End-to-end IP Service Quality and Mobility

- Lecture #3 -

Special Course in Networking Technology

S-38.215

vilho.raisanen@nokia.com

Agenda

- Goal of this lecture.
- End-to-end service quality requirements.
- Link to service event characteristics.
- Analysis of mobile services.

Vilho Räisänen

Planned contents & draft schedule

1.	Introduction	Jan 13th
2.	Characteristics of mobile applications	Jan 20th
3.	**Service quality requirement characterizations**	Jan 27th
4.	Challenges of mobile environment	Feb 3rd
5.	Mobility and QoS in GPRS	(Feb 10th)
6.	Mobility and QoS in 3GPP systems	(Feb 17th)
7.	Mobility and QoS with Mobile IP	(Feb 24th)
8.	Mobile IP QoS enhancements	(Mar 3rd)
9.	Edge mobility	(Mar 10th)
10.	Inter-system mobility	(Mar 17th)
11.	End-to-end QoS management	(Mar 31st)
12.	Summary	(Apr 7th)

Dates in parentheses to be confirmed

Goals of this lecture

- Understand the most important service quality requirements of different application types.
  - Service-specific requirement vs. end user experience.
  - Example scenarios.
  - Enumeration of relevant factors.
  - Definitions relevant to engineering of service quality support.
- Define the service quality requirement for mobile services.

Viewing angles at e2e service quality

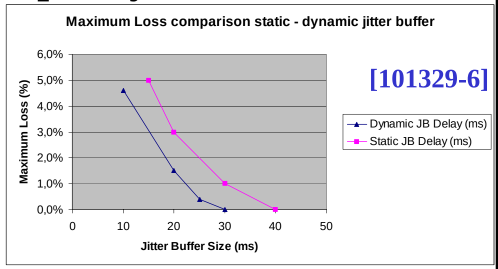
- E2e quality requirements **inherent to service instances**.
  - Services events with step-like utility curve have minimum requirements for e2e quality in the endpoint: CBR voice, …
- **End user experience** of e2e service quality.
  - Service event types with 2-∞ steps in the utility function benefit from better-than-minimum service quality.
- Operator's viewpoint: e2e user experience of service quality level is **engineered**, taking into account **service event type specific requirements**.
  - Engineering is done on aggregate level, including availability, continuity, and blocking aspects.
  - Minimum requirements for endpoints needed.

Scenario #1: G.711 VoIP telephony

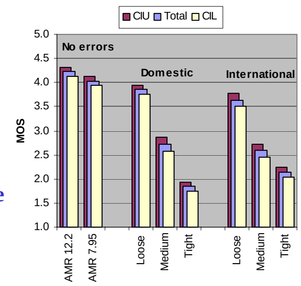
- Service availability and continuity must be sufficiently high.
- Call set-up time must be sufficiently short.
- Telephony interactive => end-to-end delay must be < 400 ms, preferably < 250 ms [Y.1541,101329-2].
  - Delay variation compensation is part of the end-to-end delay budget.
- Sufficient token rate needed for media stream.
- No error concealment in codec => packet loss percentage must be small.
  - Dejittering buffer may give rise to effective packet loss.
  - Packet loss should not be too correlated.

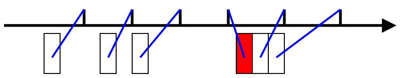
G.711 VoIP telephony, cont'd

- Operation of basic RTP media client:
  – Receive a sample (with timestamp TS).
  – Play out at time TS + Δ.
- Dejittering buffer can use the following operations:
  – Shift buffer in time (affects e2e delay).
  – Resize dejittering buffer (affects e2e delay).



[101329-6]





[Lakaniemi *et al.*, Subjective VoIP speech quality evaluation…, Proc. *ICC2001*]
[Ramjee *et al.*, Adaptive playout mechanisms…, Proc. *Infocom'94*.]

Scenario #2: browsing

- Availability, continuity.
  – Users tend to build preferences towards sites (bookmarked).
  – Availability and continuity important especially for portals.
- Browsing must be interactive.
  – Reaction to clicking a link should follow promptly.
  – It's the total downloading time that counts =>
    - Throughput for large downlink content engineered.
- HTTP run on top of TCP
  – TCP throughput considerations need to be taken into account.

TCP throughput

- Basic TCP: flow control based on advertised window size in ACKs.
- TCP throughput model of Padhye et al:
  - Based on Reno variant.
  - Model with loss rate and RTT as parameters.
- RTT and throughput consistency enhance overall throughput.

$$B(p) \approx \min\left(\frac{W_{\max}}{RTT}, \frac{1}{RTT\sqrt{\frac{2bp}{3}} + T_0 \min\left(1, 3\sqrt{\frac{3bp}{8}}\right) p\left(1+32p^2\right)}\right)$$

B=throughput
p=probability that packet is lost
W=window size
RTT=round-trip delay
b=number packets acknowledged by ACK
T_0=length of the first timeout period

[J. Padhye, V. Firoiu, D. Towsley and J. Kurose, Modeling TCP Reno Performance…, *IEEE/ACM Transactions on Networking*, April 2000]

E2e service quality characteristics

- Aggregate level characteristics:
  - Availability.
- Service instance level characteristics:
  - Continuity.
  - Service instantiation time.
  - Throughput consistency.
- Service event level characteristics:
  - E2e latency.
  - E2e event loss.
  - (BER).
- Note: TSpec parameters are part of service event characteristics!

Requirements for generalized “RSpec”

- Should accommodate current services
  - Conferencing
  - Streaming
  - Interactive
  - Data
- Should also be sufficiently generic to accommodate future services.
- Core set of service quality requirement characteristics should be small.
  - Access technology specific amendments used as necessary.

On temporal correlations

- Latency:
  - E2e latency can be measured for a service event or part thereof -> d_i.
  - Measurement for a set of service events -> time series $\{d_i\}$.
  - Delay jitter = temporal correlation of $\{d_i\}$.
- Packet loss:
  - Service event is lost / not lost -> $\{l_i\}$.
  - Measurement for a set of service events -> $\{l_i\}$.
  - Loss correlation = temporal correlation of $\{l_i\}$.
- Throughput consistency:
  - Temporal correlation of $\{b_i\}$.

[K.Lai, M. Baker, *Measuring bandwidth*, **in** *Proc. Infocom'99*, **p. 235 ff.]**

Availability

- Availability is measured on aggregate service level as per cents of total time.
- Depending on service, availability may be defined in different ways:
  - All service events belonging to the service instance must be present.
  - Some service events may be optional.
- Example: ACME kryptonite detectors
  - Seeing marketing video considered essential for service event, ditto for other service events.
  - Availability = 100% - (V||P||C)
    - V = video unavailability
    - P = payment unavailability
    - C= confirmation unavailability

Service instantiation time

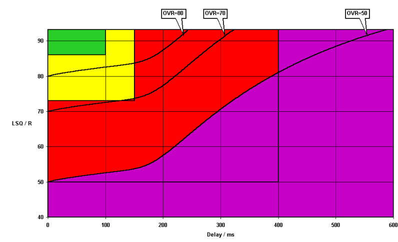
- Some services have typically well-defined service instantiation time
  - Telephony.
    - Guidelines for IP telephony still in the making.
- For other kinds of services, interactivity is important for service instantiation.
  - Browsing.
  - Streaming.
- In all cases, service instantiation time should be engineered for carrier-grade services.
  - Consistency important: e.g., 95% percentile for service instantiation time distribution = X seconds.

Latency

- Inherent latency requirements:
  - PDUs of IP media streams of conferencing applications.
- Designed latency:
  - Interactive IP control traffic (RTSP, SIP, HTTP GET, WAP).
- Designed throughput:
  - Data traffic (FTP, HTML).

	3 (WIDE BAND)	2 (NARROWBAND)			1 (BEST EFFORT)
		2H (HIGH)	2M (MEDIUM)	2A (ACCEPTABLE)	
End-to-end Delay	< 100 ms	< 100 ms	< 150 ms	< 400 ms	< 400 ms
NOTE:	The delay for best effort class is a target value.				

[101329-1]



[101329-7]

Vilho Räisänen

Event loss

- Overall effect of event loss depends on L4 and L2 protocol.
  - TCP: retransmissions hidden from application up to maximum retry limit.
  - UDP: IP-level packet losses visible to end application.
  - L2 reliability.
- Loss can also take place because of bit errors (cf. next slide).
- PDU loss usually degrades service quality of all services.
- For media streams, correlated packet loss degrades end user experienced quality even with loss concealment schemes.
  - VoIP: lower audio quality.
  - Packet video: skipped / distorted frames.

[V. Räisänen, *On end-to-end analysis of packet loss*, **Computer Communications, in press.]**

Vilho Räisänen

BER

- Bit Error Rate (BER) most important for wireless links.
- UDP: checksum computed over header & data.
  - If checksum does not match one computed from L4 SDU, datagram rejected.
- RT data on wireless links: checking of headers needed, retransmissions not good for end-to-end latency.
  - Higher bits of audio samples are more important than lower bits.
  - UDPLite: checksum covers header + predefined number of the highest bits of payload.

[L-A.Larzon, *The UDPLite protocol*, **draft-larzon-udplite03.txt.]**

Throughput consistency

- Minimum token rate required by media streams of real-time IP applications:
  - IP telephony.
  - Streaming.
- Throughput consistency also yields higher overall throughput for TCP-based data transfer.
  - For long data transfers, overall throughput most important => effect of individual bursts of lower throughput.

Estimates for end user experience

- End user experience of service quality can be estimated:
  - With measurements.
  - Via modelling.
- Measurements:
  - Mean Opinion Score (MOS) for telephony [P.800].
- Modelling:
  - Transmission planning: E-model [G.109].
  - Cognitive modelling: PESQ [P.862] etc.
- Can be generalized to other services apart from voice?

[G.109]

Overall Transmission Quality Rating	90 £ R < 100	80 £ R < 90	70 £ R < 80	60 £ R < 70	50 £ R < 60
User's satisfaction	Very satisfied	Satisfied	Some users dissatisfied	Many users dissatisfied	Nearly all users dissatisfied

Data transfer

- Desirable end user experience:
  - Availability designed.
  - Data transfer commences quickly.
  - Overall duration predictable.
- Conclusions:
  - Designed delay for small messages in UL direction.
  - Downlink:
    - Designed delay, relatively consistent.
    - Throughput relatively consistent.
    - Packet loss designed, relatively consistent.

Interactive applications

- Desirable end user experience:
  - Availability relatively high (designed).
  - Replies to requests take place interactively.
  - Duration for downloading predictable.
- Conclusions:
  - Availability needs special attention.
  - Designed delay for small messages in UL direction.
  - Downlink:
    - Designed delay, relatively consistent.
    - Throughput relatively consistent.
    - Packet loss designed, relatively consistent.

Simple SIP services

- Desirable end user experience:
  - Availability high.
  - Service instantiation relatively fast.
  - Some service events must have relatively low latency.
- Conclusions:
  - Availability needs special attention.
  - Designed delay and loss performance for small messages.
  - For large content (e.g. pictures), as with downloading.

Streaming

- Desirable end user experience:
  - Availability high.
  - Service instantiation interactive.
  - Media quality high (more important for audio than for video).
- Conclusions:
  - Availability needs special attention.
  - Designed delay and loss performance for RTSP.
  - Latency for media streams relatively small.
  - Constant token rate desirable
    - Packet loss allowed if retransmissions possible.
  - BER may be allowable for media streams.

Vilho Räisänen

Multimedia conferencing

- Desirable end user experience:
  - Availability high.
  - Service instantiation quick.
  - Media quality high (more important for audio than for video).
- Conclusions:
  - Availability needs special attention.
  - High delay and loss performance for SIP signalling.
  - Latency for media streams small.
  - Minimum token rate required for audio component, desirable for video component.
  - BER allowable for media streams.

Vilho Räisänen

Summary

- Service quality requirements: service type inherent vs. end user experience related.
- Most important service quality characteristics:
  - Availability.
  - Service instantiation time.
  - Throughput consistency.
  - E2e latency.
  - E2e event loss.
  - (BER).
- Temporal correlations.
- Modelling of end user experience.