On Autonomic QoS Provisioning for Voice over IP Services

Thomas Michael Bohnert, Edmundo Monteiro
DEI / CISUC
University of Coimbra
Polo II - Pinhal de Marrocos
3030 - 290 Coimbra
{tbohnert,edmundo}@dei.uc.pt

Abstract— **In this paper we present a framework for autonomic QoS provisioning for VoIP services. The central role is held by a Measurement Based Admission Control algorithm which incorporates three innovations. First, the mathematical underpining obeys a non-parametric approach, removing the dependency on a priori assumed characteristics of the underlying stochastic process, i.e. those of the work arrival process. Thus, we tackle a major issue of common parametric MBAC.**

The second and the third enhancement is embodied in the combination of a closed-loop control based on perceived QoS. Typically, MBAC algorithms do not validate there decision and thats why many algorithms miss QoS targets due to non-stationaries in work arrival processes. As a metric for performance evaluation we devised a new approach based on the Emodel, an ITU-T standard for quantifying QoS assessment based on human perception.

The contribution is the mathematical framework for both, the non-parametric MBAC and its closed-loop control but also an analysis based on simulations.

I. Introduction

Public approval plus virtually ubiquitous accessibility supported by a versatile set of technological drivers make the Internet evolving faster than ever into what it has been designed for, a multi-service communication infrastructure serving a global population. The robust but equally flexible architecture lends itself perfectly for a broad spectrum of applications out of which probably the most heeded one right now is voice communication, widely known as Voice Over IP (VoIP). Traditionally, an exclusive domain of large telecomunication companies and carried over huge, dedicated PSTNs[1], VoIP services are continuously being taken over by companies operating on Internet infrastructure. Actually, VoIP is to be seen as the first real-time service on a global scale.

Services like VoIP with inherent, stringent network requirements, however, urge for a structural move of the Internet architecture, away from a pure best-effort service towards Quality of Service (QoS) enhancements and the most accepted approach for QoS support is the Differentiated Services (DiffServ) [1] architecture.

DiffServ is typically applied only on access networks, i.e. on network edges, as the network core infrastructure is overprovisioned to bear with large traffic volumes and as service differentiation is considered to be of little advantage [2] in the core. Contrary to the latter, in access networks with highly time varying loads, overprovisioning incurs the risk of large revenue loss for Internet Service Providers (ISP) as resources are being unused over long periods, e.g. during night hours.

DiffServ provides scalable QoS by classifying applications into classes and works on their respective traffic aggregates, rather than on single flows. Aggregates are consequently treated differential by class priority meaning that resource assignment, like for example buffer and forwarding capacity, is according to expected traffic volume and class priority. DiffServ therefore enables the Internet to provide more or less fine grained QoS guarantees for real-time services along with reduced complexity and scalability.

Explicit but static assignment of resources to traffic classes, however, is insufficient and another entity called Admission Control (AC) is mandatory for QoS provisioning. In order to maximise revenue, ISPs typically attach more customers to access links than they can carry simultaneously since the probability of concurrent access converges to zero with an increasing number of attached terminals. A general rule is to attach as many customers to an access router as possible given the constraint that a maximum of 3 percent of customers are rejected in the busiest hour of a day, resulting in up to an 50:1 oversubscription ratio. The role of AC is exactly the access regulation on oversubscribed links. Generally speaking, AC is a policy based decision algorithm to protect traffic of a class from QoS degradation in times of high contention. In short, a new resource consumer, i.e. a flow, is admitted to a link if its characteristic in terms of resource demand superimposed with that of ongoing, previously admitted consumers is up to an extent, such that a given QoS level can be granted for the whole traffic aggregate. Though the need for AC appears just natural, AC has not been defined for the DiffServ architecture [1].

In this paper we present a framewok for autonomic QoS provisioning for VoIP services based on DiffServ with AC as the central component. In Sec. II, we present the mathematical model which follows the Measurement-based Admission Control (MBAC) approach. Contrary to common frameworks where QoS metrics are physical parameters, in Sec. III we introduce an closed-loop control extension based on so-called *subjective* QoS provisioning. Though our framework is yet at an early stage, we present and discuss first performance results of the MBAC algorithm in Sec. IV. Current conclusions and an outlook in Sec. V closes this paper.

II. A Non-Parametric Model for Measurement-based Admission Control

Measurement-based AC has been introduced to loosen dependencies on accurate source, and moreover network modelling. In fact, characteristics of traffic aggregates can be extremely complex to model and strongly dependent on random, time varying factors like number of cascaded queues, congestion control protocols, application mixtures and features and even human interaction. As a response to these findings, the rationale of MBAC is to replace a priori assumptions by actual measurements taken in real time from traffic aggregates, i.e. the underlying stochastic process. Statistics estimated from these measurements are then fed in purpose build stochastic queueing models to estimate needed resources to cater an aggregate

[1]Public Switched Telephone Network

according to present QoS requirements.Precise estimation therewith becomes a crucial matter for MBAC.

Generally, two properties of stochastic processes are decisive, the marginal distribution and the autocorrelation function. While the latter is an estimate in any case, a priori assumptions about the marginal distribution are particular significant. To illustrate this, simply consider the Exponential and Normal distribution. Both are fully defined by either only the first, or by the first and the second moment. Estimates taken from a sample applied to either model, however, yields very different probabilities. This general fact is directly related with MBAC as we will illustrate below.

Assume $\{A\}$ as the work arrival process, and let $A[s,t]$ be the amount of work arriving in the interval $(s,t]$. Further let $A_t = A[-t,0]$ such that the queue length at time zero is

$$\Omega = \sup_{t\geq 0}(A_t - Ct). \tag{1}$$

with C denoting link capacity. The probability that the queue length exceeds ω is herewith

$$P\{\Omega > \omega\} = P\{\sup_{t\geq 0}(A_t - Ct) > \omega\}. \tag{2}$$

As (2) is difficult to compute, we use the lower bound approximation

$$P\{\sup_{t\geq 0}(A_t - Ct) > \omega\} \geq \sup_{t\geq 0} P\{A_t > \omega + Ct\}. \tag{3}$$

With $\rho = \omega + Ct$, and let $F_t(x) = P\{A_t > x\}$ be the CDF of $\{A_t\}$ we get

$$P\{\Omega > \omega\} \geq \sup_{t\geq 0} P\{A_t > \omega + Ct\} = \sup_{t\geq 0}(1 - F_t(\rho)). \tag{4}$$

Eventually, the r.h.s of Equ. 4 reveals the dependencies on assumptions about F_t, the integral of the margingal distribution of $\{A\}$.

The illustrated argument is generally valid for all probability models and hence, thereof build MBAC algorithms. The generally adopted approach to deal with that issue obeys that of parametric statistics. Parametric in statistical terms and in this context means, that the nature of the CDF is assumed to be known. Motivated by the Central Limit Theorem for instance, many MBAC models assume Gaussianity (or Normality), like [3] [4] [5] [6] and [7], while for instance [5] assumes a Gumbel distribution [5] motivated by Extreme Value Theory.

Particularly the Gaussian approximation has become very popular recently with rising aggregation levels in access networks and due to its conformance with Self-Similarity (SS) and Long-Range-Dependency (LRD) findings in network traffic [8] [9] [10] [11]. An analytical argument in favour of is documented in [12] while [2] further supports it empirically based on real traffic traces, though for backbone traffic and for large time scales. The latter results, however, are questionable as common Goodness-of-fit techniques were applied though known to be error prone for large samples and in the context of SS and LRD [13, Chap. 10] [14] [15] [16, Page 33]. In [14] a more profound mathematical approach has been applied to test the Gaussian approximation. The result is rather inconclusive, stating that under certain conditions the approximation can be met but also grossly missed, particularly for small time scales.

In contrast to Gaussian approximation, the authors of [17] [18] state traffic obeys a Poisson law and is non-stationary. However, these results have been recently called to be far unrealistic [15]. Finally, for the sake of further highlighting diversity of findings, we cite [19] which recommends the Gamma distribution as the *best choice on average* for a comprehensive set of recorded traffic traces.

This short discourse indicates the uncertainty associated with a priori assumptions. We conclude that there is evidence for the application of non-parametric approaches, i.e. independent and autonomous models flexible to adapt to any condition. Indeed, several histogram based MBAC algorithms, like for instance [20] [21] [22], have been introduced in the past. However, non of the papers justifies the choice for histograms by statistical uncertainty, but use them rather implicitely. This is also evidenced by the fact that in each case, histograms are exclusively used in an ad-hoc fashion and parameters, namely bin width and bin origin, are chosen intuitively. That this salvages the risk of great imprecision is widely unknown though elaborately presented in [16, Chap. 3].

A non-parametric method to estimate unknown densities is Kernel Density Estimation. Following [16, Chap. 6, Equ. 6.1] the definition reads

$$\hat{f}_h(y) = \frac{1}{nh}\sum_{i=1}^{n} k\Big(\frac{y - Y_i}{h}\Big) \tag{5}$$

where Y is a random variable with random sample $\{Y_1, \ldots, Y_n\}$, h is the window width, also called the *smoothing* or *bandwidth* and k is the *kernel*, which has to satisfy the condition

$$\int_{-\infty}^{+\infty} k(u)du = 1 \tag{6}$$

and is therefore often chosen from a density function. In brief, the KE can be considered as a sum of superimposed bumps centred at the observations Y_i. The shape of the bumps is defined by the kernel function while the window width h defines their width [23]. From (5) we get the distribution estimator by integration

$$\hat{F}_h(y) = \frac{1}{nh}\int_{-\infty}^{y} \hat{f}_h(u)du = \frac{1}{nh}\sum_{i=1}^{n} K\Big(\frac{y - Y_i}{h}\Big) \tag{7}$$

where

$$K(x) = \int_{-\infty}^{x} k(u)du \tag{8}$$

is the integrated kernel.

The standard metric for accuracy evaluation for this type of estimator is the *Mean Integrated Squared Error* (MISE) defined as

$$MISE(h) = E\int\{\hat{F}_h(y) - F(y)\}^2 dy. \tag{9}$$

Letting $hn^{1/2} \to \infty$ as $n \to \infty$ yields the *Asymptotic Mean Integrated Squared Error* (AMISE). It has been shown in [24] that the AMISE is minimised by setting h equal to

$$h^o = (\Upsilon(K)/R_1)^{1/3} n^{-1/3} \tag{10}$$

with $\Upsilon(K)$ being a function of the kernel only. For instance, using a standard normal density as kernel, $k(x) = \phi(k)$ and $K(x) = \Phi(x)$, yields $\Upsilon = 1/\sqrt{\pi}$.

From (10) we see that provided an accurate estimate of R_1, the so-called *roughness* defined as

$$R_m = \int_{-\infty}^{+\infty} (f^{(m)}(y))^2 dy \tag{11}$$

where f^m denotes the mth derivative of f, yields an estimator with a convergence rate of $n^{-1/3}$. This is slower as the rate of n^{-1} for parametric estimators under optimal conditions, i.e. a full match between reality and model. This condition is, however, practically never given. Quantification of (11) is discussed in Sec. IV.

Finally, with (7) and (10) put in (4) we get

$$\hat{P}_{loss} = P\{\Omega > \rho\} = \sup_{t\geq 0}\Big\{1 - \frac{1}{nh^o}\sum_{i=1}^{n} K\Big(\frac{\rho - A_t^i}{h^o}\Big)\Big\} \tag{12}$$

with $\{A_t^1, \ldots, A_t^n\}$ being a sample of the arrival process $\{A\}$ at time scale t and setting ω in $\rho = \omega + Ct$ to be the buffer size. Using (12) and further assuming r_p to be the peak rate of the flows, we can finally define our admission criteria. Knowing that a new flow's

worst case behaviour in terms of resource demand is Constant Bit Rate (CBR), ρ is set to $\rho = \omega + Ct - r_p t$ and hence

$$Admitt_{bool} = \begin{cases} true & if \quad \hat{P}_{loss} < P_{loss} \\ false & if \quad \hat{P}_{loss} \geq P_{loss} \end{cases} \tag{13}$$

III. Closed Loop Control based on Subjective QoS

The major design goal of our MBAC is autonomy as elucidated in the previous section. By its nature, its performance only depends on proper smoothing (h^o) which is being estimated from a given sample. There is no need for *"human fine tuning of model parameters without any intuitive meaning"*, one of the major criticism about MBAC [25]. But there is also a commonality with general approachs, its probabilistic nature and the associated risk of local imprecision. Under optimal conditions, i.e. taking estimator consistency as granted, $\hat{P}_{loss}$ converges to P_{loss} for an infinite sample, but for realistic sample sizes, $\hat{P}_{loss}$ poses a RV by itself varying with each estimate. Time varying estimates, and hence MBAC decision arguments, incur the risk of over admission and thus, QoS degradation.

To account for this matter we further extend our algorithm towards a closed-loop control. The idea is to monitor current experienced QoS and compare it with the predicted level. In case of discrepanciy, i.e. large deviations, the algorithm does autonomously adjust. This differentiates our approach from traditional MBAC, which is a pure preventive congestion control mechanism, relying solely on prediction and model precision.

To do so, we first have to define a metric and a natural choice is the difference between measured packet loss and the predicted $\hat{P}_{loss}$. This standard procedure obeys the paradigm of *Intrinsic QoS*, i.e. quality is evaluated on the basis of physical parameteres and predefined thresholds, the general IETF[2] metric for quality assessment [26].

Quality assessment for telephony (or VoIP), however, is a highly subjective matter and in any case, the ultimate measure is human satisfaction. A standard metric for the assessment of service quality based on human perception, so-called *Subjective or Perceived QoS*, is the Emodel. Started as a study by the ETSI, it has been formally published as a standard by the ITU-T [27]. It provides an unique method for objective mouth-to-ear transmission quality assessment based on human perception and is defined as

$$R = R_0 - I_s - I_d - I_e + A \tag{14}$$

In (14), R denotes the psychoacoustic quality score between 0 and 100. It is an additive, non linear quality metric based on a set of impairment factors. Noise and loudness effects are represented by R_0, where I_s poses speech signal impairment like for example PCM quantising distortion. Both are intrinsic to the speech signal processing itself. Impairment imposed by the information transport is represented by I_d, which stands for speech signal delay impairment and I_e for "equipment" such as IP networks. Eventually, A is the advantage factor, a compensator for poor quality along with improved convenience (e.g. cell phones). To assess the quality of a VoIP call, one has to compute the individual components of (14) and add them up to get the final score. The relation of R and the final human satisfaction has been defined in [27] and is depicted in Fig. 1.

Applying (14) as a metric for our purpose calls for discussion. How to compute the elements of the r.h.s of (14) and on what scale do we apply the metric, i.e. on flow or aggregate scale?

Regarding the r.h.s elements of (14), I_s and R_0, as not related to impairment incurred by the transport media, are set to default values, $R_0 - I_s = 94$ [28]. Further, assuming worst cases, A is set to zero. Thus, only I_d and I_e are remaining and the metric now reads

$$R = 94 - I_d - I_e \tag{15}$$

[2]Internet Engineering Task Force

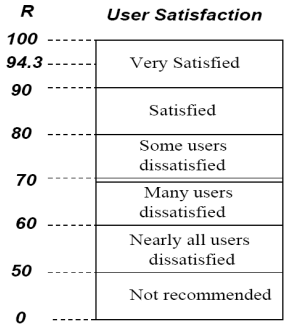

Fig. 1. Emodel Rating according to G.107/Annex B

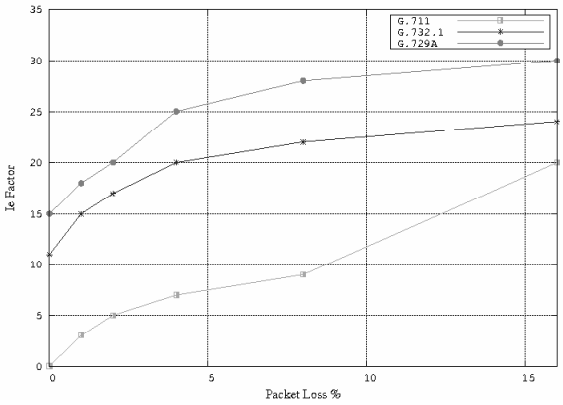

Fig. 2. Packet Loss and I_e Mapping

By setting the buffer length, ω in (4), to a fixed value we can pose a upper limit on delay impairment and thus, I_d is set to the worst case value, derived from [28, Sec. E]

$$I_d = 4.(\omega/C) \tag{16}$$

Eventually, what remains is the calculation of I_e which represents the impairment incurred by packet loss. By continously capturing loss events, I_e does function as the desired relation to the estimated $\hat{P}_{loss}$ using (12). The relation of measured loss and I_e is generally non-linear and VoIP codec dependent and we opt for the widely accepted mapping presented in [28] and depicted in Fig. 2.

Up to now, we basically devised a feedback algorithm based on measured loss events but yet there is a time relation missing. Measured packet loss percentage (or estimated probability) alone does not provide enough information for quality assessment. Equal loss percentage but different distributions of loss events over time, e.g. bursty loss at call end in contrast to uniformly distributed loss over a complete holding time, leads to different QoS rating by individuals.

This exactly reveals the conceptual weakness inherent to commonl MBAC algorithms purely based on intrinsic QoS, i.e. estimated QoS parameters as for instance loss and delay probability. General best practice is therefore to configure admission criteria that high, e.g. $P_{loss} = 10^{-6}$ to 10^{-9} such that loss events basically never take place. Hence, masking any timing effect.This has consequences as these rare events are basically never present in samples and moment estimates (e.g. for parametric models) taken from the latter do not represent the true underlying stochastic process.

Moreover, when working on traffic aggregates such low probabilities are even more questionable since in case of a loss event, only a small subset of flows is affected, leading to even lower loss probabilities for the majority of flows. This does certainly contribute

to customer satisfaction, but as admission rates are low for such a configuration ISPs risk lost revenue due to underutilised resources. In fact, this is even more relevant for VoIP applications, which are up to a certain extent tolerant to packet loss if loss clusters are not too dense.

In the literature of subjective QoS assessment, the impact of loss event distribution is well known; for an elaborate overview see for example [29]. To incorporate this in our model, we adapt the concept of loss gaps and bursts similar to the model presented in [28]. Loss events are captured and based on a preset time distance grouped in bursts or in gaps. Whenever, there is transition from a loss gap to a burst or vice versa, the subjective quality is computed using the loss percentage and the mapping presented in Fig. 2. More formally, let

$$E[I_e] = 1/T \sum_{i}^{N} I_{e,i} * t_i \tag{17}$$

be the mean loss impairment over a window of the last T seconds, where t_i is the sojourn time in either gap or burst state. Putting (16) and (17) in (15) we get our final quality score which reads

$$R(T) = 94 - 4.(\omega/C) - 1/T \sum_{i}^{N} I_{e,i} * t_i \tag{18}$$

By continuous computation of this score, we gain a long term picture about QoS levels experienced by customers. For example, setting T in (18) to 210s, the typical holding time for business calls, the rating reflects the quality a terminating call would have received. If the rating is below user satisfaction, i.e. $R(T) < 80$, admission of new flows is temporarily suspended.

Yet we did not discuss the scale on which we apply the QoS assessment metric. Our goal was to devise a framework for DiffServ, whose principal concept is aggregation for scalability reason. In adherence to this concept, we also apply our model on aggregate scale as this does not affect our quality rating. Loss events on aggregate scale can be assumed to be distributed over subsets of individual flows and thus, the experienced quality of a single flow can be assumed to be far better compared to that of the aggregate with a high probability. Thus, our model can be seen as a worst case model and does therefore provide some safety margin. Furthermore, if applied on flow scale, we would have to average individual ratings what in turn yields a rating on aggregate scale.

Plugging all components together we have the following functional. Whenever a flow requests admission, we use (12) to check if the network has enough resources on a short term. This is achieved by keeping the sample sizes limited such that the algorithm does maintain its reactivity and flexibility, the main goal of MBAC. This is the preventive functional based on prediction of resource demand.

If the first condition is met, a second conditions is verified for long term QoS, which can vary from the former due to the illustrated effects in the previous paragraph. This is the reactive functional, based on measured loss events.

Only ff both conditions are met, the flow is admitted, else rejected. Therefore, the algorithm can be considered as largely autonomous due to its independence from human "fine tuning" and as equipped with a closed-loop control mechanism to deal with misleading conditions, i.e. missing preset performance targets due to prediction errors in (12) or simply due to local effects of the loss process.

IV. Performance Evaluation of the MBAC Component

The main contribution of this paper is the conceptual and mathematical framework of the presented closed-loop MBAC tailored to subjective QoS provisioning for VoIP services. A complete and thorough investigation of its features, advantages and possibly weaknesses is currently going on and some indicative results are presented here in advance. Our aim is to support our concept quantitatively and to provide insight for evaluation purposes and community feedback.

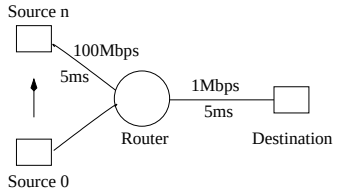

Fig. 3. Simple many sources, one router and one destination topology

The argument in favour of a non-parametric approach was uncertainty about the marginal distribution at small time scales $t = \tau$ in (12). Thus, before we can evaluate the performance, the question is what is the time scale range of interest? To answer this question, we again consider the context of human perception. It is known that the critical time for voice intelligibility is in the range of 40 to 80 ms, the duration of a single phoneme. Thus, the operational time scale of the algorithm can also only be in this range, i.e. 20ms up to a maximum of 100ms. More is not meaningful as the entropy a conversation could experience in such a large window is far from what is acceptable. Finally, we annotate that the maximum one way round trip time for VoIP should not exceed 150ms. This makes larger time scales on a single node somewhere on the paths even less meaningful.

To evaluate the algorithm we developed a module for the NS-2[3] simulator. As central for the algorithm, our focus at this stage is the performance of the density estimator, i.e. Equ. 12. As network topology we use a simple many sources, one router and one destination topology, see Fig. 3 for details. Sources are modeled using standard On/Off models to simulate realistic VoIP encoders with Voice Activity Detection (VAD) and On/Off distributions are either Exponential (EXPOO) or Pareto (POO). For both source models the mean On time is set to 600ms and Off to 300ms respectively [29]. On call level, Poisson distributed call arrivals are configured with mean 3.75/s and exponentially distributed holding times with mean 210s, the typical setting for business environments [29]. In On state, sources emit packets of 125 Bytes length and with rate of 64Kbps, simulating a G.711 codec and one voice frame per packet.

Finally, the router queue is set to 15 packets herewith limiting maximum delay to 15ms. Here we assume 10 hops end-to-end as typical for the Internet and a maximum delay of 150ms one way as the desired upper limit. A busy hour is simulated setting total simulated time to 4100s where the first 500s are disregarded to evaluate the system in equilibrium.

To gain detailed insight, we evaluate the algorithm for particular time scales separately and the results are listed in Tab. IV. Performance metrics are, target loss threshold $P_{l,t}$, total loss ratio $\hat{P}_{loss}$ and link utilisation U_l as standard for MBAC evaluation.

The results provide some interesting insight. For a very small time scale τ, included for instructive reasons and out of the range of interest, the algorithm is able to achieve high link loads U_l, i.e. high admission rates (ISPs objective) but misses the configured target loss probability $P_{l,t}$ (Customers objective) up to three orders of magnitude for both source models. For $\tau = 0.04$, a scale in the range of interest, the algorithm does perform much better, i.e. achieving high link loads and does approach $P_{l,t}$ quite closely. On the largest time scale, $\tau = 0.08$, the algorithm becomes too restrictive for strict QoS targets in the order of 10^{-3}. In conclusion, what can be stated is, that the algorithm does perform well for QoS targets up to 10^{-2} and $\tau \geq 0.04$.

Are there explanations for the imprecision? The bottleneck capacity is 1Mbps and since the MBAC works on this scale, the maximum arrival rate fluctuates around this value. That means that in average

[3]http://www.isi.edu/nsnam/ns/

TABLE I

SIMULATIONS RESULTS FOR THREE DIFFERENT TIME SCALES $t = \tau$

$\tau = 0.01$	POO		EXPOO	
$P_{l,t}$	$\hat{P}_{loss}$	$\hat{U}_l$	$\hat{P}_{loss}$	$\hat{U}_l$
10^{-1}	$4.7 * 10^{-1}$	1.00	$4.6 * 10^{-1}$	1.00
10^{-2}	$3.8 * 10^{-1}$	0.99	$3.8 * 10^{-1}$	0.99
10^{-3}	$1.8 * 10^{-1}$	0.70	$3.0 * 10^{-1}$	0.99
$\tau = 0.04$				
10^{-1}	$1.1 * 10^{-1}$	0.97	$1.1 * 10^{-1}$	0.97
10^{-2}	$4.0 * 10^{-2}$	0.92	$2.0 * 10^{-2}$	0.89
10^{-3}	$\leq 10^{-5}$	0.50	$\leq 10^{-5}$	0.50
$\tau = 0.08$				
10^{-1}	$4.0 * 10^{-2}$	0.93	$3.0 * 10^{-2}$	0.91
10^{-2}	$5.0 * 10^{-4}$	0.92	$2.0 * 10^{-4}$	0.65
10^{-3}	$\leq 10^{-5}$	0.50	$\leq 10^{-5}$	0.50

around $C*0.01$ bits can be sent per τ, corresponding to $\tilde{1}0$ packets in the case of $\tau = 0.01$. This is a very small value and does transform the marginal distribution of the arrival process in a rather discret one, in turn leading to imprecision since the estimator assumes continuous forms.

The main reason, however, lies in the estimation of the optimal smoothing parameter in Equ. 12, h^o. As mentioned in Sec. II, h^0 is a function of (11), the roughness of $f(x)$ and for this first study, we used a Gaussian scale reference estimate $h_\sigma^{ref}(R_1(f)) = (\hat{\sigma}^3 4)^{1/3} n^{-1/3}$ with $f \sim \phi(x)$. This estimate is too imprecise for very small time scales as it depends on the variance which is limited for this time scales and does assume near Gaussian shapes.

Finally, some comments are added to this discussion. In the first place, using a Gaussian reference estimate is contradictory to our main goal, loosen dependency on statistical assumptions. For this reason, we are currently evaluating the so-called data driven bandwidth selection methods based on the cross-validation principle, see for example [16, Chap. 6.5] for an introduction. Moreover, for this evaluation we did not enable the feedback algorithm. If enabled, it would prevent the algorithm from over admission, i.e. missing the target loss probability. Though this is the desired feature, it can only be the goal to optimise the predictor, i.e. Equ. 12, for instance for stability reasons. Hence, in this evaluation we only focused on the performance of the Kernel Estimator, i.e. Equ. 12 to validate our argumentation.

V. CONCLUSION

In this article we presented the framework for autonomic QoS provisioning with an MBAC algorithm as central component. Three major innovations do differentiate our algorithm from common MBAC approaches. The first one is its independence on statistical assumptions. In Sec. II we provided evidence about the dependency of MBAC approaches on the marginal distribution of the work arrival process and also related uncertainty about its nature based on a set of recent publications. We concluded that a non-parametric approach is to be investigated and developed an algorithm based on Density Estimation.

Naturally, estimation of parameters is afflicted by chance and thus subject to local imprecision, as discussed in Sec. III. To overcome this issue, the second enhancement is the development of an closed-loop control, a step ahead from classical open-loop MBAC.

A third innovation is its focus on the so-called subjective QoS. Rather than being built on metrics with purely physical interpretations, the feedback mechanism presented does account for human satisfaction. This has been motivated by the fact that the functional between a QoS parameter and human satisfaction is non-linear. We showed that intrinsic QoS is inappropriate for VoIP quality assessment and devised a model which does quantify human satisfaction based on measurements using the ITU-T Emodel.

The MBAC model of the framework has been implemented in a simulator and analysis showed that the it generally performs very well. This is interesting as yet not all improvements have applied for this analysis. But also some issues have been indentified, one of the goals of this investigation. Solutions are currently being evaluated and will be published in a follow-up of this work.

ACKNOWLEDGEMENT

This work was supported by the Portuguese Ministry of Science and High Education (MCES), and by European Union FEDER under program POSI (project QoSMap).

REFERENCES

[1] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss. *An Architecture for Differentiated Services, RFC 2475*. IETF Internet Engineering Task Force, dec 1998.

[2] C. Fraleigh, F. Tobagi, and C. Diot. Provisioning IP backbone networks to support latency sensitive traffic. In *INFOCOM 2003. Twenty-Second Annual Joint Conference of the IEEE Computer and Communications Societies. IEEE* , volume 1, pages 375–385, April 2003.

[3] S. Georgoulas, P. Trimintzios, and G. Pavlou. Joint Measurement-and Traffic Descriptor-based Admission Control at Real-Time Traffic Aggregation. In *ICC '04: IEEE International Conference on Communications*, 2004.

[4] M. Grossglauser and D. N. C. Tse. A framework for robust measurement-based admission control. *IEEE/ACM Transactions on Networking*, 7(3):293–309, 1999.

[5] J. Qiu and E. W. Knightly. Measurement-based admission control with aggregate traffic envelopes. *IEEE/ACM Trans. Netw.*, 9(2):199 – 210, 2001.

[6] T. K. Lee and M. Zukerman. Practical approaches for connection admission control in multiservice networks. In *Proceedings of IEEE ICON '99*, pages 172 – 177, Brisbane, Australia, 1999.

[7] S. Floyd. Comments on measurement-based admissions control for controlled-load services. In *Sally Floyd. Comments on measurement-based admissions control for controlled-load services. submitted to CCR*, 1996.

[8] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson. On the self-similar nature of Ethernet traffic (extended version). *IEEE/ACM Trans. Netw.*, 2(1):1 – 15, 1994.

[9] V. Paxson and S. Floyd. Wide-Area Traffic: The Failure of Poisson Modeling. In *SIGCOMM*, pages 257–268, 1994.

[10] M. E. Crovella and A. Bestavros. Self-similarity in World Wide Web traffic: evidence and possible causes. In *SIGMETRICS '96: Proceedings of the 1996 ACM SIGMETRICS international conference on Measurement and modeling of computer systems*, pages 160 – 169, New York, NY, USA, 1996. ACM Press.

[11] G. Mao. Finite Timescale Range of Interest for Self-Similar Traffic Measurements, Modelling and Performance Analysis. In *IEEE International Conference on Networks, ICON'03*, pages 7 – 12, Sydney, 2003.

[12] R. G. Addie. On the applicability and utility of Gaussian models for broadband Traffic. In *Proceedings of the International Conference on ATM*, Colmar, France, 1998.

[13] J. Beran. *Statistics for Long-Memory Processes*. Chapman & Hall / CRC, 1 edition, 1994.

[14] J. Kilpi and I. Norros. Testing the Gaussian approximation of aggregate traffic. In *Internet Measurement Workshop*, pages 49 – 61. ACM, 2002.

[15] W. Willinger, D. Alderson, and L. Li. A pragmatic approach to dealing with high-variability in network measurements. In Alfio Lombardo and James F. Kurose, editors, *Internet Measurment Conference*, pages 88–100. ACM, 2004.

[16] *Multivariate Density Estimation*. Probability and Mathematical Statistics. John Wiley & Sons, 1 edition, 1992.

[17] T. Karagiannis, M. Molle, M. Faloutsos, and A. Broido. A Nonstationary Poisson View of Internet Traffic. In *Infocom*, 2004.

[18] J. Cao, W. S. Cleveland, D. Lin, and D. X. Sun. On the nonstationarity of Internet traffic. In *ACM SIGMETRICS '01: Proceedings of the 2001 ACM SIGMETRICS international conference on Measurement and modeling of computer systems*, pages 102 – 112, New York, NY, USA, 2001. ACM Press.

[19] A. Scherrer, N. Larrieu, P. Owezarski, and P. Abry. Non Gaussian and Long Memory Statistical Characterisation for Internet Traffic with Anomalies. Technical report, Ecole Normale Superieure de Lyon, 2005.

[20] M. Zukerman and P. W. Tse. An Adaptive Connection Admission Control Scheme for ATM Networks. In *Icc (3)*, pages 1153 – 1157, 1997.

[21] K. Gopalan, T. C. Chiueh, and Y.-J. Lin. Probabilistic delay guarantees using delay distribution measurement. In *ACM Multimedia*, pages 900 – 907, 2004.

[22] M. Menth, J. Milbrandt, and S. Oechsner. Experience Based Admission Control. In *The Ninth IEEE Symposium on Computers and Communications ISCC2004*, Alexandria, Egypt, 2004.

[23] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Monographs on Statistics and Applied Probability. Chapman & Hall, 1 edition, 1986.

[24] M. C. Jones. The performance of kernel density functions in kernel distribution function estimation. *Statistics & Probability Letters*, 9(2):129–132, February 1990.

[25] L. Breslau, S. Jamin, and S. Shenker. Comments on the Performance of Measurement-Based Admission Control Algorithms. In *Infocom*, pages 1233 – 1242, 2000.

[26] J. Gozdecki, A. Jajszczyk, and R. Stankiewicz. Quality of service terminology in IP networks. *IEEE Communications Magazine*, 41(3):153–159, March 2003.

[27] ITU-T. *The Emodel: A computational model for use in transmission planning*. ITU-T Recommendation G.107. ITU-T, December 1998.

[28] A. Clark. Modeling the effects of burst packet loss and recency on subjective voice quality. IP Telephony Workshop, March 2001.

[29] A. P. Markopoulou, F. A. Tobagi, and M. J. Karam. Assessing the quality of voice communications over internet backbones. *TON*, 11(5):747–760, October 2003.