Charging for Quality-of-Experience: A New Paradigm for Pricing IP-based Service

Peter Reichl and Florian Hammer

Telecommunications Research Center Vienna (ftw.)
Donaucitystr. 1, A-1220 Vienna, Austria

{reichl|hammer}@ftw.at

Abstract

Over the last couple of years, the interdisciplinary research area of "Internet Economics" has received rapidly increasing attention, especially regarding various proposals for Quality-of-Service (QoS) aware charging of IP services. However, the continuously strong trend towards flat-rate pricing in today's fixed and mobile networks indicates that QoS differentiation does not provide a suitable economic framework for the trade-off between quality delivered by the provider and willingness-to-pay from the customer's side. Therefore, in this paper we argue for a paradigm shift towards Quality-of-Experience (QoE) as an alternative framework for pricing service quality according to the user perception. To this end, we provide a comprehensive survey about various charging schemes for Internet services, discuss the evolution from QoS to QoE and finally describe in detail two different proposals for QoE-aware Internet charging, i.e. (1) an instrumental approach complementing standard charging architectures with an additional module for translating network QoS into perceptual QoS, and (2) a novel mechanism called "Reactive Charging" which is based on direct user feedback for both perceived quality and willingness-to-pay.

1. Introduction

Over the last decade, an almost philosophical discussion has dominated the Internet community as soon as it came to the issue of "Quality-of-Service" (QoS). Whereas the Internet originally is based on best-effort packet transmission, enhanced architectures like IntServ [1] or DiffServ [2] have been proposed to allow for quality discrimination and QoS guarantees. In the context of fixed-mobile convergence, this evolution has by now become relevant also for mobile networks, where, e.g., the 3GPP IP Multimedia Subsystem (IMS) is supposed to bridge the traditional gap between circuit-switched and packet-switched networks and thus will provide the future platform for a QoS-aware All-IP infrastructure [3].

On the other hand, there is no use in providing differentiated service quality levels without price discrimination. Therefore, in parallel the question of how to charge for such QoS-enabled services has led to a significant increase of interest in Internet pricing and tariffing. Eventually, this has resulted in the new interdisciplinary research area of "Internet Economics", which investigates communication networks from an economical rather than from a technical perspective and allows for innovative solutions in network management and control [4].

Traditionally, QoS delivered in a network is specified in terms of network parameters like packet loss rate, link bandwidth, delay, delay jitter etc. Consequently, Internet pricing mechanisms start usually from one or more of these parameters in order to derive useful pricing concepts. As a result, a broad variety of more or less sophisticated QoS-aware charging mechanisms has been proposed during the last years (for an overview see, e.g., [5] and [6] and references therein). In practice, however, hardly any of these approaches has undergone serious attempts to be realized, instead we continuously observe a strong trend towards flat rate pricing which is both relatively easy to be implemented in the network and transparent to the end user.

More recently, with the rise of Voice-over-IP (VoIP), there has appeared a second notion of "service quality" which focuses on the quality experienced by the end user ("Quality-of-Experience", QoE) rather than on pure engineering parameters. This concept is based on performing comprehensive user trials, and the respective metrics are usually expressed in terms of "Mean Opinion Scores" (MOS). A variety of measurement techniques has been developed for that purpose, including subjective methods and instrumental algorithms like [7] or [8].

In this paper we describe a straightforward consequence of this situation, which in our opinion, however, has not been recognized appropriately by the research community so far. We argue that price discrimination should be based more on the quality as perceived by the user (QoE) than on the quality as delivered by the network (QoS). To this end, section 2 briefly surveys the portfolio of quality-aware pricing schemes currently under discussion, and section 3 sketches the evolution from network-based QoS to user-centered QoE. The rest of the paper is devoted to two different proposals for realizing this fundamental paradigm shift: section 4 describes how to integrate an instrumental QoE-monitoring algorithm into a standard QoS-aware charging architecture, and section 5 proposes a novel pricing mechanism based on direct user feedback. Section 6 concludes the paper with an outlook on current and future work.

2. A Brief History of Internet Economics

2.1. Definition and Requirements

"Research on Internet Economics attempts to improve our understanding of the Internet as an economic system..." [9]. This almost classical description formulated by McKnight and Bailey more than a decade ago is still providing a unifying framework for interdisciplinary research activities in various contexts like call admission control, network management and control, routing and load balancing, and last not least pricing and charging for QoS-enabled IP services. In contrast to Internet economy, this research field does not deal with how to make money through the Internet, but rather focuses on "the efficiency of markets as resource allocation mechanisms" [10] in the context of packet networking, based on fundamental game-theoretic concepts like utility, Pareto efficiency, Nash equilibria etc. Among the different applications of the results, the area of charging and pricing is of particular importance. Therefore, in this section we provide a brief survey of important pricing schemes developed over the last couple of years.

Following [11], we distinguish the following three basic requirements for Internet pricing mechanisms:

- *Network efficiency* as determined by maximal utilization of network resources (equivalent to maximizing the total revenue for the provider) is the primary interest of network operators. Thus, the pricing scheme provides an important direct interface for the relationship between user behavior and network status, indicating as different problems as network congestion due to overloaded resources, lack of traffic balance in the network, violation of Service Level Agreements (SLA) by the customer etc.
- *Usability* of a charging mechanism is an important criterion for the end customer and includes the transparency of the pricing mechanism, the predictability of the charges to be paid, the degree of flexibility required for dynamic tariff schemes, the ease-of-use of the charging interface, the reliability of the accounting, the preference for one-stop billing (i.e. one single bill summarizing the different business relationships with multiple service providers) etc.
- *Technical feasibility* is still a major issue, despite of the standardization of charging architectures as flexible as in the case of the IMS [13]. The vast range of technical conditions in a network may induce varying availability, granularity and reliability of the accounting data which themselves have to serve as the primary input to any charging tool. Additionally, the potential complexity of tariffs is strictly limited by requirements on storage space, CPU time, signaling traffic, algorithmic complexity etc.

It is a non-trivial task to design a charging scheme which fulfills all three basic requirements (plus some more less important ones, like standard compliance, incentive compatibility etc., with which we do not deal here explicitly), and any such proposal has to find a balance between those conflicting goals. Therefore, in [14] we have coined the term "*NUT Trilemma*" in order to describe this fundamental trade-off between *Network*, *Usability* and *Technology* which forms the basic starting point for the rest of the paper.

2.2. A Survey on Internet Pricing Mechanisms

The interest in pricing and charging mechanisms specifically for IP-based services started around 1993, when Cocchi et al. formulated a first important pricing paradigm, i.e. *Edge Pricing* [15]. The fundamental idea here is to charge the user only by the first Internet Service Provider (ISP) along the data path, even if also services from other providers are used. The charging information may be transmitted as part of a signaling protocol like RSVP (Resource ReSerVation Protocol). In this way, multilateral contracts are transformed into a series of bilateral agreements, thus reducing the complexity and at the same time enhancing the transparency for the user.

The basic idea of *Congestion Pricing* is to adapt the charge for resource usage to the current status of resource utilization. Hence, if a resource (e.g. a link) is congested, the price to be paid is increasing. This approach has been first proposed by MacKie-Mason and Varian, who in their seminal paper [16] have introduced the concept of "*Smart Market*" as an efficient pricing structure for managing congestion, encouraging network growth and guiding resource utilization according to the users' needs. Under the basic assumption that bandwidth is a scarce network resource (which is certainly still the case also for today's wireless links), ideal prices reflect the resource costs generated by the user, e.g. for network infrastructure, network connection and the social costs of delaying packets sent by other users during periods of congestion. To this end, [16] has proposed to allow the price for sending a packet to vary on a short time-scale, reflecting the actual congestion situation in the network. This is achieved by the use of *auctions*: each packet is expected to carry a bid field in the header, and the packet is admitted to the network if the bid exceeds the current market price, i.e. the marginal cost of transportation. The appropriate auction scheme for this kind of applications is called "Generalized Vickrey Auction" (GVA) where packets are not charged their actual bid, but the (lower) market-clearing price. Note that this type of auctions is "incentive compatible", i.e. it is optimal for the users to bid truthfully (according to their honest evaluation). Later, the Progressive Second-Price Auction mechanism (PSP) due to Lazar and Semret [17] has extended this approach to the case of traffic flows over arbitrarily divisible resources, which afterwards has triggered much more related work on auction schemes like MIDAS (Multilink Distributed Auction Scheme) [18], SAM (Second-chance Auction Mechanism) [19] and the multi-bid auction scheme due to [20], to name but a few.

Resource Pricing as proposed by Gibbens and Kelly [20] aims at achieving the economic efficiency of the smart market by a simple packet marking scheme. Starting from the standard TCP behavior concerning rate adaptation in case of packet loss, they propose to mark each packet traversing an overloaded resource (independently of whether the packet is lost or not). This produces the precise shadow price rate for the end-to-end flows, forcing the end-nodes to react appropriately, e.g. by adapting their sending rates. Note that this mechanism can be viewed as a special case of *Proportionally Fair Pricing* [22] where for the case of elastic traffic, rate allocation is performed in proportion to the users' willingness-to-pay, thus driving the system towards a stable optimal solution with respect to global welfare.

The Resource Pricing approach has been realized in the European IST project M3I (Market-Managed Multiservice Internet, http://www.m3i.org), using the ECN (Explicit Congestion Notification) field of the TCP header for the marking procedure. For further details on this project and on *ECN Pricing* in general we refer to [23].

Another important class of charging schemes is characterized by the fact that they are based on some sort of prior contract between customer and provider (e.g. in the form of a Service Level Agreement SLA) which usually includes a traffic profile binding the customer and a QoS/service specification to be delivered by the provider. A typical example has been the *Expected Capacity Pricing* [24] framework which aims at providing differentiated QoS with high predictability while still running usual best effort. To this end, based on the definition of service profiles for each user, demands are distinguished as either being within the profiles and or outside. Giving priority to traffic which respects the agreed-upon profiles allows the network to offer different levels of service with high predictability.

In contrast, [25] does not classify users but services. The QoS is identical within each service class for all customers, but the more expensive a service class is, the better the offered service. This scheme is based on the nominal bit rate (NBR) as fundamental parameter for a monthly fee. In case of congestion, the system starts to discard packets based on the preference and delay indication bits carried by the packets, and preferably from flows with actual bit rate much larger than the agreed NBR.

Extending this type of SLA-based charging schemes with respect to the integration of multiple time-scales has led to the idea of the *Cumulus Pricing Scheme* (CPS, [26]) which combines a long-term flat fee for SLA-compatible traffic with a mid-term feedback mechanism in case of over- or underusage of the resources, which itself is detected through short-term traffic monitoring. This scheme has been designed in a particularly transparent and user-friendly way, because misbehaving traffic initially yields only some sort warning messages without change of the flat fee agreed upon. Only if the user does not return to compliance with her announced profile over quite some time, the accumulation of the warning messages eventually leads to monetary consequences and/or a renegotiation of the traffic contract.

A somewhat similar proposal related to ECN Pricing is the *Contract and Balancing Process* due to Anderson et al. [27]. Here, the owner of the resource sells contracts to the users, and after usage, the users participate in a balancing process where they make or receive additional payments based on the proportion of marks their traffic has generated, compared to the capacity they have contracted for.

Effective Bandwidth Pricing [28] is another interesting approach. Based on the notion of "effective bandwidth" as a mathematically sophisticated concept which allows to summarize mean, peak and burstiness of the delivered traffic altogether in one single scalar parameter, the customer is asked to declare her expected value of this parameter. The tariff depends on both the customer's declaration and the actually measured effective bandwidth in a way that an honest declaration implies a minimal charge, i.e. the tariff increases linearly if the declared and the measured value for the effective bandwidth do not coincide [29].

So far, all mentioned approaches have in common that they guarantee only relative priority, no absolute QoS. Providing QoS guarantees requires first of all a network architecture which allows for explicit resource reservation, like with Integrated Services architecture [1].Then, [30] combines the concepts of smart market and effective bandwidth for guaranteeing and charging multiple QoS classes (especially for inelastic traffic) by scheduling resources in advance. Here, the solution of a multicommodity flow problem is interpreted in terms of "spot prices" for inserting or extracting traffic at certain nodes, and the marginal system cost for traffic from node A to node B is expressed in terms of only two numbers, i.e. the nodal spot prices for the source and the sink of the flows.

From a practical perspective, [31] has started to focus more on protocol aspects, especially the question how to enhance the RSVP protocol with charging information using the PATH and RESV messages. Price information can then be transmitted using a special field which is initially set to zero, while at each hop the current market price for the requested QoS is added, thus delivering an approximate picture of the market situation upon arrival of the messages at the sender and/or receiver.

Finally, we would like to mention a slightly unconventional idea which will nevertheless play an important role later in this paper. The *Paris Metro Pricing* scheme (PMP) due to Odlyzko [32] is originally based on a tariff which was used in the subway of Paris as well as in some RER trains until roughly 1999: here, trains consisted of 1st class and 2nd class carriages which were absolutely identical except for the price which for the 1st class was double the amount as for the 2nd class. As an interesting (albeit not really surprising) observation, 1st class carriages used to be always less crowded than 2nd class ones. Odlyzko's idea was to translate this scheme into the world of packet networking: thus, the network is partitioned into logical subnetworks which are identical except for the charges. Each user decides for each packet which price to pay, i.e. which subnetwork to use. As an effect, the resources of the 1st class subnetwork will be used less frequently (as they are more expensive), and thus the QoS can be safely expected to be higher than on the 2nd class subnetwork. On the other hand, as soon as the 1st class subnetwork starts to be too much crowded, the difference in QoS will decrease as will the incentive to pay more, and more and more users will return to the 2nd class. In this way, PMP possesses an inherent self-stabilizing moment which guarantees a minimal QoS differentiation between the classes in a very elegant way. Since Odlyzko's original proposal, related work like [33] and [34] has further explored the basic properties of PMP.

3. From Quality-of-Service to Quality-of-Experience

Having surveyed some of the most important Internet pricing mechanisms in section 2, we will now investigate the relationship between pricing and quality of IP-based services in more detail. For the purpose of this paper, we restrict the discussion to Voice-over-IP (VoIP) as a standard example, but of course most of the arguments are valid also in a much wider context.

In a VoIP system we can distinguish three levels of service quality: network QoS, terminal QoS and user-perceived QoS which we will introduce now in some more detail:

- *Network QoS* is usually characterized by the classical triad of packet loss rate, packet delay and delay jitter. Whereas the packet loss rate plays an important role for applications like file transfer, delay and jitter are more crucial for a real-time service like VoIP, especially with respect to the interactivity of a conversation (see e.g. [35] for a more detailed discussion).
- *Terminal QoS* relates to the role of the user terminal, where the following elements contribute to the overall speech quality of the system: the speech codec and its packet loss concealment method, the playout buffer algorithm (increasing the end-to-end delay), echo cancellation and the electro-acoustic interface, i.e., the headset, handset or speakerphone.
- *Quality of Experience* (QoE), also referred to as "perceptual QoS", is defined as "a measure of the overall acceptability of an application or service, as perceived subjectively by the end-user" [36]. Here, the communication context plays an important role in how the users perceive the quality delivered by the network and terminal. The context consists of features of the conversational situation, e.g., the purpose of the call or the user's environment, and human factors, e.g., the technological experience or the age of the user.

Whereas the assessment of network QoS is usually done in a straightforward manner by monitoring one ore more of the mentioned QoS parameters, the case of perceived speech quality is slightly more sophisticated. Here, we distinguish between subjective and instrumental quality assessment: in *subjective quality tests*, test persons either listen to a set of (reference and degraded) speech samples (listening-only tests) or have conversations over a set of connections at different conditions (conversational tests). Subjective quality testing takes a lot of time and effort, e.g., for setting up the test equipment and for carrying out the actual tests. The results of subjective tests are typically given in terms of a Mean Opinion Score (MOS). In *instrumental quality assessment* the perceived speech quality is estimated either based on the actual speech signal (signal-based models) or based on quality related parameters (parameter-based models).

For our objective of monitoring the perceived speech quality for charging purposes, single-sided speech quality assessment methods like "3SQM" (ITU-T Rec. P.563 [37]) or the non-intrusive E-Model (NIEM [38]) seem appropriate. The algorithm described in Rec. P.563 estimates the perceived speech quality by analyzing the received speech signal, whereas the NIEM predicts the perceived quality based on network and terminal parameters and parameters that are extracted from the signal itself (e.g., noise). Note that the recent study [40] has shown that the E-Model overestimates the quality impairment caused by delay, i.e., users seem to accept larger amounts of end-to-end delay.

As already mentioned above, traditional Internet pricing mechanism are usually based on one or more network QoS parameters which by means of one ore more tariff functions are eventually mapped onto some monetary result describing the charge to be paid by the user. It is interesting to note, however, that despite of the huge plethora of such proposals, today there is still a clear tendency towards flat-rate based pricing schemes, thus indicating that this type of charging scheme is – for whatever reason – not accepted by the end user. Therefore, based on the transition from network QoS to QoE sketched above, we argue for a parallel change of the central paradigm from charging for QoS towards charging for QoE. In the remainder of this paper, we discuss two potential approaches for this novel type of mechanisms: (1) a QoE-based instrumental scheme which is transparent to the end user, and (2) a proposal for reactive charging triggered by real-time QoE feedback from the end user.

4. Instrumental QoE-based Charging

Figure 1 sketches a typical architecture for a generic and modular IP charging system including the Internet Charge Calculation and Accounting Subsystem (ICCAS) as e.g. implemented in the M3I project (see [39] for further details). In our context, we focus on the mediation module whose basic purpose is to transform metered data (from the individual monitoring entities), to merge data of different meters, and to substantially reduce the amount of metered data in general. The resulting aggregated monitoring data serve as direct input for the ICCAS where accounting and charge calculation is performed. It is important to note that in the original M3I charging architecture, the Service Directory component has been included because it cannot be assumed that the end user has a full understanding of QoS and thus is able to specify detailed requirements in such a way that it can be used as input to the QoS component in the router.

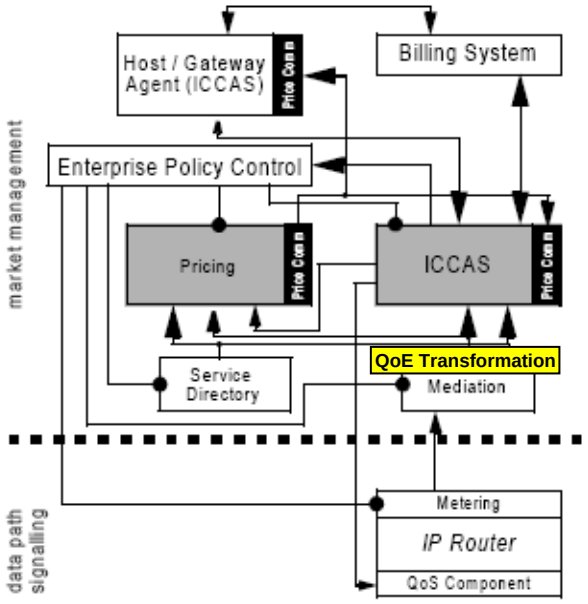

Figure 1: Architecture of the Generic and Modular Charging System of the M3I Project (taken from [39] Fig. 2) and the Additional QoE Transformation Module

The transition of such a QoS-based charging system into our new paradigm requires the integration of an additional

component, the "QoE Transformation Module" on top of the Mediation component as sketched in Fig. 1. This new module is mainly responsible for transforming the metered and mediated QoS data originating in the IP routers into Mean Opinion Scores (MOS), e.g. by using algorithms like the 3SQM mentioned already in section 3 [36]. This transformation must be performed in real-time, enabled by the fact that the mediation component already has significantly reduced and preprocessed the raw metering data delivered by the network. The resulting MOS values are used as estimators for the Quality-of-Experience related to the current network QoS, and as such serve as input to the ICCAS, where the fundamental Service Level Agreements (SLAs) are now formulated in terms of MOS rather than standard network QoS parameters like bandwidth, loss rate, delay or jitter.

Note that this approach is completely transparent to the end user and does not require fundamental changes in the existing charging architecture. The QoE Transformation Module can either be realized as a separate component with interfaces to the mediation module, the ICCAS and the Enterprise Policy Control component which is responsible for the general management and supervision of all charging entities, or it is implemented directly as part of the mediation function (as sketched in Figure 1). The other relevant change concerns the contracts between end user and provider which specify the service quality no longer in terms of network QoS, but rather for QoE, thus increasing significantly the usability dimension of the mechanism. In principle, this approach is suitable for any metering-based IP charging and accounting framework, and its integration e.g. with the IMS charging functions is subject of current work.

5. Reactive Charging

Whereas in the previous section we have discussed how to include a Quality-of-Experience evaluation in the form of an additional sublayer between the entities metering network QoS parameters and the charging module, we will now propose an alternative approach which is no longer transparent to the end user, but on the contrary is explicitly based on direct user feedback in real-time indicating the current level of Quality-of-Experience.

To this end, we focus on the following scenario: let us consider a parsimonious business customer who during a train ride through Upper Austria is having an important voice call over a packet-switched network. For most of the time, the customer is happy with the standard (best effort) quality which of course, due to geographic reasons, is subject to varying loss rates, and therefore the quality of the session as perceived by the customer is changing over time. Now we assume that at specific moments during her call, our customer would prefer to have optimal quality for a certain limited time period, e.g. in order to avoid misunderstandings in negotiations, and would be willing to pay an extra fee for this quality improvement. If the train happens to cross a tunnel at such a moment, the situation could become even worse, and the user would value very much to receive better than best-effort quality.

Figure 2 illustrates the interplay between actual QoE (in terms of a MOS value between 1 and 5) and the user's willingness-to-pay (indicated as a binary variable, i.e. either 0 or 1). Observe that there are three basic cases to be distinguished: for certain periods, the user would be willing to pay, but the QoE is sufficient, during other periods like episode 1, the QoE is low but the user is not willing to pay, whereas the interesting situation is achieved during episode 2 where the QoE is low and the user would be willing to pay for better quality.

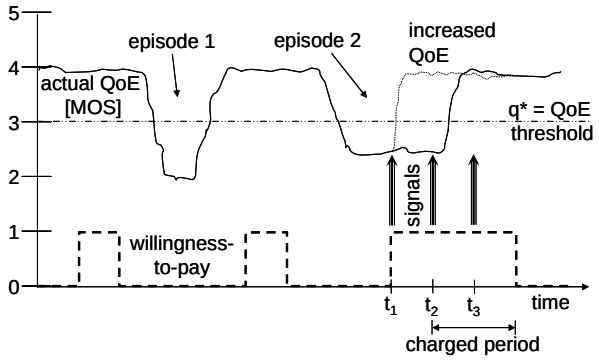

Figure 2: Interplay between QoE (straight line) and Willingness-to-Pay (dashed line)

For this scenario, there are several usability aspects to be considered:

- The mechanism indicating the user's willingness-to-pay (and thus also the charging scheme) has to be transparent and predictable. Therefore, we propose to have *a fixed charge for a fixed time period* of increased quality (e.g. 0.50 € for 2 minutes of better quality).
- The user may only be charged if she is convinced of the increased quality. Therefore, we propose to charge *ex-post*, i.e. after the end customer has evaluated the received Quality-of-Experience positively (see [41] for a more detailed discussion of ex-post pricing). This could also be realized by offering the first high-quality period for free, thus allowing the user to test the increased QoE.
- The signaling interface for both current QoE and willingness-to-pay must be kept extremely simple. Ideally, both signals coincide into one single action. Therefore, we propose as physical interface a simple button to be pressed at the mobile phone, indicating at the same time both that the QoE is bad *and* the user is willing to pay the a priori announced tariff for a better QoE.

Altogether, this leads to the following proposal for a "Reactive Charging" mechanism: as soon as the user wants better quality, she presses a button at her mobile device. Then, for the next 2 minutes she receives better quality without paying for it and has got the chance to evaluate her new perceived quality. After this initial phase, she either signals her wish to prolongate the phase of high quality, e.g. by pressing the button again – in this case she is charged a fixed fee for the subsequent high quality periods until she decides to switch back to best effort. Or she does not press the button again after the initial period, in which case she is not charged anything but switches back to best effort immediately and is not allowed to activate the button again during the next minutes.

In order to put the scheme more formally, assume the time is slotted into equal periods of length Δt, and let $q(t) \in [1, 5]$ be the QoE at time t from the perspective of the user, whereas $w(t) \in \{0;1\}$ describes her willingness-to-pay. Furthermore, we assume that $h(t) \in \{0;1\}$ characterizes the user's happiness with the quality perceived during the (preceding) slot which ends at time t. Then, the signal $\sigma(t) \in \{0;1\}$ sent at time t is defined as follows:

$$\sigma(t) = \big(q(t) < q^*\big) \wedge w(t) \wedge \big(h(t) \vee \big(1 - \sigma(t - \Delta t)\big)\big) \qquad (1)$$

where the QoE threshold q^* relates to a minimum perceived quality the user wants to receive. Hence, equation (1) states that at time t (i.e. the begin of a time slot), a signal is sent if three conditions are fulfilled: (1) the quality is below the acceptable threshold, AND (2) the user is willing to pay for better quality, AND (3) either the user has paid also during the preceding slot and has been happy with the resulting quality improvement OR everything has been ok without additional payment during the preceding slot.

We can illustrate the resulting signaling behavior easily for the example of Fig. 2: if q^*=3.0 (corresponding to the usual MOS level of acceptable quality, dashed-dotted line in Fig. 2), the user would send her first signal in the middle of episode 2 at time t_1 where the QoE is below the threshold and the willingness-to-pay starts being 1. As a result, the QoE immediately increases to a MOS value of around 4 (dotted line in Fig. 2). One slot later, at time t_2, the user is happy with the increased quality which she wants to be maintained during the next slot (willingness-to-pay still 1), therefore she sends a second signal, as well as a third one at time t_3. Afterwards, the user is no longer interested in paying for quality, hence she stops signaling (and the QoE remains on a high level anyway). As the slot between t_1 and t_2 has been offered free of charge in order to allow the user to test the increased QoE, there is a total of 2 slots the user is charged for (as she could not know at time t_3 that the QoE is back on a high level again and has sent another signal there).

To offer improved QoE during the periods of button activation, in principle we could use any mechanism for providing differentiated Quality-of-Service. As an example, we will discuss briefly the case of Paris Metro Pricing (PMP) as introduced at the end of section 2. Assume our IP network is separated into two different subnetworks: the class 1 subnetwork charges a non-negative amount c (per packet/bandwidth unit etc.), whereas the class 2 subnetwork is for free, thus resembling the current best-effort Internet. PMP in its original form is designed on a per-packet level, hence the user is supposed to decide about her willingness-to-pay for each individual packet which is then transported in class 1 if the willingness-to-pay exceeds c, and in class 2 otherwise.

It has already been mentioned in [32] that it is much more efficient not to decide about the class on a per-packet level, but at least for bulks of packets (e.g. one charge unit for the next 1000 packets). We propose to further aggregate this decision process to the level of a packet stream over time, hence the decision between class 1 and class 2 network is made at the beginning of each slot. Therefore, it is necessary to chose the slot duration Δt very carefully: if Δt is very short, both user and network are kept busy with signaling all the time, whereas if Δt is too long, the mechanism loses adaptivity, moreover this might create the somewhat perverse incentive for the network provider to offer deliberately short periods of bad quality from time to time in order to force the user to pay for at least a whole slot of QoE improvement even if the quality has improved again after a short time.

6. Summary and Conclusion

This paper is devoted to the remarkable evolution of the notion of service quality from network QoS characterized by parameters like packet loss rate, delay and jitter, to perceptual QoS (also known as Quality-of-Experience) described in terms of MOS values. We have focused on the implications of this development with regard to the question of how to charge for services with differentiated quality, and we have proposed two different approaches for new charging mechanisms taking QoE explicitly into account.

The general framework of QoE-aware charging as introduced in this paper aims at providing an initial starting point for the research to come in this new direction. Current and future work is investigating in much more detail how service differentiation based on direct user feedback can be achieved in a reliable and well-described way, especially in the framework of the current trend towards network convergence based on All-IP platforms like the 3G IP Multimedia Subsystem.

Acknowledgments

This work has been performed within the strategic ftw.-project U0-SUPRA and has been funded in the framework of the Austrian government's Kplus program.

References

[1] R. Braden, D. Clark, S. Shenker: "Integrated Services in the Internet Architecture: an Overview", IEEE RFC 1633, June 1994.

[2] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, W. Weiss: "An Architecture for Differentiated Services", IEEE RFC 2475, December 1998.

[3] G. Camarillo, M. Garcia-Martin: *The 3G IP Multimedia Subsystem (IMS)*. Wiley 2006.

[4] C. Courcoubetis, R. Weber: *Pricing Communication Networks: Economics, Technology, and Modelling*. Wiley, March 2003.

[5] M. Falkner, M. Devetsikiotis, I. Lambadaris: "An Overview of Pricing Concepts for Broadband IP Networks". IEEE Communications Survey, 2nd Q 2000, pp. 2-13.

[6] B. Stiller, P. Reichl, S. Leinen: "Pricing and Cost Recovery for Internet Services: Practical Review, Classification and Application of Relevant Models". NETNOMICS, Baltzer, Vol. 3, No. 1, March 2001.

[7] *Perceptual Evaluation of Speech Quality (PESQ)*, ITU-T Recommendation P.862.

[8] *Single Ended Method for Objective Speech Quality Assessment in Narrow-Band Telephony Applications*, ITU-T Recommendation P.563.

[9] L. McKnight, J. Bailey (eds.): *Internet Economics.* MIT Press, 1997 (ISBN 0-262-13336-9).

[10] H. Varian: "Differential Pricing and Efficiency". First Monday, 1996.

[11] P. Reichl, P. Flury, J. Gerke, B. Stiller: "How to Overcome the Feasibility Problem for Tariffing Internet Services: The Cumulus Pricing Scheme". Proc. IEEE ICC 2001, pp. 2079-2083, Helsinki, Finland, June 2001.

[12] P. Kurtansky, P. Reichl, J. Fabini, T. Lovric, B. Stiller: "Efficient Prepaid Charging for the 3GPP IP Multimedia Subsystem (IMS)". International Conference on Integrated Design and Process Technology (IDPT'06), San Diego, CA, USA, June 2006.

[13] 3GPP TS 32.260: "Telecommunication management; Charging management; IP Multimedia Subsystem (IMS) charging".

[14] P. Reichl, B. Stiller: "Edge Pricing in Space and Time: Theoretical and Practical Aspects of the Cumulus Pricing Scheme". Proc. 17th International Teletraffic Congress ITC-17, Salvador da Bahia, Brazil, December 2001.

[15] R. Cocchi, D. Estrin, S. Shenker, L. Zhang: "Pricing in Computer Networks: Motivation, Formulation and Example". IEEE/ACM Transactions on Networking, Vol. 1, No. 6, December 1993, pp. 614 - 627.

[16] J. MacKie-Mason, H. Varian: "Pricing Congestible Network Resources", IEEE Journal on Selected Areas in Communications, Vol. 13, No. 7, 1995, pp 1141 – 1149.

[17] A. A. Lazar, N. Semret: "Design and Analysis of the Progressive Second Price Auction for Network Bandwidth Sharing". Telecommunication Systems, Special Issue on Network Economics, 2000.

[18] C. Courcoubetis, M. P. Dramitinos, G. D. Stamoulis: "An Auction Mechanism for Bandwidth Allocation over Paths". International Teletraffic Congress ITC-17, Salvador da Bahia, Brazil, Dec 2001, pp 1163-1174.

[19] P. Reichl, S. Bessler, B. Stiller: "Second-chance Auctions for Multimedia Session Pricing". Proc. MIPS 2003, Naples, Italy, Nov. 2003.

[20] P. Maillé, B. Tuffin: Multi-Bid Auctions for Bandwidth Allocation in Communication Networks. IEEE Infocom 2004, Hong Kong, March 2004.

[21] R. J. Gibbens, F. P. Kelly: "Resource Pricing and the Evolution of Congestion Control". Automatica, vol. 35 (1999), pp. 1969-1985.

[22] F. P. Kelly, A. K. Maulloo, D. K. H. Tan: "Rate control in communication networks: shadow prices, proportional fairness and stability". Journal of the Operational Research Society, 49:237-252, 1998.

[23] B. Briscoe, V. Darlagiannis, O. Heckman, H. Oliver, V. Siris, D. Songhurst, B. Stiller: "A Market Managed Multi-Service Internet (M3I)". Computer Communications 26 (4), pp. 404–414, February 2003.

[24] D. Clark, W. Fang: "Explicit Allocation of Best-Effort Packet Delivery Service". IEEE/ACM Transactions on Networking, Vol. 6, No. 4, August 1998.

[25] K. Kilkki et al.: "Internet Charging Reconsidered". 5th Annual Networld and Interop Engineers Conference. Las Vegas, Nevada, U.S.A., May 1998.

[26] P. Reichl, D. Hausheer, B. Stiller: "The Cumulus Pricing Model as an Adaptive Framework for Feasible, Efficient and User-friendly Tariffing of Internet Services". Journal of Computer Networks, Elsevier, 2003.

[27] E. Anderson, F. P. Kelly, R. Steinberg: "A contract and balancing mechanism for sharing capacity in a communication network". Management Science 52 (2006) 39-53.

[28] F. P. Kelly: "Charging and accounting for bursty connections". In: [9], pp. 253-278.

[29] C. Courcoubetis, F. P. Kelly, V. Siris, R. Weber: "A Study of Simple Usage-Based Charging Schemes for Broadband Networks". Telecommunication Systems, 15 (3-4): 323-242, 2000.

[30] J. MacKie-Mason: "A Smart Market for Resource Reservation in a Multiple Quality of Service Information Network". University of Michigan, September 1997.

[31] G. Fankhauser, B. Stiller, C. Vögtli, B. Plattner: "Reservation-based Charging in an Integrated Services Network". 4th INFORMS Telecommunications Conference, Bocca Raton, Florida, U.S.A., March 1998.

[32] A. Odlyzko: "Paris Metro Pricing for the Internet". Proc. ACM Conference on Electronic Commerce (EC'99), ACM, 1999.

[33] R. Gibbens, R. Mason, R. Steinberg: „Internet service classes under competition". IEEE Journal on Selected Areas in Communications, 18:2490-2498, 2000.

[34] D. Ros, B. Tuffin: "Mathematical Model of Paris Metro Pricing Scheme for Charging Telecommunication Networks". Computer Networks, vol. 46, pages 73-85, 2004.

[35] F. Hammer, P. Reichl, A. Raake: "The Well-Tempered Conversation. Interactivity, Delay and Perceptual VoIP Quality". Proc. IEEE ICC 2005, Seoul, Korea, May 2005.

[36] International Telecommunication Union, "Definition of Quality of Experience", ITU-T Delayed Contribution D.197, Source: Nortel Networks, Canada (P. Coverdale), 2004.

[37] International Telecommunication Union: "Single-ended method for objective speech quality assessment in narrow-band telephony applications", ITU-T Recommendation P.563, May 2004.

[38] International Telecommunication Union: "The E-model, a computational model for use in transmission planning", ITU-T Recommendation G.107, March 2005.

[39] B. Stiller, J. Gerke, P. Reichl, P. Flury: "Management of Differentiated Services Usage by the Cumulus Pricing Scheme and a Generic Internet Charging System". Proc. 7th IEEE/IFIP Integrated Network Management Symposium (IM 2001), Seattle, Washington, U.S.A., pp. 93-106, May 2001

[40] A. Raake: "Predicting Speech Quality under Random Packet Loss: Individual Impairment and Additivity with other Network Impairments", Acta Acustica united with Acustica, Vol. 90, No. 6, pp. 1061-1083, Dec. 2004.

[41] P. Reichl, B. Stiller: "Nil Nove Sub Sole: Why Internet Charging Schemes look like as they do". Proc. 4th Berlin Internet Economic Workshop IEW'2001, Berlin, May 2001.