The Futility of QoS

Yaakov (J) Stein <author@dspcsp.com>

Abstract

I first establish that QoS parameters (such as link continuity, packet loss ratio, and one-way delay) are not interesting in their own right, but only as proxies for QoE. I then present a few well-known examples of explicit functional relationships wherein QoS parameters determine QoE for traditional communications services. Through a sequence of "thought experiments" I demonstrate that such relationships between QoS and QoE can't hold in general for modern rich communications services with on-path processing functions. The conclusion is that QoS may be meaningless for such rich services, and its measurement and maintenance a futile exercise.

Introduction

More and more communications services and services with communications components are being made available essentially without cost, and many consumers take advantage of free WiFi, free email (e.g., Gmail), free voice calls (Skype), free video conferencing (Zoom), free storage (dropbox), free web site hosting (Wix), etc.

The catch, of course, is that these are "best effort" services without any service guarantees (you get what you pay for). While consumers have grown accustomed to free best effort services, they *are* willing to pay for services with quality guarantees (pledged as part of Service Level Agreements, SLAs) when these are needed.

It makes sense that the higher the quality of a service the more consumers will be willing to pay for it. We can imagine that below a certain service quality the service would be expected to be free of charge; at perfect quality (corresponding to a communications medium with no perceivable degradation or delay) the maximum charge (determined by the service's economic value) can be levied; and some cost interpolation holds at intermediate qualities.

Such a charging model requires numerically quantifying service quality.

The Quality of Experience (QoE) of a telecommunications service is variously defined as :

- *The degree of delight or annoyance of the user of an application or service.* [ITU-T P.10]
- *The degree of satisfaction of the user of an application or service. It results from the fulfilment of his or her expectations with respect to the utility or enjoyment of the application or service in light of the user's personality and current state.* [ITU-T P.915]
- *The overall acceptability of an application or service as perceived subjectively by the end-user.* [ITU-T P.10 Amd 2]

An early example of a numeric QoE value is the 5-point **M**ean **O**pinion **S**core for telephony quality voice [ITU-T P.800]. So called toll-quality (MOS=4) was the value of a standard telephony service (less than 5, due to the spectral bandwidth being limited to under 4 kHz), while under MOS=3 is considered unacceptable for most applications. The original cellular voice quality of about 3.5 (due to speech coding employed to reduce digital data rate) was rendered acceptable due to the added benefit of mobility; in fact, although improved speech coding technologies made higher than toll quality possible (via wideband codecs) for cellular, many mobile operators didn't offer it due to consumers having become accustomed to lower quality.

In all variants of its definition, QoE focuses on the holistic subjective service experience and hence requires considering psychophysics, cognitive science, social psychology, and economics in addition to objective criteria. Direct assessment of QoE involves exposing a sufficiently large group of people to the service and averaging their subjective scorings (which is why it is called MEAN Opinion Score). Subjective QoE is thus expensive to evaluate, impractical to gauge in real-time, and impossible to scale-up to a publically available service.

However, the situation is not as bleak as we have just implied. It turns out to be possible to objectively predict the subjective QoE based on received signals. For example, ITU-T's PSQM and PESQ [ITU-T P.862] estimate MOS for telephone grade speech by modeling the human auditory perception system. PEAQ [ITU-R BS-1387] similarly estimates QoE for broadcast quality audio. ITU-T also produced several methods for objective assessment of television picture quality [ITU-T J.148] and PEVQ [ITU-T J.247] for general video. ITU-T's G.1010 [ITU-T G.1010] discusses many further applications, such as voice messaging, streaming audio, web-browsing, bulk data transfer, email, e-commerce, interactive games, SMS, and instant messaging. Yet, these mechanisms are computationally intensive and often require comparing the original non-degraded signal to the degraded one (except for less accurate single-ended methods, such as [ITU-T P.563] for telephony quality audio). This *can* be accomplished, e.g., by periodically sending known signals through the end-to-end communications path, but such mechanisms are at best awkward to implement.

QoS as proxies

The traditional solution to the challenge of pricing service according to quality has been to define Quality of Service (QoS) parameters, which are values that are assumed to correlate with the desired QoE and which *are* readily measured. For packet networks QoS parameters may include **P**acket **L**oss **R**atio, one-way delay, two-way delay or **R**ound **T**rip **T**ime, **P**acket **D**elay **V**ariation, and other measures. In essence the collection of measured QoS values is assumed to act as a proxy for true QoE.

Once such QoS parameters have been defined, several tasks remain:

- how to economically *measure* QoS parameters,
- how to *collect* and/or *report* QoS measurements,
- how to *guarantee* QoS levels,
- how to estimate QoE based on QoS.

The first two tasks are the responsibility of Operations, Administration, and Maintenance (OAM) protocols [RFC 7276]. The third is triggered by detection of approaching SLA nonconformance and dictates mechanisms such as Automatic Protection Switching.

The fourth task is to objectively estimate of the average subjective QoE from the measured QoS parameters. Extensive studies have come up with a plethora of formulas relating N QoS parameters to QoE for particular applications:

(1) $\mathbf{QoE = f\ (application;\ QoS_1,\ \ldots\ QoS_N)}$

where the function f was initially taken as a linear, logarithmic, exponential, or power law in the QoS parameters with free parameters that are optimized based on experiments. More recently neural networks have been trained for such tasks.

An early example of such a mapping is VQmon [ETSI TS 101 329-5 Annex E] which predicts quality of VoIP based on codec type, packet loss statistics, and delay. Interested readers may consult [Fiedler et al].

QoE vs QoS for modern communications services

As can now be understood, the *only* reason to measure QoS parameters, price a service according to QoS parameters, and build mechanisms to ensure QoS parameters, is the relationship between these parameters and subjective QoE. One can envision many parameters that are readily measurable, but only those that impact subjective QoE are significant.

However, the aforementioned relationships have only been established for *traditional* communications services.

By a traditional service I mean a *pure transport* service, characterized by transporting bits

- from site X to site Y (or between N>2 sites),
- with data rate at least R,
- with latency no more than L.

Today's communications services can be much *richer*, i.e., characterized by transporting bits

- from site X to site Y (or between N>2 sites),
- with application information rate at least R,
- with experienced latency no more than L,
- while performing (virtual) network functions on the information along the way.

Network functions that may be performed on information during its end-to-end transport include firewalls and various other security functions, WAN optimization, video transcoding and transrating, various proxies, etc.

I will prove in the following that there is no general relationship of the form (1) for conventional QoS parameters and rich communications services, and that hence QoS and SLAs based on QoS are in general meaningless for such services. And NFV makes the situation even worse.

Our proof will be via a sequence of *thought experiments* (AKA *gedanken experiments*), as are frequently employed in theoretical physics (such as Maxwell's demon, Einstein's elevator, and Schrödinger's cat). In each such thought experiment we will choose a well-known QoS parameter and find a network function for which that QoS parameter is irrelevant (or even counter-intuitive). While I only present here thought experiments relating to packet loss, delay, and loss of link continuity, it is not difficult to construct similar arguments for other QoS parameters.

Packet loss* can be *problematic

Thought Experiment 1 Firewall

It is a universal maxim that packet loss leads to QoE degradation; more specifically, increased PLR means decreased QoE. A firewall, and more generally an Intrusion Prevention System, is a function that intentionally discards packets that it deems to be malicious, thus leading to an increased PLR. Since discarding these packets are in the user's best interest, the subjective QoE by definition increases. Thus with a firewall increased PLR can counter-intuitively lead to increased QoE.

Packet loss can be *meaningless*

Thought Experiment 2 TCP proxy

A TCP proxy is a function placed near the middle of the end-to-end TCP session to improve TCP throughput by reducing the sensed RTT. A TCP proxy terminates the TCP sessions towards both hosts, maintaining the transmitted byte-stream, but not its segmentation into IP packets. Thus, while two packets may enter the proxy and either one or three may emerge. Thus, PLR can be extremely high (or paradoxically even negative) while the QoE improves.

The reader will immediately comprehend a fallacy in this argument – with such re-segmentation packet counts become meaningless and we need to switch to measuring *traffic volume* (the *number of bytes* received irrespective of their packetization).

So, we'll next check if *traffic volume loss* is a useful QoS parameter.

Volume loss can be *meaningless*

Thought Experiment 3 WAN optimization – data compression

Data compression can mean many different things, including:

- lossless data compression,
- data deduplication,
- lossy audio or video compression.

All of these mechanisms decrease the traffic volume without affecting QoE (the last of the three may decrease QoE but the algorithms are designed to make this degradation imperceptible).

So when employing data compression traffic volume is as meaningless as packet counts. The remedy would seem to lie in completely abandoning byte volumes and measuring *Shannon information.* Surely that must be meaningful!

Information loss can be *meaningless*

Thought Experiment 4 WAN optimization – caching server (CDN)

A caching server stores information that may be consumed multiple times but by different consumers. When a flow contains information that has been previously cached near the destination, zero information may be transferred from the information source to the cache, yet the consumer's QoE remains unaffected.

So, even measured loss of Shannon information content can't always be used as an end-to-end QoS parameter!

Synthetic OAM packets can't help

Network engineers may object to our line of reasoning and assert that PLR is certainly well-defined, and the fault lies totally with our measurement methodology! The proper way to measure PLR in such cases is to introduce synthetic OAM packets designed to bypass the computational functionality and thus measure true end-to-end transport PLR!

That argument is completely true, and completely irrelevant!

Recall that we aren't interested in measuring QoS parameters as an academic exercise. The purpose of measuring them is to predict QoE on user traffic. Traffic that does not traverse all the elements of user packets i.e., that is not *fate sharing* with true user traffic, can't assist in the prediction of the QoE of such user traffic!

Delay may be *problematic*

Now that we have seen that packet loss and its derivatives do not always correlate with QoE, let's move on to the second most conventional QoS parameter, namely end-to-end propagation delay.

Of course, many of our previous examples already cast doubt on the possible meaningfulness of delay. If packets are re-segmented as in thought experiment 2 (TCP proxy) then we would need to measure the delay of individual bytes, not packets. If the packet contents changes as in thought experiment 3 (data compression) then it becomes meaningless to measure byte delay as well. If packets are not even sent in the first place, as in thought experiment 4 (caching server) then how can propagation delay even be defined?

But there are even stronger arguments against the prospects of delay correlating with QoE !

Delay can be *meaningless*

Thought Experiment 5   Web browsing

Studies show that users

- are usually satisfied if web pages stabilize in less than 2 seconds
- are usually frustrated if web pages don't stabilize within 8 seconds

where stabilize means that the main graphic elements have reached their final positions (although pictures may not yet be full quality).

The browser used by the end-user to view a web page is in fact a software function that is part of the end-to-end service. Such browsers may run software (e.g., javascript), the code of which is downloaded as part of the data stream. This software may in theory add unbounded run-time before the web page finally stabilizes. Thus, delay from a get request to page stabilization, and hence the QoE, is not uniquely determined by network delay.

Link failure can *improve QoE*

Thought Experiment 6   Rerouting or protection switching

It would seem obvious that path continuity (the statement that information sent is actually received) is even more fundamental to QoE than packet loss and delay. Certainly a failed communications service is useless.

However, consider a rich communications service that initially traverses some set of links and benefits from some software functionality located on server A. Due to a link failure along the path, the service is automatically rerouted and no longer passes through server A, and the software network functionality is repositioned to server B.

Furthermore, for the purpose of the argument let's assume that server B happens to perform the desired functionality better, either due to upgraded software or to more available CPU power and/or memory and/or storage. In such a case the QoE improves.

We conclude that a link failure may actually lead to QoE *improvement* !

How does NFV affect these results?

One may be able to find *workarounds* that mitigate the effects of these results, when the processing functions are static known functions placed at known locations.

But NFV facilitates

- continually developing new functionalities and upgrading existing ones
- dynamically inserting/moving/reconfiguring functionalities

so that

- we can't make assumptions on what network functionalities do, and
- we can't make assumptions as to where network functionalities are placed.

So, with NFV we must pessimistically assume that any of the aforementioned problems may occur anywhere along the end-to-end network path!

Summary

Some of these thought experiments may seem to be contrived or outliers, and I certainly do not claim that none of the known relationships between QoS parameters and QoE may in fact continue to hold for many cases.

However, we have seen that the cardinal QoS parameters of link continuity, packet loss ratio, and one-way delay can not be guaranteed to correlate with QoE for modern rich communications services. This undermines the justification for OAM mechanisms that measure QoS parameters and for control plane mechanisms that strive to maintain QoS parameters in some range of values, driving us towards SLAs directly specifying QoE measured at end-points.

Bibliography

- **ETSI TS 101 329-5 Annex E** Telecommunications and Internet Protocol Harmonization Over Networks (TIPHON) Release 3; End-to-end Quality of Service in TIPHON systems; Part 5: Quality of Service (QoS) measurement methodologies (2000-11)
- **Fiedler et a**l A generic quantitative relationship between quality of experience and quality of service, Markus Fiedler; Tobias Hossfeld; Phuoc Tran-Gia, IEEE Network, 24:2, pp 36 – 41 (March-April 2010)
- **ITU-R BS-1387** Method for objective measurements of perceived audio quality (2001-11)
- **ITU-T G.1010** End-user multimedia QoS categories (2001-11)
- **ITU-T J.148** Requirements for an objective perceptual multimedia quality model (2003-05)
- **ITU-T J.247** Objective perceptual multimedia video quality measurement in the presence of a full reference (2008-08)
- **ITU-T P.10** Vocabulary for performance, QoS and QoE (2017-11)
- **ITU-T P.10 Amd 2** Amendment 2: New definitions for inclusion in P.10 (2008-07)
- **ITU-T P.563** Single-ended method for objective speech quality assessment in narrow-band telephony applications (2004-05)
- **ITU-T P.915** Subjective assessment methods for 3D video quality (2016-03)
- **ITU-T P.800** Methods for subjective determination of transmission quality (1996-08)
- **ITU-T P.862** Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs (2001-02)
- **RFC 7276** An Overview of OAM Tools (June 2014)