EQ: A QoE-Centric Rate Control Mechanism for VoIP Calls

CING-YU CHU, New York University
SHANNON CHEN, University of Illinois at Urbana-Champaign
YU-CHUAN YEN, University of Southern California
SU-LING YEH, HAO-HUA CHU, and POLLY HUANG, National Taiwan University

The rising popularity of data calls and the slowed global economy have posed a challenge to voice data networking—how to satisfy the growing user demand for VoIP calls under limited network resources. In a bandwidth-constrained network in particular, raising the bitrate for one call implies a lowered bitrate for another. Therefore, knowing whether it is worthwhile to raise one call's bitrate while other users might complain is crucial to the design of a user-centric rate control mechanism. To this end, previous work (Chen et al. 2012) has reported a log-like relationship between bitrate and user experience (i.e., QoE) in Skype calls. To show that the relationship extends to more general VoIP calls, we conduct a 60-participant user study via the Amazon Mechanical Turk crowdsourcing platform and reaffirm the log-like relationship between the call bitrate and user experience in widely used AMR-WB. The relationship gives rise to a simple and practical rate control scheme that exponentially quantizes the steps of rate change, therefore the name—exponential quantization (EQ). To support that EQ is effective in addressing the challenge, we show through a formal analysis that the resulting bandwidth allocation is optimal in both the overall QoE and the number of calls served. To relate EQ to existing rate control mechanisms, we show in a simulation study that the bitrates of calls administered by EQ converge over time and outperform those controlled by a (naïve) greedy mechanism and the mechanism implemented in Skype.

CCS Concepts: • **Networks → Application layer protocols**;

Additional Key Words and Phrases: VoIP, QoE, Skype, Rate Control, Proportional Fairness

ACM Reference format:
Cing-Yu Chu, Shannon Chen, Yu-Chuan Yen, Su-Ling Yeh, Hao-Hua Chu, and Polly Huang. 2018. EQ: A QoE-Centric Rate Control Mechanism for VoIP Calls. *ACM Trans. Model. Perform. Eval. Comput. Syst.* 3, 1, Article 4 (February 2018), 20 pages.
https://doi.org/10.1145/3170430

All authors have contributed equally and should be considered co-first authors.

This work is supported by the Ministry of Science and Technology of Taiwan, under grant MOST 105-2628-E-002-009-MY3 and grant MOST 106-2918-I-002-034.

Authors' addresses: C.-Y. Chu, NYU Tandon School of Engineering, 6 MetroTech Center, Brooklyn, NY 11201; email: cyc391@nyu.edu; S. Chen, Facebook, 1 Hacker Way, Menlo Park, CA 94025, USA; email: shann@fb.com; Y.-C. Yen, Department of Computer Science, 941 Bloom Walk, Los Angeles, CA 90089-0781; email: yeny@usc.edu; S.-L. Yeh, H.-H. Chu, and P. Huang, National Taiwan University, No. 1, Sec. 4, Roosevelt Rd., Taipei 10617, Taiwan; emails: suling@g.ntu.edu.tw, haochu@ntu.edu.tw, huang.polly@gmail.com.

ACM 2376-3639/2018/02-ART4 $15.00
https://doi.org/10.1145/3170430

1 INTRODUCTION

Effective end-to-end data rate control is one of the classic problems in data networking. In a typical VoIP service, the analog audio signals are converted to a digital stream before being sent through the network. Such conversion is done by the speech codec implemented in the VoIP service. Most speech codecs are capable of compressing the stream into different rates. Increasing the sending rate typically enhances the quality users perceive. The question is whether the rate added to a VoIP call justifies the quality enhanced. Furthermore, given the fixed amount of capacity in a network, extra bandwidth for some calls implies less bandwidth for others. The subtlety we aim to address is that a rate control mechanism needs to satisfy not only an individual but also the bigger society of VoIP callers. This requires a good understanding of (1) how users perceive voice quality at different sending rates across popular codecs and (2) how to allocate bandwidth such that given the same amount of capacity we admit as many users as possible and achieve high overall satisfaction.

The first issue is fundamental to the rate control problem. Early efforts studying the relationship between QoS and QoE (Fiedler et al. 2010; Reichl Tuffin et al. 2010) lack support from real user studies. Later, the user studies in Chen et al. (2012) and Chen et al. (2014) revealed a log-like relationship between the bitrate and user experience in Skype calls. The skepticism, however, comes from whether such relationship holds across voice codecs in general. This study reviews the work on SILK codec (Vos et al. 2010) and investigates further the widely used AMR-WB (3GPP 2011). Both of these codecs are of the wideband genre, which allows voice encoding at variable rates in the range of wideband frequencies. SILK is the de facto codec in Skype, whereas AMR-WB is used to encode phone calls over the cellular network. The 60-participant user study via Amazon's Mechanical Turk (Amazon) reaffirms a significant log-like relationship between the sending rate and the user experience for AMR-WB. This study quantitatively derives a model of sending rate to perceived call quality for the codec. Provided the log-like relationship is persistent across two widely used codecs, the models derived from the user studies provide the utility functions for the second issue addressed in this study.

Inspired by the observed log-like relationship and the theory of proportional fairness, we propose a simple mechanism, i.e., EQ, that exponentially quantizes the steps of rate change. In that, the range of the sending rate is divided into a discrete number of levels. The step between two levels grows exponentially from lower-rate to higher-rate levels. A call's sending rate is controlled to the level no greater than the available bandwidth. The amount of spare bandwidth required to boost a high-rate call to the next level is exponentially higher than that for a low-rate call, which makes it harder to raise the level for high-rate calls. In other words, this mechanism favors the allocation of bandwidth to low-rate calls in the spirit of proportional fairness, i.e., trying to allow all calls a minimal level of service.

Through a formal analysis, we prove that the resulting bandwidth allocation of EQ converges to proportional fairness over time. The proof is driven by the log-like relationship in the bitrate and user experience. That is, given a certain amount of bandwidth, the gain in overall user experience is higher when allocating the bandwidth to low-rate calls. To further support the analysis, we apply the model derived from SILK to conduct the simulation. The results show that the calls, regardless of the initial sending rates, stabilize over time, and the proposed mechanism outperforms a (naïve) greedy mechanism, as well as Skype's own rate control mechanism, in both overall QoE and number of calls served.

The contribution of this work is threefold: (1) measuring AMR-WB via a crowdsourcing platform and reaffirming the user experience to the sending rate relationship; (2) proposing a quantization mechanism that is proportionally fair, simple, and practical to implement; and (3) showing through analysis and simulation that the proposed mechanism optimizes for both performance (overall QoE) and fairness (number of calls).

The rest of the article is organized as follows. Section 2 reviews related research on multimedia networking and psychophysics. Section 3 derives the relationship of QoE to different sending rates. Section 4 presents a rate control mechanism toward sustainable VoIP services. Section 5 details a proof of the proposed mechanism being proportionally fair. Section 6 presents a simulation comparison of the proposed mechanism to Skype's, and Section 7 shows, via simulation, the convergence property of the proposed mechanism. Finally, Section 8 concludes the article.

2 RELATED WORK

The problem of rate control for multimedia services has long been a subject of study. The following discussion provides a literature walkthrough of the subthemes in this study: performance assessment, multimedia rate control, and psychophysics.

2.1 Performance Assessment

Whether a network application provides *good enough* service is traditionally measured by throughput, loss rate, delay, and delay jitter. These measurements are referred to as the Quality of Service (QoS). For years, measured QoS represented the performance of a network application. The quality of the best-effort data, transported via TCP, is measured mainly by throughput, whereas the quality of real-time streaming media, transported via the user datagram protocol (UDP), is measured by additional metrics such as loss rate, delay, and delay jitter.

A growing trend is to measure the quality of network services by the Quality of Experience (QoE). As defined by the International Telecommunication Union (ITU), QoE is "the overall acceptability of an application or service, as perceived subjectively by the end-user." This concept "includes the complete end-to-end system effects" and "those that may be influenced by user expectations and context" (ITU-T 2006).

QoE measurements are difficult to acquire without application-level support (Chen et al. 2006; ITU-T 1996). Measuring responsiveness might require data content analysis, and the mean opinion score (MOS) requires user feedback. Rate adaption mechanisms based on such measurements are inherently difficult to implement. Furthermore, the delay caused by acquiring the measurement might be longer than the granularity of network dynamics, rendering the approach impractical. Recent studies (Fiedler et al. 2010; Kelly 1997; Chen et al. 2006) have attempted to map the objective, network-centric QoS to the direct, user-centric QoE, thus enabling the practical implementation of a user-centric rate adaptation mechanism.

2.2 Multimedia Rate Control

The debate about how multimedia content should be delivered dates back to the 1990s. Bolot and Turletti (1994) proposed a rate control mechanism for video packets that can adapt the output rate of video coders using feedback information. In these early methods, the receiver side fed back to the sender consists of delay and loss information, and the sender controls for delay or bandwidth requirements. The performance-driven approach was the main method for years until the concept of TCP friendliness was developed.

2.2.1 TCP Friendliness. In addition to maximizing QoS or QoE under fluctuating network conditions, another objective of adaptation is to achieve TCP friendliness. With the increasing popularity of real-time multimedia applications that adopt UDP as the underlying transportation protocol, UDP traffic is becoming a major part of the Internet. Because UDP is not aware of network congestion and cannot change its rate accordingly, it might lead to the starvation of TCP traffic or even congestion collapse (Floyd and Fall 1999) if UDP continues to send data using the original rate during network congestion. Therefore, researchers have proposed several TCP-friendly mechanisms for UDP traffic to achieve congestion control.

Inspired by earlier TCP-friendly research (Floyd et al. 2000; Rizzo 2000), some TCP-friendly protocols have been proposed and listed as standards of the Internet Engineering Task Force (IETF) (Handley 2003). The design of congestion control mechanisms in later studies (Yan et al. 2006; Wakamiya 2000) considered application-level QoS and the characteristics of media content. Researchers have subsequently proposed mechanisms that achieve both media and TCP friendliness. Another study (Cicco and Mascolo 2008) has modeled the congestion control algorithm of Skype. These results suggest that Skype adopts a loss-driven mechanism and its sending rate matches the available bandwidth.

Previous research (Bu et al. 2006) has shown that VoIP services are by nature TCP friendly when taking the human factor into account. That study indicates that users prefer to drop calls when the call quality suffers from inferior network conditions. Motivated by this observation, that study presents a probabilistic model to describe the user behavior of dropping calls, and compares the responsiveness of both TCP and user back-off during network fluctuation. Their results show that the spontaneous user back-off mechanism is more responsive than TCP, making VoIP services friendly to TCP traffic.

2.2.2 Proportional Fairness. Because of the considerable amount of real-time multimedia traffic today, being friendly to TCP traffic might not be enough. Another emerging issue is the fairness among interapplication sessions. Max-min fairness (Bertsekas and Gallager 1992) is traditionally chosen as the optimality criterion, but Kelly (1997) proposed a proportional fairness measure that provides an attractive tradeoff between the maximum overall objective function (which represents the overall utility) and user fairness by exploiting the temporal diversity and game-theoretic equilibrium in a multiuser environment. Kelly et al. (1998) suggested a simple algorithm that converges to the proportionally fair rate vector. The basic idea of proportional fairness is as follows: for a proportionally fair allocation, any positive change of a user in the allocation of resources must result in a negative change for the system. This approach should maximize the sum of logarithmic terms, which can be written as

$$\mathrm{P} = \max \sum_{\mathrm{i} \in \mathrm{U}} \ln \mathrm{x_i},$$

where $\mathrm{x_i}$ represents the resource allocated to user i, and U represents the whole user set.

Because of the good properties provided by proportional fairness, researchers have proposed several scheduling schemes to achieve resource allocation in a proportionally fair manner. One study (La and Anantharam 2002) proposed a utility-based rate control mechanism that does not require feedback in an effort to achieve proportional fairness. Xue et al. (2006) presented a resource allocation method for enforcing proportionally fair wireless ad hoc networks. Other studies (Kim et al. 2004; Kwan et al. 2009) have applied the concept of proportional fairness to multicarrier and LTE systems to achieve fair scheduling. The proposed quantization mechanism, with the idea and preliminary results appearing first in Yen et al. (2013), is one such mechanism that aims at allocating network bandwidth proportionally fair. Extended here is a complete evaluation, including (1) the proof of proportional fairness in EQ and (2) the simulations that showcase EQ's convergence properties.

2.3 Psychophysics

The Weber-Fechner law (Boff 1986) provides a plausible explanation of the logarithmic relationships observed between various QoS and QoE metrics. Given a stimulus S and its quantitated perception P, the relationship between S and P is as follows:

$$dP = k \times \frac{dS}{S},$$

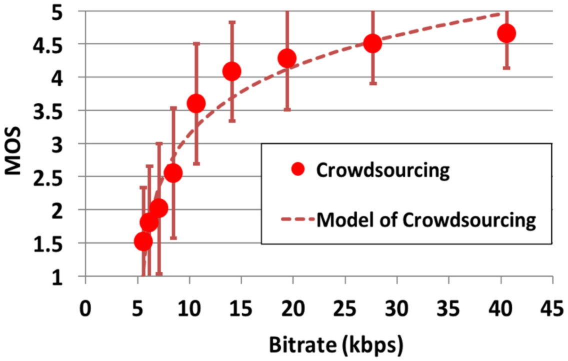


Fig. 1. MOS-bitrate plot of SILK.

where dP and dS are the differences of stimulus and perception, respectively, and k is a constant scale factor. Integrating both sides leads to the following:

$$P = k \times \ln S + c,$$

where c is a constant introduced by integration. The Weber-Fechner law produces plausible results for a wide range of human perception, including hearing, vision, taste, sense of touch and heat, and even temporal, spatial, and numerical cognition (Shen 2003; Scheirer 1998; Moyer and Landauer 1967; Longo and Lourenco 2007).

2.4 Skype User Study

The issue with early works (Fiedler et al. 2010; Kelly 1997; Chen et al. 2006) attempting to map QoS to QoE is that the proposed QoE metrics, e.g., response time or call duration, are each partial to the overall user experience. How these QoE metrics compose the overall user experience remains an open issue.

To better represent overall user experience, Chen et al. (2012) presented the first user study on Skype calls. In that work, the user experience is measured by MOS scores provided by human participants. A log-like relationship is observed between the MOS and the call bitrates. Chen et al. (2014) extended the user study further to investigate how users react to bitrate changes. In that, the same log-like relationship between the MOS and the call bitrate is observed in an independent user study. Despite the effort, one common criticism is that the number of user scores collected per bitrate is small: only 12 user scores in Chen et al. (2012) and 14 in Chen et al. (2014).

To address the challenge of collecting user scores at scale, Yen et al. (2013) attempted to crowdsource the user scores online via platforms such as Amazon Mechanism Turk. The results echoed those of the in-lab user experiment, while the cost and time consumed conducting the crowdsourced user experiment is substantially lower. The MOS-bitrate model, derived from a 60-participant user study, is the state of the art. It confirms again the log-like relationship as shown in Figure 1. The x-axis indicates the sending rate and the y-axis indicates the corresponding MOS. The error bar indicates the range of the average user scores ± one standard deviation. The regression test shows that the log-like relationship is statistically significant with a high R-squared value: 0.95. The relationship is captured in this formula: MOS = $\gamma \times \ln \text{br} - \alpha + \beta$, where br is the bitrate, and the coefficients (α, β, γ) are (4.75, 1.57, 0.95), respectively.

In this work, we take the same crowdsourcing approach and investigate how users rate the AMR-WB-encoded calls. The above model is also used in our simulations to evaluate how the proposed quantization mechanism compares to Skype's rate adaptation mechanism.

Table 1. Sending Bitrates of AMR-WB (kbps)

6.6	8.85	12.65	14.25	15.85
18.25	19.85	23.05	23.85	

3 MODELING AMR-WB CALLS

This section verifies whether the Weber-Fechner Law applies to voice data encoded by the AMR-WB (3GPP 2011) codec. The objective is to reaffirm the log-like relationship between the MOS and call bitrates. Therefore, the Weber-Fechner Law is not a special case, but more general across two widely used codecs, i.e., SILK and AMR-WB.

3.1 Codec

To explore the dependency of voice codecs on the Weber-Fechner Law in voice quality perception, we conducted another user study for AMR-WB. AMR-WB and SILK codecs are perfect targets demonstrating the relationship between user experiences, and the call bitrate is general for a number of reasons: (1) The SILK and AMR-WB codecs both have a significant user base. The former is adopted by Skype, a state-of-the-art native VoIP service. The latter is a well-accepted standard in the public domain, and is widely used in most high-end digital phones and GSM- and 3G-capable devices. Given the increase in data network access through mobile devices such as smartphones and tablet PCs, the portion of the AMR-WB voice data running over the Internet is likely on the rise. (2) Traditional narrowband audio codecs, such as G.729 (ITU-T 2012) used in the public switched telephone network (PSTN), can only provide limited quality of voice because of coding constraints. As the Internet infrastructure continues to evolve to accommodate more bandwidth and allow for a higher level of mobility, the use of narrowband codecs is likely to fade gradually. One limitation to point out is that despite being multirate codecs, AMR-WB is capable of generating streams of nine different bitrates, whereas SILK is a variable bitrate codec that outputs streams of arbitrary bitrates from 5kbps to 40kbps.

3.2 Methodology

Crowdsourcing. Similar to the experiment settings in SILK (Yen et al. 2013), we collect data from real humans through user studies to model the relationship between user experience and call bitrate. The crowdsourcing platform is Amazon Mechanical Turk for its large and diverse user pool. To ease the effort composing the web page for the audio user study, we adopt CrowdMOS (Ribeiro et al. 2011), a tool allowing publishing various types of subjective tests. As a result, participants from all over the world can complete the assigned user tests collaboratively.

Audio Source and Test Tracks. Following the recommendations of ITU-T P.830 (ITU-T 1996), the source material used in this study consists of a number of simple, short, meaningful sentences with no obvious contextual connections. Two female and two male speakers were recruited to record the voice audio samples. The length of the source material was 30 seconds, and the recording quality was at 44.1kHz sampling rate with 16 bits per sample. For the test tracks, we encoded the audio source with all nine supported sending rates listed in Table 1.

Participants. This experiment recruited 60 participants. In addition to asking the participants to score the test tracks, we asked the participants to provide information such as age, gender, device used, and frequency of using VoIP calls to get a profile of the participant population. The information was summarized in Table 2. In this test, we managed to recruit, in slightly more than a week, a wide span of participants, from young to aged, male to female, and frequent to occasional users.

Table 2. Profile of Test Participants

Audio Codec	**AMR-WB**					
Age	Avg.	28.6	Min	19	Max	55
Gender	Male		36	Female		24
Device	Earphone		33	Speaker		27
Frequency of using VoIP	Everyday					14
	At least once a week					22
	At least once a month					24

Each test track was rated using a 5-point MOS, with 5 indicating the most desirable quality and 1 indicating the least desirable quality. The original audio source was presented to the participants at the beginning of the experiment for reference. Other than the reference track, the remaining test tracks were randomly ordered to avoid time-dependent bias. A participant rated no more than 11 audio tracks, which took less than 10 minutes to complete.

Score Calibration. This study used a modified Absolute Category Rating with Hidden Reference (ACR-HR) method (ITU-T 2008) to calibrate scores that might be biased because of fatigue. The main idea of the ACR-HR method is to play a reference track before each test track so that the participants can rate the MOS of each track based on the corresponding reference. The ACR-HR method provides a good way to compare the scores of different tracks but also increases the test duration. As a compromise to the original ACR-HR, which essentially doubles the experimentation time, we modified the ACR-HR to insert one high-quality (the original audio source) reference track for every five test tracks.

As for data processing, a differential quality score (DMOS) was computed between each track and its corresponding reference track (i.e., the reference track (MOSREF) played immediately before the test track (MOSTEST)) using the following formula:

$$\text{DMOS} = \text{MOS}_{\text{TEST}} - \text{MOS}_{\text{REF}} + 5.$$

Because the DMOS might exceed 5, we apply a 2-point crushing function to prevent DMOS from unduly influencing the overall MOS:

$$\text{DMOS}_{\text{crushed}} = \frac{7 \times \text{DMOS}}{2 + \text{DMOS}} \text{ when DMOS} > 5.$$

3.3 Results

Figure 2 shows the MOS-bitrate relationship for tracks encoded by AMR-WB. The x-axis indicates the sending rate and the y-axis indicates the corresponding MOS. The error bar indicates the range of the average user scores ± one standard deviation.

As we cross-compare Figures 1 and 2, it appears that the codec design plays a role in the MOS-bitrate relationship. The MOSs for AMR-WB span in a rather limited range (tightly between 3.5 and 4.5). This could be because AMR-WB targets wireless channels where the available bandwidth is typically thin. The emphasis of the codec design is 15kbps and below, at which point AMR-WB achieves a better MOS, increasing the bitrate. On the other hand, there is little change in MOSs above 15kbps. This can be explained given that the extra bits added above 14.25kbps are in fact all redundant bits (Scheirer 1998).

Despite the flatter shape, as compared to the curve in SILK, the log-like relationship is still pronounced. The regression test shows that the log-like relationship is substantial with a high R-square value—0.90—and the relationship can be modeled as follows:

$$\text{MOS} = \gamma \times \ln(\text{br} - \alpha) + \beta,$$

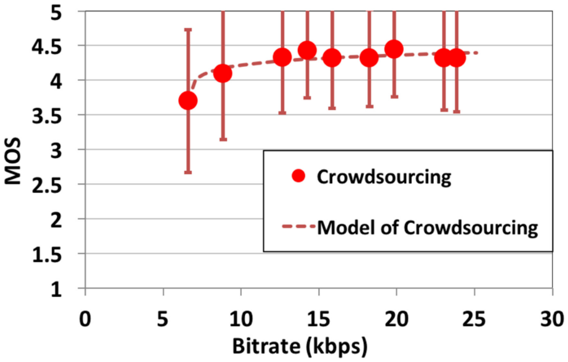


Fig. 2. MOS-bitrate plot of AMR-WB.

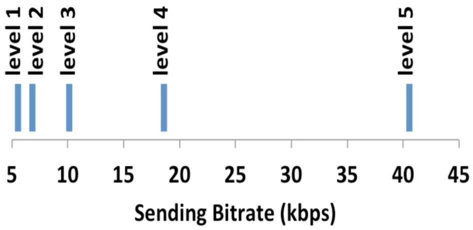


Fig. 3. Exponential quantization of the sending rate.

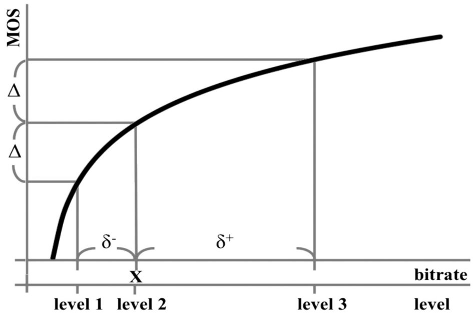


Fig. 4. MOS-bitrate relationship.

where br is the bitrate, and α, β, and γ are codec-dependent coefficients derived from logarithmic fits. Specifically, for AMR-WB, $(\alpha, \beta, \gamma) = (6.57, 4.09, 0.11)$.

4 RATIONAL AND PROPOSED MECHANISM

One implication of the logarithm in the MOS-bitrate relationship of VoIP calls is that the perceptual improvement decreases when the bitrate of the call session is already high. This is remarkable in that, from the traditional QoS point of view, cutting the bitrate for one call and reallocating the same amount elsewhere makes no difference for the network as a whole. However, the MOS-bitrate relationship indicates that the same action could well improve the overall QoE. In that, cutting the bitrate from a high bitrate call might be negligible, while reallocating the same amount to a low bitrate call is much more sensible and satisfying.

A simple, crisp rationale we can draw from above is that a rate control scheme should favor low bitrate calls to achieve the best overall user experience. Conversely, when the bitrate is already high, small increments should be avoided because of the relatively low cost-effectiveness. Thus, an increment should be large enough relative to the current call bitrate, to justify an equal level of change in QoE. This gives rise to the nonuniform quantization in bitrates. More specifically, and according to the log-like relationship observed in VoIP calls, the bitrates are best quantized exponentially.

The proposed mechanism is as simple as exponentially quantizing the bitrates based on the model reviewed in Section 2.4. Figure 3 shows an example for Skype/SILK calls. In that, the bitrates are quantized into five exponentially distributed levels. The bitrates are computed as follows. The

MOS range is first evenly divided into five levels. By transposing the parameters of the model, the bitrate can be expressed as below:

$$br = \alpha + e^{\frac{MOS-\beta}{\gamma}}.$$

Then, apply the MOS value for each level. The formula returns the corresponding bitrate. The difference between two adjacent levels is greater for higher-rate levels and smaller for lower-rate levels. Under this scheme, a call is allowed to transmit voice data at one of the five discrete levels at any given point in time. When the network bandwidth increases, the bitrate increases to the highest level allowed. Conversely, when the network bandwidth decreases, the bitrate drops by one level.

The proposed mechanism is simple and distributed, and it produces bandwidth allocations that improve two important objectives: (1) fairness (the number of calls served) and (2) performance (the accumulated user experience). The analytical proofs and simulations in the next sections would support the claim.

5 ANALYSIS

This section proves that the proposed exponential quantization scheme approaches optimality for both (1) the number of calls served and (2) the accumulated user experience.

5.1 Number of Calls Served

The bitrate allocation of calls changes as the amount of available bandwidth changes and the available bandwidth varies along the fluctuation of background traffic. The following paragraphs present an analysis of the behavior of the proposed scheme in two cases of available bandwidth changes: (1) an increase of B and (2) a decrease of B.

Case 1: Available bandwidth increasing by B. As Figure 4 shows, we set x to be the sending rate of a call before the available bandwidth increases. The sending rate difference between the call's current level and its next higher level is δ^+, and the MOS difference between the call's current level and its next higher level is Δ. Hence, the relationship between δ^+ and Δ in the exponential quantization scheme is[1]

$$\Delta = \ln(x + \delta^+) - \ln(x) = \ln\left(1 + \frac{\delta^+}{x}\right)$$
$$\Rightarrow \delta^+ = x(e^{\Delta} - 1).$$

The amount of increase is B, so only calls with $\delta^+ \leq$ B may be able to improve their sending rates. The proposed model shows that the relationship between the bitrate and the MOS is logarithmic. Thus, exponential quantization implies that Δ is constant for all calls, and only calls with bitrate x $\leq \frac{\mathrm{B}}{\mathrm{e}^{\Delta}-1}$ have the chance to advance to the next level. In other words, this scheme favors calls with lower quality, including calls just starting up. Conversely, calls with higher quality are prevented from hoarding resources. This deterministic discrimination ensures an increasing number of calls served over time.

Case 2: Available bandwidth decreasing by B. When the available bandwidth decreases and a call can no longer sustain its sending rate, it reduces its bitrate by one level according to the exponential quantization scheme. Because existing calls do not detect the change simultaneously, the choice of which calls should sacrifice their bandwidth to fulfill the reduction of B is random. Despite this

[1] We have omitted the coefficients of the exact MOS-bitrate model because they are merely matters of unit conversion and shifting. Instead, what we need to emphasize here is the general relationship between physical and perceptual quantities.

randomness, the scheme still favors calls with lower qualities because the bitrate levels are nonuniformly quantized. For two calls with sending rates x and y before the reduction, denote the amount of backing-off as δ_x^- and δ_y^-, respectively, reacting to the bandwidth change and the MOS difference between the call's current level and its next lower level as Δ. Under exponential quantization,

$$\ln(x) - \ln(x - \delta_x^-) = \Delta = \ln(y) - \ln\left(y - \delta_y^-\right)$$
$$\Rightarrow \frac{\delta_x^-}{x} = \frac{\delta_y^-}{y}.$$

This equation shows that $\delta_x^- > \delta_y^-$ if $x > y$, and vice versa. Thus, a call with higher quality tends to cut its sending rate more, and hence permits low-quality calls from giving up too much of their already scarce bandwidth. By protecting the low-quality calls deterministically, the proposed scheme ensures a reduced number of call drops over time.

Combining the analysis of both cases, the exponential quantization scheme favors low-quality calls. Along with the fluctuation of available bandwidth, the sending rates of calls are balanced when more calls can be sustained, thereby achieving the first goal.

5.2 Accumulated User Experience

This subsection first introduces the concept of *Proportional Fairness* (Kelly 1997) to facilitate the proof in this part. The definition and relevant corollary are enlisted as follows.

Definition 1 (Proportional Fairness). Given a set of N calls, a bandwidth allocation $\vec{x} = (x_1, x_2 \ldots x_N)$ is said to be feasible if the sum of all sending rates $(\sum_{i=1}^{N} x_i)$ does not exceed the total available bandwidth. A feasible resource allocation $\vec{x}$ is defined as proportionally fair if for all feasible allocation $\vec{y}$,

$$\sum_{i=1}^{N} \frac{y_i - x_i}{x_i} \leq 0.$$

Corollary 1. A resource allocation $\vec{x}$ is proportionally fair if and only if, for all feasible allocation $\vec{y}$,

$$\sum_{i=1}^{N} \ln(y_i) \leq \sum_{i=1}^{N} \ln(x_i).$$

Corollary 1 shows that achieving a proportionally fair resource allocation is equivalent to maximizing the sum of the logarithm of sending rates for all calls. According to the proposed model of MOS-bitrate mapping, the sending rate logarithm is proportional to the corresponding MOS value. Thus, proving that the proposed exponential quantization achieves proportional fairness is equivalent to proving that we have achieved the second goal in this study: maximizing the accumulated MOS of all users.

The following paragraphs prove that under the proposed exponential quantization scheme, resource allocation among calls approaches proportional fairness over time. For times t_0 and t_1, at which the available bandwidths are the same, the resource allocation $\vec{y}$ at t_1 is always proportionally fairer than the resource allocation $\vec{x}$ at t_0.

The following proof consists of three stages. First, we define the notations and derive the properties used afterward. Second, we prove the first theorem using the proposed exponential quantization scheme. This theorem guarantees that the total number of level-up events (P) is no less than the number of level-down events (K) during time interval $[t_0, t_1]$. A level change event indicates when a call changes its bitrate level by one. For the case in which one call advances n levels at once and the case in which n calls proceed to the next level, n level-up events are accounted for. Finally, we prove the second theorem, which shows that the allocation of the proposed exponential quantization scheme approaches proportional fairness.

5.2.1 *Definitions and Corollaries.*

Definition 2 (PF Index). For two feasible allocations $\vec{x}$ and $\vec{y}$, define PF($\vec{x}, \vec{y}$) as

$$PF\left(\vec{x},\ \vec{y}\right) = \sum_{i=1}^{N} \frac{y_i - x_i}{x_i},$$

where PF($\vec{x}, \vec{y}$) is the contribution to proportional fairness when the allocation changes from $\vec{x}$ to $\vec{y}$. By Definition 1, if $\vec{x}$ is already proportionally fair, then PF($\vec{x}, \vec{y}$) ≤ 0 for all feasible $\vec{y}$.

Definition 3 (U(n)*).* Let $\tilde{x}$ be an allocation and let $\hat{x}$ be another allocation like $\tilde{x}$ except that only one call i has increased its sending rate n levels higher. Define

$$PF\left(\tilde{x},\ \hat{x}\right) = U\left(n\right).$$

Definition 4 $D(n)$. Let $\tilde{x}$ be an allocation and let $\check{x}$ be another allocation like $\tilde{x}$ except that only one call i has decreased its sending rate n levels lower. Define

$$PF\left(\tilde{x},\ \check{x}\right) = D\left(n\right).$$

Corollary 2. *$U(n) = e^{n\Delta} - 1 \geq 0$, independent of any allocation and any call to raise n levels.*

Proof. Denote the sending rate of call i as x in $\tilde{x}$ and x + δ_n^+ in $\hat{x}$. Definitions 2 and 3 and the proposed MOS-bitrate model show that

$$n\Delta = \ln\left(x + \delta_n^+\right) - \ln\left(x\right) = \ln\left(1 + \frac{\delta_n^+}{x}\right)$$
$$\Rightarrow e^{n\Delta} - 1 = \tfrac{\delta_n^+}{x} = U\left(n\right) \geq 0,$$

where Δ is defined as the MOS difference between two levels. □

Corollary 3. *$D(n) = e^{-n\Delta} - 1 \leq 0$, independent of any allocation and any call to lower n levels.*

Proof. Similar to the proof for Corollary 2. □

Corollary 4. *$|U(n)| \geq |D(n)|$.*

Proof.

$$\left|U\left(n\right)\right| \geq \left|D\left(n\right)\right| \Leftrightarrow e^{n\Delta} - 1 \geq 1 - e^{-n\Delta}$$
$$\Leftrightarrow \frac{e^{n\Delta} + e^{-n\Delta}}{2} \geq \sqrt{e^{n\Delta}e^{-n\Delta}} = 1$$ □

Corollary 5. *$U(a) + U(b) \leq U(a+b)$.*

Proof.

$$U\left(a\right) + U\left(b\right) \leq U\left(a+b\right)$$
$$\Leftrightarrow \left(e^{a\Delta} - 1\right) + \left(e^{b\Delta} - 1\right) \leq \left(e^{(a+b)\Delta} - 1\right)$$
$$\Leftrightarrow \left(e^{a\Delta} - 1\right)\left(e^{b\Delta} - 1\right) = U\left(a\right) \cdot U\left(b\right) \geq 0$$ □

Corollary 6. *$D(a) + D(b) \leq D(a+b)$.*

Proof. Similar to the proof for Corollary 5. □

5.2.1 *$P \geq K$.*

Theorem 1. *In time interval $[t_0,\ t_1]$, the total number of level-up events P is no less than the number of level-down events K.*

Proof. Because the available bandwidths at t_0 and t_1 are the same, the total amount of bandwidth increased and decreased during $[t_0,\ t_1]$ must also be the same. According to the description in

Case 1 of Section 5.1, the proposed scheme favors calls with poor quality. In addition, in Case 2, the decision of a call to decrease its sending rate to accommodate available bandwidth reduction is a random process. Thus, the average bandwidth released by a call when available bandwidth decreases exceeds the average extra bandwidth occupied by a call when the available bandwidth increases. Hence, given that the amount of available bandwidth increases and decreases are the same during $[t_0,\ t_1]$, there are more level-up events than level-down events (i.e., $P \geq K$). □

Lemma 1. *If we denote* q_i *as the difference of levels of call* i *in* $\vec{x}$ *and in* $\vec{y}$, *and denote* $P' = \sum_{q_i>0} |q_i|$ *and* $K' = \sum_{q_i<0} |q_i|$, *then* $P' \geq K'$.

Proof. P' and K' are the net level changes after some of the increasing and decreasing changes have canceled out. Thus,

$$P - P' = K - K' \Rightarrow P' - K' = P - K \geq 0 \Rightarrow P' \geq K'.$$

5.2.3 Approaching Proportional Fairness.

Theorem 2. *For times t_0 and t_1, at which the available bandwidths are the same, resource allocation $\vec{y}$ at t_1 will always be proportionally fairer than resource allocation $\vec{x}$ at t_0 under the proposed exponentially quantized rate control scheme (i.e., $PF(\vec{x},\ \vec{y}) \geq 0$).*

Proof. First, note that $PF(\vec{x}, \vec{y})$ is equal to

$$PF\left(\vec{x},\ \vec{y}\right) = \sum_{q_i>0} U\left(|q_i|\right) + \sum_{q_i<0} D\left(|q_i|\right).$$

Corollaries 5 and 6 show that U(n) and D(n) are minimized when the P' level-up events and the K' level-down events are evenly scattered. Thus,

$$\sum_{q_i>0} U\left(|q_i|\right) \geq U\left(1\right) \times P'$$

$$\sum_{q_i<0} D\left(|q_i|\right) \geq D\left(1\right) \times K',$$

and hence,

$$PF\left(\vec{x},\ \vec{y}\right) \geq U\left(1\right) \times P' + D\left(1\right) \times K'.$$

Corollaries 2, 3, and 4 show that $U(1) + D(1) \geq 0$ and $U(1) \geq 0$. Lemma 1 further indicates $P' \geq K'$. Therefore, $PF(\vec{x},\ \vec{y}) \geq 0$. □

The proposed exponential quantization scheme makes $PF(\vec{x},\ \vec{y})$ a nonnegative value and hence approaches a proportionally fair resource allocation among calls. Corollary 1 shows that the accumulated MOS is maximized when proportional fairness is reached, thereby achieving the second goal of this study.

6 SIMULATION

While the above section proves that EQ converges to optimality in (1) the number of calls served and (2) accumulative MOS, the goal of this section is to examine quantitatively how EQ compares to two other mechanisms. In the following, we describe the simulation methodology and the three rate controls. The results show that EQ outperforms the Naïve and Skype mechanisms in both number of calls served and accumulated MOS, echoing the properties proven in Section 5.

6.1 Call-Level Simulations

Network. What sets the rate control mechanisms apart is how they react to the changes in available bandwidth at the network bottleneck. Similar to prior works evaluating rate control mechanisms (Floyd et al. 2000; Ghobadi et al. 2016), we simulate calls as they traverse through a high utilization link. The bottleneck link capacity is set to 155Mbps (capacity of an OC-3 link). The average background traffic rate is set to 124Mbps, resulting in an average link utilization of 80%. Note that, under low utilization, supplying all calls with a constant 40kbps bitrate will work as well as any existing mechanisms. What really sets a robust rate control mechanism apart from others is whether the mechanism handles not just the easy situations but also the challenging ones. The settings are intended to drive the simulated network to cases where rate control mechanisms are needed the most.

Background Traffic. To capture the variability of Internet traffic in the background, our simulator implements the model proposed in Fraleigh et al. (2003), in which the traffic over the optical carrier (OC) backbone links in an ISP was measured, analyzed, and identified as Fractional Brownian Motion (FBM). Our implementation of the FBM model is adopted from (Dieker 2002).

Calls. Each call lasts for 300 seconds. Note that the call duration of 300 seconds is shorter than usual and challenging for the control mechanism to cope with, as the average Skype call duration has grown from 5 minutes in 2006 to 27 minutes in 2016 (Chen et al. 2006; Statistics 2016).

The maximum bitrate of a call is set to 40kbps. To stress the bottleneck link, all calls arrive at once at the beginning of the simulation. The starting bitrate is determined by the available bandwidth at the time the call starts. During the simulation, each call adapts periodically to the available bandwidth (i.e., capacity-background traffic-existing calls) based on the mechanism at hand. In case there is not sufficient available bandwidth to allow the lowest possible bitrate, the call is dropped. The adaptation period defaults to 1 second unless specified otherwise.

Number of Calls. To investigate how the various mechanisms scale with growing call traffic, we vary the number of calls simulated. The number of calls increases from 1,000 to 10,000 in a 1,000-call increment. For each number of calls, the simulation is repeated 100 times with random background traffic. In the end, 10 sets of 100 repetitions, a total of 1,000 simulation runs, are conducted.

Metrics. To echo the analysis in Section 5, we compare the mechanisms using two metrics: (1) the number of calls served and (2) the accumulated MOS score. The number of calls served is recorded by counting the number of calls remaining at the end of the simulation. The bitrate of a call is recorded per adaptation period. The bitrate samples are further converted to MOS scores using the Skype/SILK's MOS-bitrate model as reviewed in Section 2.4. A call's MOS score is the average of the MOS scores sampled throughout the call duration. Finally, the accumulated MOS score is the sum of all calls' MOS scores per simulation run.

6.2 Mechanisms in Comparison

Naïve Mechanism. The Naïve mechanism in this study implements an intuitive, fair control method that determines the level of call rate increase or decrease based on the amount of increase or decrease in available bandwidth divided by the number of calls.

Quantization Mechanism. As described in Section 4, the quantization mechanism implemented in the simulator was identical to the one shown in Figure 3, with exactly five levels. When there is an increase in available bandwidth, the bitrate is set to the level closest to, but not exceeding, the available bandwidth. When there is a decrease, the bitrate is set one level lower.

Skype Mechanism. This study includes Skype's rate control mechanism for comparison. Despite various attempts to decipher Skype's mechanism (Baset and Schulzrinne 2006; Cicco et al. 2007; Huang et al. 2010; Bonfiglio 2008), existing findings are based on the older version of Skype with fixed-rate audio codecs such as G.729. The exact rate adaptation scheme that the latest Skype uses

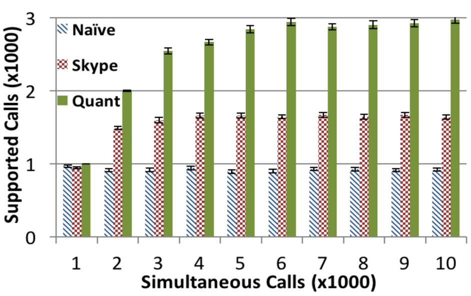


Fig. 5. Number of calls served.

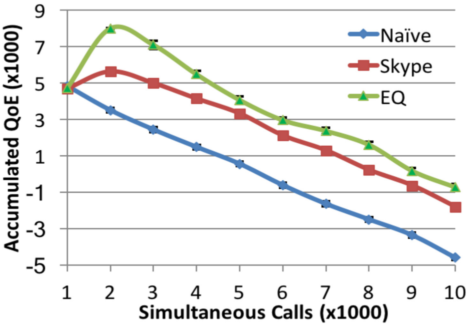


Fig. 6. Accumulated QoE.

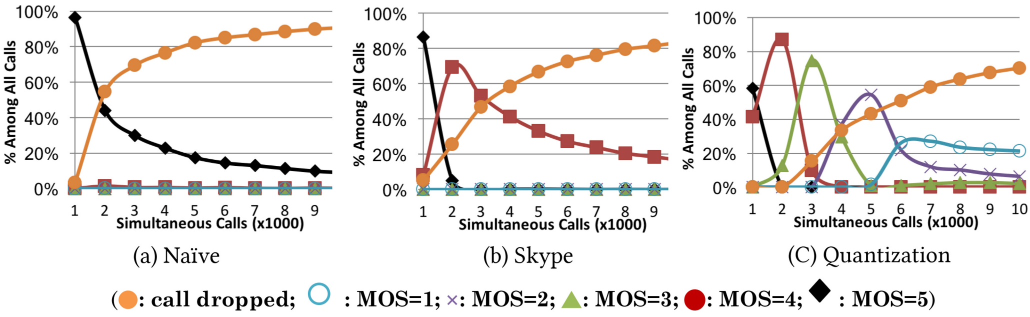
has not been systematically investigated. As such information is crucial in facilitating the simulation comparison, we backward-engineer Skype's rate adaption mechanism. In that, we enforce Skype to tune its call bitrate by emulating a wide range of available bandwidth. The relationship between the available bandwidth and call bitrate is measured and implemented as a table in the simulator. A Skype call is simulated by looking up the corresponding bitrate, provided the current available bandwidth. For further details of the backward engineering effort, please refer to the appendix.

6.3 Simulation Results

6.3.1 Number of Calls Served. Figure 5 is a bar chart showing the average number of calls supported as the number of calls increases. The x-axis indicates the number of simultaneous calls running through the bottleneck link in each simulation set. The y-axis plots the average number of calls over 100 runs per simulation set, and the error bar plots the 95% confidence interval. The Naïve mechanism serves the least amount of calls throughout all call populations, whereas the proposed quantization mechanism serves the highest amount. Skype offers new calls a higher chance of joining because of its slower response to spared bandwidth (see Appendix A.2). This characteristic allows Skype to support more calls than the Naïve mechanism. However, it fails to compete with the quantization mechanism because it does not discriminate calls while allocating spare bandwidth, i.e., missing the opportunities to keep the low bitrate calls alive.

6.3.2 Accumulated MOS. The accumulated MOS is derived by summing up the MOS scores of all calls. As a call adapts its rate, an MOS score is derived from the MOS-bitrate relationship. The MOS of a call is calculated by averaging the MOS scores recorded throughout the call. One issue to address is that the MOS score for the dropped calls is not 0. Based on the derived model of SILK, a sending rate of less than 4.091kbps takes the corresponding MOS to negative infinity. This means that the quality of dropped calls perceived by users is significantly worse than calls with low quality. In other words, a significantly small negative value is more characteristic of dropped calls. To facilitate quantitative comparison, this section assigns a conservative negative value, -1, to the dropped calls and discusses the effects of using a small negative score later.

Figure 6 shows the average accumulated MOS scores over 100 repetitions. The x-axis represents the number of simultaneous calls per simulation set and the y-axis represents the accumulated MOS. The error bar indicates the 95% confidence interval for the mean. The Naïve mechanism has the lowest accumulated MOS of all cases and the quantization mechanism the highest. Based on the number of calls dropped (Figure 5), the lower the negative MOS score for dropped calls is, the further the quantization mechanism outperforms the other two mechanisms. As discovered in



(a) Naïve (b) Skype (C) Quantization

Fig. 7. MOS distributions.

Appendix A.2, Skype's rate control mechanism is conservative in the timing in which a call adapts to the new bitrate. The new bitrate assignment is, however, almost aggressive as the Naïve mechanism. This indicates why Skype stands in between the other two mechanisms in accumulated MOS.

More specifically, in the 1,000-call scenario, the available bandwidth is abundant to support 1,000 simultaneous calls, as one can see in Figure 5 that there are barely call drops in the 1,000-call scenario. In the 2,000-call and 3,000-call scenarios, Skype begins to drop calls while the quantization mechanism is able to sustain more calls of lower bitrates. For simulation sets of 4,000 simultaneous calls and above, the performance gap, although statistically significant, is not as large. The network is badly saturated in these scenarios and there is little space for any rate control mechanism to maneuver. This leads to another potential use of the MOS-bitrate model: call admission control for local subnets. For a network adopting the Skype or quantization mechanism, allowing 3,000 or more calls does not yield a higher level of overall user experience.

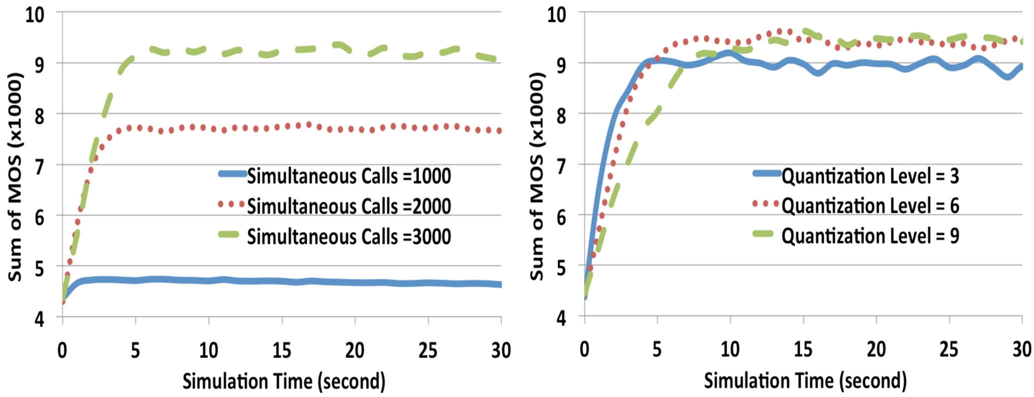
6.3.3 Call-MOS Distribution. The MOS distribution in Figure 7 shows the percentage of calls with different MOS values at the end of the simulation. The x-axis represents the number of simultaneous calls, and the y-axis represents the percentage of calls at each MOS value.

Figure 7(a) shows that all the MOS values of the Naïve mechanism are larger than 4. When the number of simultaneous calls increases, even the MOS = 4 calls drop and only the MOS = 5 calls remain. The property of the thin calls being dropped as the network bandwidth fluctuates and a few fat calls hoarding the network resources in the end is distinctive. The users of the dropped calls are frustrated, yet the users of the fat calls do not perceive a proportional improvement in perceptual quality. This explains the low number of calls supported and the low overall MOS, indicating why the Naïve mechanism is unable to dynamically adjust the allocation of network resources to improve the user experience.

Figures 7(b) and 7(c) show a clear shift in the MOS values of calls for the Skype and quantization mechanisms. To serve more calls simultaneously, the MOS values of most calls decrease gradually as the number of calls increases. Both mechanisms back off high-quality calls so that more calls are allowed to join. The proposed quantization mechanism adapts more effectively, explaining its advantage in the number of calls supported and the accumulated MOS over Skype.

7 CONVERGENCE PROPERTY

As proven in Section 5, the proposed exponential quantization scheme adapts and converges to a steady, proportional fair state over time. To validate this property, we conducted a separate set



(a) Varying number of simultaneous calls (Quantization level = 5)

(b) Varying number of quantization levels (Simultaneous calls = 3000)

Fig. 8. Convergence property of the Exponential Quantization scheme.

of simulations. In that, all calls started at the same time and ran infinitely long. The starting rates of the calls are randomly distributed within the 0kbps to 40kbps range, including scenarios where calls were clustered in two extreme groups, i.e., 0kbps and 40kbps groups. The adaptation frequency is set to once per second.

Shown in Figure 8(a) is the summation of MOSs (y-axis) of all calls over time (x-axis) for the clustered scenario, representing the worst-case behavior. One can observe that the calls find the optimal rates in a few seconds and the summation of MOSs converges to and stays at the maximum through the rest of the simulation. This shows not only that the calls stabilize well but also that the exponential quantization mechanism leads to the best overall QoE.

Note that the number of calls influences the convergence speed. The convergence time is longer when there are more calls. It is because only a portion of the calls gets to increase their sending rates when a high-rate call drops its sending rate by only one level within one adaptation cycle. It could take multiple adaptation cycles for the high-rate calls to tune properly back down.

A lower adaptation frequency would lengthen the time needed for convergence. As can be observed from Figure 8(a), with a 1-second adaptation period, the call rates converge to steady state in 3 to 5 seconds. That means it takes about three to five cycles to stabilize. The influence of the adaptation frequency to the convergence speed is linear to the adaptation period. Note also that the 1-second adaptation period is already longer than usual. The round-trip time of IP packets is at the scale of milliseconds. As such, the sender can be informed of available bandwidth changes within milliseconds, which is way below 1 second. In practice, EQ could converge well below 1 second, which is way shorter than any sensible conversation one would like to carry out. Similarly, during a long call, bandwidth changes will be adapted also quickly within a second.

Not so obvious is that the number of quantization levels affects also the speed of convergence. As shown in Figure 8(b), as the number of quantization levels increases, it takes more time to reach convergence. The phenomenon is attributed to the gap of adjacent bitrate levels. When there are fewer quantization levels in the fixed bitrate range, the gap tends to be larger and the amount of bandwidth released by calls with higher quality would be higher. This allows calls with lower quality to raise their bitrates to a higher level more easily and facilitates the convergence process.

8 CONCLUSION AND OUTLOOK

With the goal of devising a rate control mechanism for VoIP calls, this study investigates (1) how users perceive voice quality at different sending rates and (2) how to allocate bandwidth to retain users rather than to lose them.

Targeting the first issue, this study extends previous work measuring SILK/Skype call quality and investigates the most frequently used voice codecs: AMR-WB for general voice data. The user study reaffirms a significant log-like relationship between the data rate and the user experience and develops a model of sending rate to perceived call quality.

Inspired by the log-like relationship observed and enlightened by the theory of proportional fairness, this study devises a simple mechanism, named EQ, that exponentially quantizes the steps of rate change. The analysis proves that the resulting bandwidth allocation of EQ on calls with a logarithmic QoE-bitrate relationship converges to proportional fairness over time, and hence optimizes both the overall QoE and the number of calls served. Through simulations of SILK calls, we find that EQ outperforms Skype's rate adaptation mechanism in both the overall QoE and the number of calls served.

We unfortunately could not simulate AMR-WB calls. The main challenge is that although AMR-WB is a variable bitrate codec, it does not allow arbitrary bitrate encoding such as SILK does. That is, the bitrate levels that are exponentially distributed in EQ are simply not implementable using AMR-WB as of today. Furthermore, simulating hypothetical AMR-WB calls over EQ versus the current rate adaption mechanism in the cellular network does not prove EQ better either. One would have trouble isolating whether the mechanism in the cellular network is limited because of AMR-WB's bitrate setting or the adaptation strategy. That said, we have observed codecs evolving over time. In fact, SILK is one such result. Perhaps with EQ better known to the world, the standardization body would push to extend AMR-WB for a more flexible bitrate setting and adopt a rate adaptation mechanism such as EQ to allow more calls while maintaining the overall user experience.

APPENDIX

Here, we detail the experiments and findings unveiling Skype's rate adaptation mechanism.

A.1 Measurement Testbed

Considering Skype's internal as a black box, we set up a series of dummynet (Rizzo 1998) experiments to identify Skype's sending rate at different available bandwidths. The measurement testbed is illustrated in Figure A.1. We set up two Skype nodes within a subnet, and then insert a

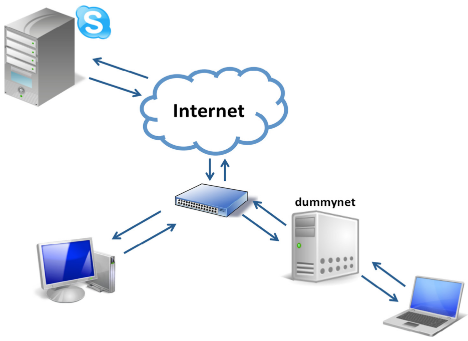


Fig. A.1. Skype measurement testbed.

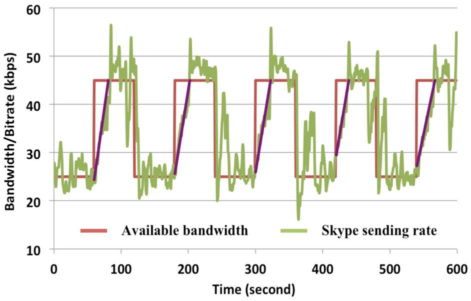


Fig. A.2. Skype's sending rate under available bandwidth dynamics.

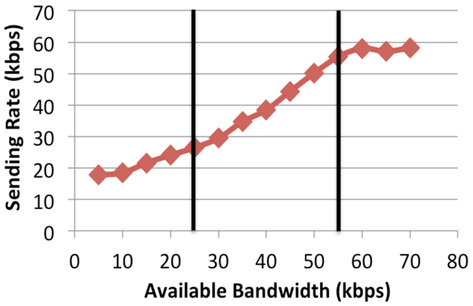


Fig. A.3. Skype's sending rate.

dummynet node in the middle. This allows controlling of the available bandwidth in between the Skype nodes. While varying the available bandwidth at the dummynet node, we measure the bitrate at the sender node. The gateway node blocks all external traffic except Skype's control messages. This allows the two Skype nodes to contact Skype's login server, so the two nodes can establish a call afterwards. This also avoids traffic coming from outside to complicate the measurement. The measured sending rate and response time per available bandwidth serve as the base of the Skype simulations.

A.2 Measurement Results

The measurement results reveal the characteristics of the response time and sending rate: (1) With regard to the response time, Skype's rate control is somewhat conservative. Figure A.2 shows, in an example scenario, how Skype adjusts its sending rate (green fluctuating line) to track the available bandwidth set by the dummynet node (red square wave), which alternates between 25kbps and 45kbps every 60 seconds. Skype increases its sending rate slowly when more bandwidth becomes available. It drops its sending rate immediately when the available bandwidth falls below the current sending rate. (2) In an episode of rate increase, the response time needed is proportional to the rate difference before and after the adaptation. The increasing speed is indicated by the linear fit (purple line segment) in Figure A.2. In the subsequent simulations, we implement Skype's rate adaptation obeying the identified increasing speed. (3) With regard to the sending rate, the target sending rate does not necessarily match the available bandwidth. Figure A.3 shows Skype's sending rates under different available bandwidths. As indicated in the figure, Skype estimates the available bandwidth properly and controls its sending rate to match the available bandwidth until it reaches its maximum, approximately 60kbps. (4) One curious behavior is that Skype overshoots the sending rate when the available bandwidth is below 25kbps. This design may introduce packet losses. Focusing on the adaptation to available bandwidth in this work, we leave the effect of packet losses for future work. The MOS of the simulated Skype call is derived from the measured sending rate. This overestimates the Skype call quality to a certain degree and gives the Skype results a certain degree of advantage over the proposed mechanism.

A.3 Discounting Packet Header Overhead

We use the MOS-bitrate model of SILK to score the simulated calls. One particular caution we take here is to discount the bandwidth consumed by the packet headers. This is not quite as straightforward as to subtract a fixed amount of bitrate from the measured sending rate given the available bandwidth.

Table A.1. Packet Header Overhead

Available Bandwidth	Packet HeaderOverhead
>40kbps	17.18kbps
17kbps ≤ AB ≤ 40kbps	8.79kbps
<17kbps	6.30kbps

SILK takes the frame rate as one of its input parameters. This allows Skype to control the amount of bandwidth consumed by the packet header. We find in our measurement that the frame rate differs with the available bandwidth. To articulate the score per simulated call, we subtract the bandwidth consumed by the packet header given the available bandwidth. The rules derived from the measurement are listed in Table A.1.

REFERENCES

3GPP. 2011. 3GPP TS26.171: Speech codec speech processing functions: Adaptive multi-rate - wideband (AMR-WB) speech codec; general description.

ITU-T. 1996. ITU-T Recommendation P.800, Methods for Subjective Determination of Transmission Quality.

ITU-T. 1996. ITU-T Recommendation P.830, Subjective Performance Assessment of Telephone-band and Wideband Digital Codecs.

ITU-T. 2006. ITU-T Recommendation P, 10/G.100. Vocabulary for Performance and Quality of Service.

ITU-T. 2008. ITU-T Recommendation P.910, Subjective Video Quality Assessment Methods for Multimedia Applications.

ITU-T. 2012. ITU-T Recommendation G.729: Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear prediction.

Amazon Mechanical Turk. Retrieved from https://www.mturk.com/mturk/welcome.

S. A. Baset and H. Schulzrinne. 2006. An analysis of the Skype peer-to-peer internet telephony protocol. In *Proceedings of IEEE International Conference on Computer Communications (INFOCOM'06).*

D. Bertsekas and R. Gallager. 1992. *Data Networks.* Prentice-Hall, Englewood Cliffs, NJ, 1992.

K. R. Boff, L. Kaufman, and J. P. Thomas. 1986. *Handbook of Perception and Human Performance.* Wiley-Interscience.

J. Bolot and T. Turletti. 1994. A rate control mechanism for packet video in the internet. In *Proceedings of IEEE International Conference on Computer Communications (INFOCOM'94).*

D. Bonfiglio, M. Mellia, M. Meo, N. Ritacca, and D. Rossi. 2008. Tracking down Skype traffic. *IEEE International Conference on Computer Communications (INFOCOM'08).*

T. Bu, Y. Liu, and D. Towsley. 2006. On the TCP-friendliness of VoIP traffic. In *Proceedings of IEEE International Conference on Computer Communications (INFOCOM'06).*

C.-N. Chen, C.-Y. Chu, S.-L. Yeh, H.-H. Chu, and P. Hunag. 2012. Measuring the perceptual quality of Skype sources. In *Proceedings of ACM Communications and Computer Networks (SIGCOMM, W-MUST'12).*

C.-N. Chen, C.-Y. Chu, S.-L. Yeh, H.-H. Chu, and P. Hunag. 2014. Modeling the QoE of rate changes in Skype/SILK VoIP Calls. *IEEE/ACM Transactions on Networking* 22, 6 (December 2014), 1781–1793.

K.-T. Chen, C.-Y. Huang, P. Huang, and C.-L. Lei. 2006. Quantifying Skype user satisfaction. In *Proceedings of ACM Communications and Computer Networks (SIGCOMM'06).*

L. D. Cicco, S. Mascolo, and V. Palmisano. 2007. An experimental investigation of the congestion control used by Skype VoIP. In *Proceedings of Wired/Wireless Internet Communications (WWIC'07).*

L. De Cicco and S. Mascolo. 2008. A mathematical model of the Skype VoIP congestion control algorithm. In *Proceedings of IEEE Conference on Decision and Control (CDC'08).*

T. Dieker. 2002. Simulation of Fractional Brownian Motion. Master's thesis, University of Twente, the Netherlands.

M. Fiedler, T. Hossfeld, and P. Tran-Gia. 2010. A generic quantitative relationship between quality of experience and quality of service. *IEEE Network* 24, 2 (2010), 36–41.

S. Floyd and K. Fall. 1999. Promoting the use of end-to-end congestion control in the internet. *IEEE/ACM Transactions on Networking* 7, 4 (1999), 458–472.

S. Floyd, M. Handley, J. Padhye, and J. Widmer. 2000. Equation-based congestion control for unicast applications. In *Proceedings of ACM Communications and Computer Networks (SIGCOMM'00).*

C. Fraleigh, F. Tobagi, and C. Diot. 2003. Provisioning IP backbone networks to support latency sensitive traffic. In *Proceedings of IEEE International Conference on Computer Communications (INFOCOM'03).*

M. Ghobadi, R. Mahajan, A. Phanishayee, H. Rastegarfar, P.-A. Blanche, M. Glick, D. Kilper, J. Kulkarni, G. Ranade, and N. Devanur. 2016. ProjecToR: Agile reconfigurable datacenter interconnect. In *ACM SIGCOMM*.

M. Handley, S. Floyd, J. Padhye, and J. Widmer. 2003. TCP friendly rate control (TFRC): Protocol specification. *RFC* 3348 (2003).

T.-Y. Huang, P. Huang, K.-T. Chen, and P.-J. Wang. 2010. Can Skype be more satisfying? A QoE-centric study of the FEC mechanism in the internet-scale VoIP system. *IEEE Network* 24, 2 (2010), 42–48.

F. Kelly. 1997. Charging and rate control for elastic traffic. *European Transactions on Telecommunications* 8 (1997), 33–37.

F. P. Kelly, A. Maulloo, and D. Tan. 1998. Rate control for communication networks: Shadow price proportional fairness and stability. *Journal of the Operational Research Society* 49 (1998), 237–252.

H. Kim, K. Kim, Y. Han, and S. Yun. 2004. A proportional fair scheduling for multicarrier transmission systems. *IEEE Vehicular Technology Conference (VTC'04).*

R. Kwan, C. Leung, and J. Zhang, 2009. Proportional fair multiuser scheduling in LTE. *Signal Processing Letters* 16 (2009), 461–464.

R. J. La and V. Anantharam. 2002. Utility-based rate control in the Internet for elastic traffic. *IEEE/ACM Transactions on Networking* 10, 2 (2002), 272–286.

M. R. Longo and S. F. Lourenco. 2007. Spatial attention and the mental number line: Evidence for characteristic biases and compression. *Neuropsychologia* 45 (2007), 1400–1406.

R. S. Moyer and T. K. Landauer. 1967. Time required for judgments of numerical inequality. *Nature* 215, 5109 (1967), 1519–1520.

P. Reichl, S. Egger, R. Schatz, and A. DAlconzo. 2010. The logarithmic nature of QoE and the role of the weber- fechner law in qoe assessment. In *Proceedings of ICC'10.*

P. Reichl, B. Tuffin, and R. Schatz. 2010. Economics of logarithmic quality-of-experience in communication networks. In *Proceedings of the IEEE Conference on Telecommunications Internet and Media Techno Economics.* CTTE, 2010.

F. Ribeiro, D. Florencio, C. Zhang, and M. Seltzer. 2011. CROWDMOS: An approach for crowdsourcing mean opinion score studies. In *IEEE ICASSP.*

L. Rizzo. 1998. Dummynet and forward error correction. In *Proceedings of the USENIX Annual Technical Conference.*

L. Rizzo. 2000. Pgmcc: A TCP-friendly single-rate multicast congestion control scheme. In *Proceedings of ACM Communications and Computer Networks (SIGCOMM'00).*

E. D. Scheirer. 1998. Tempo and beat analysis of acoustic musical signals. *Journal of the Acoustical Society of America* 103, 1 (1998), 588–601.

J. Shen. 2003. On the foundations of vision modeling: I. Weber's law and Weberized TV restoration. *Physica D: Nonlinear Phenomena* 175, 3–4 (2003), 241–251.

Statistic Brain. 2016. Retrieved from http://www.statisticbrain.com/skype-statistics/.

N. Wakamiya, M. Murata, and H. Miyahara. 2000. On TCP-friendly video transfer with consideration on application level QoS. In *Proceedings of IEEE International Conference on Multimedia and Expo (ICME'00).*

K. Vos, S. Jensen, and K. Soerensen. 2010. Internet-draft: Draft-vos-silk-02, SILK speech codec. *Internet Engineering Task Force (IETF).*

Y. Xue, B. Li, and K. Nahrstedt. 2006. Optimal resource allocation in wireless ad hoc networks: A price-based approach. *IEEE Transactions on Mobile Computing* 5, 4 (2006), 347–364.

J. Yan, K. Katrinis, M. May, and B. Plattner. 2006. Media- and TCP-friendly congestion control for scalable video streams. *IEEE Transactions on Multimedia* 8 (2006), 196–206.

Y.-C. Yen, C.-Y. Chu, C.-N. Chen, S.-L. Yeh, H.-H. Chu, and P. Huang. 2013. Exponential quantization: User-centric rate control for Skype calls. In *Proceedings of the 31st ACM Annual Conference of the Special Interest Group on Data Communication (ACM SIGCOMM'13)*, Poster Session, Hong Kong, August.

Y.-C. Yen, C.-Y. Chu, S.-L. Yeh, H.-H. Chu, and P. Huang. 2013. Lab experiment vs. crowdsourcing: A comparative user study on skype call quality. In *Proceedings of the 9th Asian Internet Engineering Conference (AINTEC'13).*

Received February 2016; revised May 2017; accepted November 2017