactive digital services* such as *Internet browsing* and *E-commerce* on the Web.
  - **Loss is not acceptable**.
  - As with computer interfaces in general, delays need to be **within a few seconds**.
  - As response is delayed beyond the target, users' short-term memory and attention spans are stressed. The task the user is trying to complete becomes more challenging (i.e., user error increases and satisfaction drops).
  - Delays beyond the target zone can be mitigated by providing feedback to the user that their request is pending. This feedback *can help reduce frustration* but it *cannot lengthen short-term memory* nor the fundamental interruption in the flow of the task for the user.

QoS target zones (cont'd)



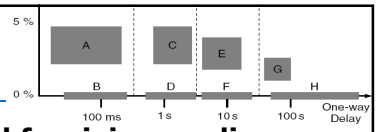
- **Zone E: QoS for social and forgiving media**
  - Services involving *one-way streaming analogue* source material such as *audio* and *video*.
  - The distinction with Zone C is that the content is *more voluminous or continuous* in nature (e.g. Internet radio) and hence more *difficult to re-start or playback*.
  - **Start-up delay can be up to ten or so seconds** given that the duration of the stream is likely to be orders of magnitude longer.
  - The stresses for the user are not a function of neurology but of *expectation based on experience*.
  - The role of feedback about progress to the user while the stream is starting up can significantly mitigate the usability of the service (e.g., „Buffering…").
  - If the delay is roughly an order of magnitude beyond the target then users will suspect the service is not working.

QoS target zones (cont'd)



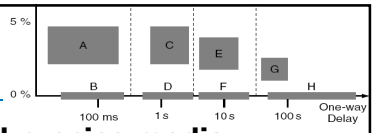
- **Zone F: QoS for social and precise media**
  - Similar to Zone E except that the *source is either digital* or of a *static* (non-streaming and persistent) nature such as *still image downloading* or *FTP downloads* (e.g. software downloading).
  - Unlike Zone E, **loss is not acceptable**.
  - The **start-up delay** is similar to Zone E in the **ten second range**.
  - Start-up delay outside the target zone can be handled identically to Zone E services.
  - The transmission of content in Zone F is usually a finite task. This means the transfer of material will come to an end because the data file is finite in size.
  - *Forecasting completion* of content transfer provides an opportunity for a progress indicator to the end-user to show how close to termination the request is (e.g. count down indicators or completion bars).

QoS target zones (cont'd)



- **Zone G: QoS for postal and forgiving media**
  - *Non-digital content* such as *fax* is the typical service.
  - The end product is static and persistent in nature. This makes errors more noticeable which is why the **acceptable loss is lower**.
  - The acceptable delay is much larger (i.e. anywhere from **20 seconds to a minute and a half**).
  - Unless the sender makes immediate contact with the receiver the delay is not perceived at all.
  - Mitigating delays well beyond the target zone may be achieved by providing feedback to the sender.

QoS target zones (cont'd)



- **Zone H: QoS for postal and precise media**
  - *Digital services* such as *email* have this target.
  - **Loss must be 0 %**.
  - Acceptable delay is hugely variable and provides a very broad target ranging **from a few minutes to hours**.
  - Mitigating delays beyond expectation is something that cannot typically be done in real time but rather by enabling users to *query the status or progress* of their mail message if they wish.

QoS handling in abnormal situations

- All QoS targets must be met under *„normal operating conditions"*
- …BUT it is important how to handle QoS in *abnormal* or *unusual circumstances*.

QoS in overload

- Telecommunications systems are vital tools in response to *emergencies* and natural disasters.
- …BUT, the more severe the scale of an emergency or disaster the more telecommunications networks are likely to be *overloaded*.
- Other sources of overload include
  - „mass calling events",
  - „denial of service attacks",
  - etc.

QoS in overload



- The scale of overload experienced can easily be *orders of magnitude* beyond the normal maximum load.
- In overload situation, there are two fundamental approaches to protect QoS:
  - **increasing the capacity** of the system, and/or
  - providing some form of **load control**
    - e.g., load shedding, load rebalancing, or admission controls.
- Adding capacity is a sound and simple strategy so long as it is *technically feasible*!
- In the absence of load control, *QoS can collapse* when offered traffic goes beyond system's capacity.

QoS in overload



- Principles to protect QoS in overload:
  - where possible, *ensure carrying capacity of the system is greater* than the offered traffic;
  - carried traffic must *always meet QoS* targets;
  - carried traffic should be handled in an *egalitarian* manner;
  - offered traffic should be handled in an *egalitarian* manner so long as it is within the capacity of the system;
  - offered traffic beyond the capacity of the system needs to be *segregated* into *carried* and *non-carried* traffic;
  - segregation of traffic (e.g., load shedding or admission controls) should be done as *close to the source* as possible.

QoS in failure conditions



- Telecommunications networks are designed to minimize any outages due to hardware or software failures.
- When a *service is recoverable* the question becomes *how long does the recovery take* to come into effect
  - E.g. the re-routing of traffic or the swapping in of standby components?
- This *recovery time* introduces a *delay* into the service.
- For any service involving streaming of audio or video or for media in the Perceptual category this recovery delay will usually be an *interruption* and will be *noticed by the end-user*.

QoS exceptions



- <u>Note</u>: If a service involves more than one media type used concurrently, the *strictest target* should be applied to all of them.
  - E.g., in remote collaboration it is needed to keep all media synchronized.
- <u>Note 2</u>: There is nothing preventing an SLA from specifying a QoS target which is *stricter* than the minimum required for human factors reasons.
  - E.g., telemedical Xray, HiFi sound in PC cards…