Self-Admission Control for IP Telephony Using Early Quality Estimation

Olof Hagsand[1], Ignacio Más[1], Ian Marsh[2], and Gunnar Karlsson[1]

[1] Department of Microelectronics and Information Technology
Royal Institute of Technology (KTH)
S-16440 Kista, Sweden
[2] Swedish Institute of Computer Science
Box 1263
SE-164 29 Kista, Sweden

Abstract. If quality of service could be provided at the transport or the application layer, then it might be deployed simply by software upgrades, instead of requiring a complete upgrade of the network infrastructure. In this paper, we propose a self-admission control scheme that does not require any network support or external monitoring schemes. We apply the admission control scheme to IP telephony as it is an important application benefiting from admission control. We predict the quality of the call by observing the packet loss over a short initial period using an in-band probing mechanism. The quality prediction is then used by the application to continue or to abort the call. Using over 9500 global IP telephony measurements, we show that it is possible to accurately predict the quality of a call. Early rejection of sessions has the advantage of saving valuable network resources plus not disturbing the on-going calls.

1 Introduction

Quality of service in the Internet has been researched for the last twenty years, yet its introduction has been extremely slow. Differentiated services [1] was originally proposed in 1997 to overcome scalability problems of previous proposals. However, DiffServ is still not widely offered by Internet service providers, perhaps due to the required upgrade in network infrastructure. Our proposal offers a light QoS for multimedia stream traffic, by a regulated admission of sessions, rather than a regulation of the flow rate per session. In human terms it is better to block a call that has little chance of being completed with adequate quality rather than allowing it to start and potentially degrading the system. Therefore, admitted sessions gain by having a high probability of being completed with decent quality. All these properties can be successfully accomplished by using admission control.

The purpose of this paper is to devise an efficient and flexible admission control scheme for IP telephony. Although IP telephony is used as the example real-time application in this work, it should be clear that there are no inherent restrictions on the applicability of the admission control scheme.

The admission control can be performed without explicit support from the network [2]. The procedure is in–band probing [3], in which the first seconds of the voice

N. Mitrou et al. (Eds.): NETWORKING 2004, LNCS 3042, pp. 381–391, 2004.

transmission are used as a probe stream. A new session is established only after estimating that the state of the network is acceptable. The receiver of the call measures the packet loss ratio of the first few seconds and estimates the packet loss probability. This estimated loss probability is compared to an acceptance threshold, which determines whether the session should be established or not. Loss levels above the threshold result in blocking of the new session and the sender should wait before establishing a new session. Hence, ongoing calls are protected from new calls that could deteriorate the overall quality to an unacceptable level by placing additional load on the network. The admission control being proposed is related to the out-of-band probing scheme being developed in our group [4,5,6,7].

We claim that measurements can produce data useful for predicting future quality. However, it is important to state we use only packet loss as the quality indicator of a VoIP session in this work. Packets arrive at a receiver 50 times per second (assuming no loss) in our VoIP scheme [8], so we have frequent sampling and observation of the network state. The measured loss after an initial number of seconds (zero to ten) is compared with the loss measured over the whole session. We use the correlation between the two measurements to determine how accurate the estimation is. This is possible as we have the whole session recorded at the receiver stored available for post processing[1].

The structure of this paper is as follows. The next section gives some background on how the empirical measurements were taken; we also describe how we measure the packet loss ratio for one call. Section 3 shows the results for all considered calls and offers a statistical analysis of the accuracy of the loss estimation for different initial time intervals, as well as blocking and error probabilities. Section 4 gives some conclusions of our work, some applications and pointers to future work. A preliminary version of this work was published in [9].

2 Measurement Description

This paper uses the results of previous work where approximately 23000 VoIP calls were measured between hosts at nine academic sites [10]. The locations of the sites are shown in Figure 1. The sites were connected as a full-mesh, allowing us, in principle, to measure the quality of 72 different Internet paths. These paths represent large differences in timezones, hop counts and geographic distances.

The measurements were performed over a period of 15 weeks in the following way: A call between two hosts was initiated on an hourly basis between a sender and a receiver. The sender transmitted a sequence of pre-recorded speech samples at 64 kbps as a stream of RTP/UDP/IP datagrams. The receiver made a detailed log of the arrival process, recording the reception time of each datagram. The complete details of the measurements are described in our previous work [10].

2.1 Reducing the Sample Set

For the purposes of this paper, we needed a common basis for our analysis, and therefore selected a subset of the 23000 calls. We only used calls that experience loss, since loss

[1] see also http://www.sics.se/~ianm/COST263/cost263.html

free calls do not provide any extra information for our analysis: both the probing and the total loss rate are zero giving perfect correlation. A large percentage of the calls are in fact loss free which reduces the sample set somewhat. We attribute the large number of loss free sessions to the fact that the sites are located on well provisioned (academic) networks. This restriction resulted in a subset of 9683 calls. Despite this reduction, all nine sites are represented in the subset.

2.2 Measuring a Single Call

Figure 2 shows the loss process of a sample call as observed by a receiver. The call was made between the Argentinian and Turkish sites. The figure shows a loss pattern that is representative of many other calls in the subset. The plot shows the number of lost packets on the y-axis versus time on the x-axis. It can be seen that the number of lost packets increases almost linearly as the call proceeds.

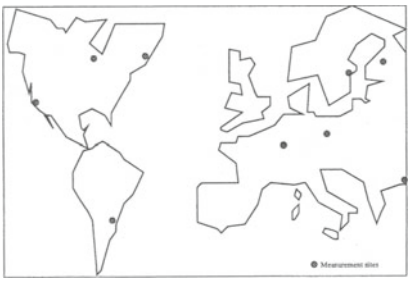

Fig. 1. Measurements were made between nine academic sites worldwide.

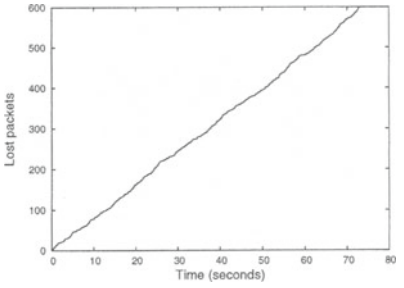

Fig. 2. Loss process of a single sample call between Turkey and Argentina.

Figure 3 shows the cumulative loss ratio for the same call. This ratio is defined as the number of lost packets divided by the number of sent packets. We show the cumulative plot to clarify how long we need to measure to obtain a good estimation of the final loss ratio. From the plot, we see that the final loss ratio for the complete call is approximately 18%.

In Figure 4 we show the first 20 seconds of the same call. From the figure, we see that the initial loss is approximately 14% after one second and 19% after ten seconds. These are early estimations of the final loss rate. We want to know how accurate such early estimations are. Therefore we need to study the relation between the loss ratio of an initial part of the call and the loss ratio of the whole duration.

3 Analysis

In the preceding section, only one call was considered. In Figures 5 and 6, the loss ratio for the whole call is plotted versus the loss ratio of an initial interval for all calls in the selected subset. In the figures, every point represents one call. The plots show that as the

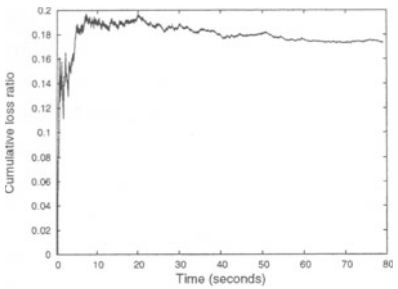

Fig. 3. Cumulative loss ratio of a single sample call between Turkey and Argentina.

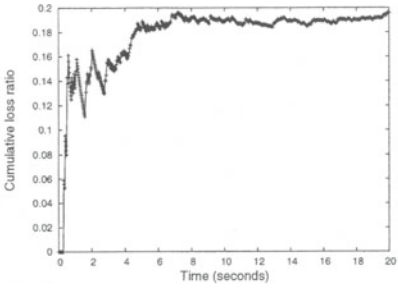

Fig. 4. Cumulative loss ratio of the same session showing only the initial portion of the call.

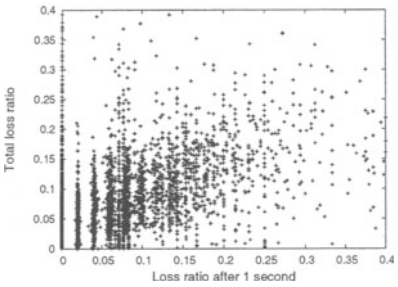

Fig. 5. Relation between the loss ratio after one second and the total loss ratio for all calls

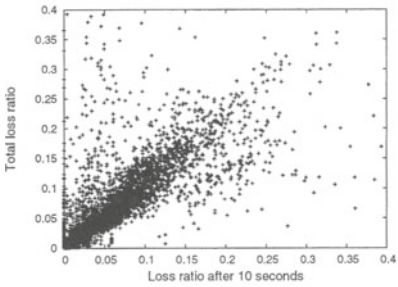

Fig. 6. Relation between the loss ratio after ten seconds and the total loss ratio for all calls

initial interval increases, the points group closer around the line $y = x$. In other words, the correlation increases and the estimation improves.

The plots in Figures 5 and 6 give an intuitive measure of the correlation between the loss ratio of the initial interval and the total call. In order to evaluate more precisely the accuracy of the estimation, we computed the actual correlation factor as a function of the initial interval. The result is plotted in Figure 7.

Figure 7 shows that the correlation factor increases as the probing interval increases. From the figure, we can clearly see that the correlation stabilizes after four seconds. This is important, because after this point no further estimates are necessary.

The relation between the loss ratio of an initial interval, l_p, and the loss ratio of the total call, l_t can be further examined by forming an error function, such as $l_t - l_p$, and analyzing it as a stochastic variable ϵ.

Figure 8 shows histograms of ϵ for probing intervals of one, four and ten seconds. In the histograms, positive values represent calls where the initial loss rate is smaller than the total loss rate, i.e., $l_p < l_t$. In other words, those calls experienced a higher packet loss after the probing: the quality of the calls deteriorated after the initial interval. Likewise, negative values represent calls where the initial loss rate is greater than the total loss

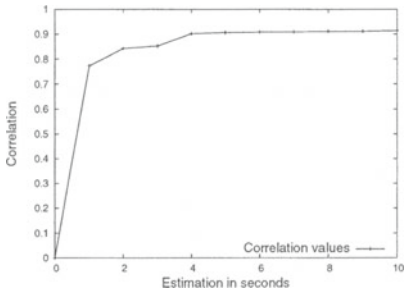

Fig. 7. The correlation factor as a function of the initial probing interval

rate, i.e., $l_p > l_t$. Note that the histograms represent probability density functions of ϵ that are not normally distributed.

Based on the values in the histograms we calculated the confidence intervals by counting the number of samples around $\epsilon = 0$ that sum up to the desired confidence level. The result is shown in Table 1.

Table 1. Table showing confidence levels and intervals of the error function $\epsilon = l_t - l_p$ for probing intervals one, four and ten seconds.

Level	interval (1 second)	interval (4 seconds)	interval (10 seconds)
0.75	$[-0.0288, 0.0336]$	$[-0.0141, 0.0187]$	$[-0.0087, 0.0151]$
0.80	$[-0.0424, 0.0436]$	$[-0.0180, 0.0260]$	$[-0.0124, 0.0213]$
0.85	$[-0.0608, 0.0568]$	$[-0.0252, 0.0344]$	$[-0.0183, 0.0299]$
0.90	$[-0.0848, 0.0752]$	$[-0.0416, 0.0496]$	$[-0.0280, 0.0424]$
0.95	$[-0.1768, 0.1144]$	$[-0.1200, 0.0800]$	$[-0.0888, 0.0696]$
0.99	$[-0.4000, 0.2536]$	$[-0.3200, 0.2216]$	$[-0.2480, 0.2144]$

Based on the table, we can express to what degree we can trust an initial observation. For example, if we measure the loss ratio l_p of a call after four seconds, we can be 80% certain that the total loss of the call will be in the interval $[l_p - 1.8\%, l_p + 2.6\%]$.

While the confidence intervals may be useful in themselves to express confidence in an observed value, forming the cumulative distribution function (*cdf*) of ϵ is more useful when an upper bound on the final loss value is of interest. This is typically the case in admission control scenarios, where we want to block calls that we believe will experience a loss higher than a certain threshold.

Table 2 shows the *cdf* of ϵ. Using the table, we can make statements such as: Given a probing loss and a confidence level, the final loss will be bounded by the probing loss plus a value given by Table 2. Figures 9, 10 and 11 show the *cdf* of ϵ in graphical form.

The *cdf* of ϵ can directly be used for admission control purposes. The table gives us the percentage of calls that have an error less or equal to the value of ϵ:

$$P(\epsilon < l_a - l_p) \geq confidence \quad level$$

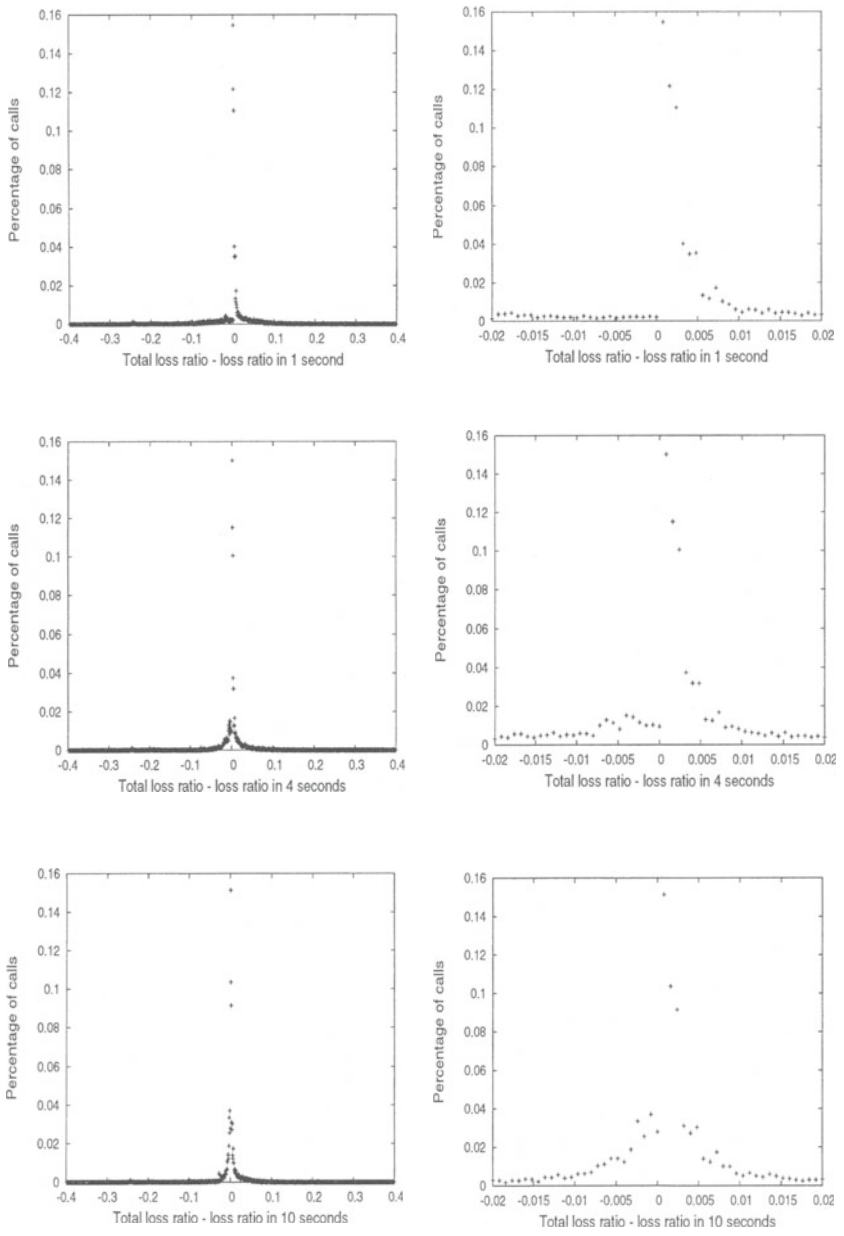

Fig. 8. Histograms of the error $\epsilon = l_t - l_p$ for initial probing intervals one, four and ten seconds. Each histogram is shown in full view on the left, while the right plot shows an enlarged region around $\epsilon = 0$.

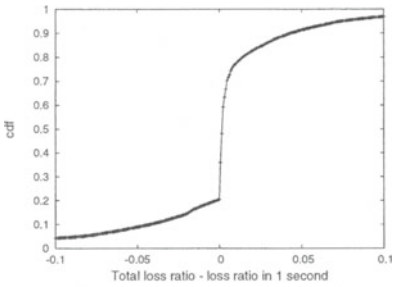

Fig. 9. Cumulative distribution function of $\epsilon = l_t - l_p$ for a probing interval of one second.

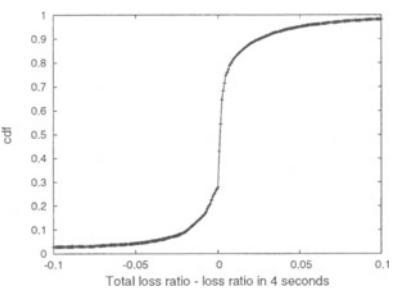

Fig. 10. Cumulative distribution function of $\epsilon = l_t - l_p$ for a probing interval of four seconds.

Table 2. Table showing cumulative values of the error function $\epsilon = l_t - l_p$ for probing intervals one, four and ten seconds.

Confidence level	1 second	4 seconds	10 seconds
0.05	−0.0848	−0.0416	−0.0280
0.1	−0.0424	−0.0181	−0.0101
0.2	−0.0016	−0.0056	−0.0038
0.3	–	–	−0.0008
0.4	–	–	–
0.5	0.0018	0.0014	0.0009
0.6	0.0025	0.0020	0.0017
0.7	0.0042	0.0038	0.0030
0.8	0.0144	0.0086	0.0069
0.9	0.044	0.0260	0.0212
0.95	0.0752	0.0496	0.0424

For example, suppose the aim of a strict admission control scheme using four seconds probing is to drop calls that have a higher risk than 10% to surpass a pre-established loss rate l_a. Retrieving the value of ϵ from Table 2 shows that $l_a - 2.6\%$ is a good threshold. A more relaxed policy could have the aim to reject all calls that have more than 90% risk to surpass l_a. In that case, again using Table 2, the threshold is $l_a + 1.81\%$. The strict and relaxed policies outlined above both have drawbacks. With a strict policy, most bad calls ($l_t > l_a$) will be blocked, along with a large number of good calls ($l_t < l_a$). A relaxed policy admits most good calls, while admitting many bad calls.

Table 12 shows a classification of calls with respect to an admission control strategy: classes AG and AB represent calls that were admitted while classes RG and RB represent calls that were blocked. Further, classes AG and RB represent categories where the admission control decision was correct. Classes AB and RG represent decisions that were wrong. An admission control policy based on probing, needs to consider the trade–off between classes.

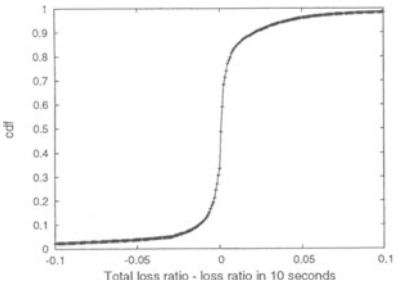

Fig. 11. Cumulative distribution function of $\epsilon = l_t - l_p$ for a probing interval of ten seconds.

	Good calls $l_t < l_a$	Bad calls $l_t > l_a$
Admitted $l_p < l_\alpha$	AG	AB
Rejected $l_p > l_\alpha$	RG	RB

Fig. 12. The table shows the different kinds of calls based on the initial estimation and the final outcome. l_α denotes an admission threshold applied after a probing interval, while l_a is the desired upper bound on the loss level.

If we return to the strict policy introduced above, it minimizes class AB while class RG is large, thus protecting on–going calls in a more successful manner whilst increasing the blocking probability. In the same way, the relaxed policy minimizes class RG, thus reducing the blocking probability at the risk of a higher number of bad calls.

To obtain absolute numbers on the number of calls in the classes, a real loss distribution has to be considered. By aiming at an upper bound of the loss rate and applying the *cdf* to that bound, it is possible to get absolute numbers of the different classes. The admission threshold can then be varied to find a desired optimum.

Figure 13 shows an example of a uniform loss distribution (calls can experience any packet loss rate between 0 and 100% with equal probability) with a desired upper bound on the loss rate l_a. The *cdf* for four seconds in Figure 10 has been superimposed[2] on the uniform loss distribution for two admission thresholds, strict policy and relaxed policy. The number of calls belonging to each class can be determined by the areas in the graph. The areas are bounded by l_α and the *cdf*. For example, it can be seen from the graph that area RG (rejected calls that turned out good) is large in the strict policy, but is small in the relaxed. Likewise, area AB (admitted calls that turned out bad) is small in the strict and large in the relaxed policies.

A uniform loss distribution is evidently unrealistic, but the same methodology can be applied for a real loss distribution. We have applied the method to the complete set of 9683 error-free calls in the measurements in the case of four seconds probing and calculated the percentage of calls that fall in each of the areas. The rest of this section deals with this case.

Figure 14 gives the blocking probability (RG+RB) for the complete sample space. From the figure it can be seen that rejecting calls that experience an initial loss rate equal or higher than 10% gives a blocking probability of around 15%, while a more stringent packet loss rate threshold would result in a rapidly increasing blocking probability. Note however, since the error-free calls are omitted, the blocking probability is overly pessimistic. We would expect a lower blocking probability with a factor of around three if the error free calls were included.

[2] Note, the *cdf* is reflected around x=0.

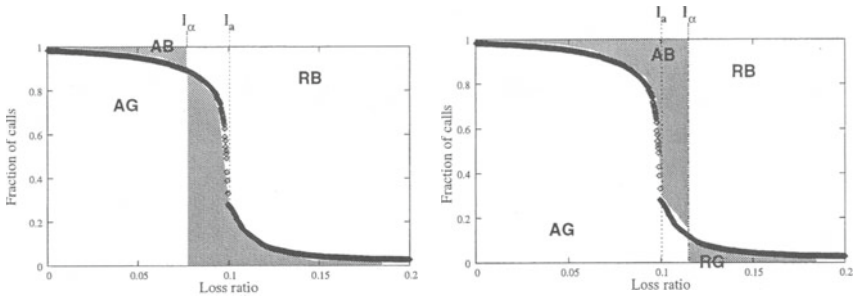

Fig. 13. Example showing the result of imposing admission control decision in the strict (left) and the relaxed (right) admission policy with a uniform loss distribution. The desired upper bound on packet loss is l_a and the imposed threshold is l_α.

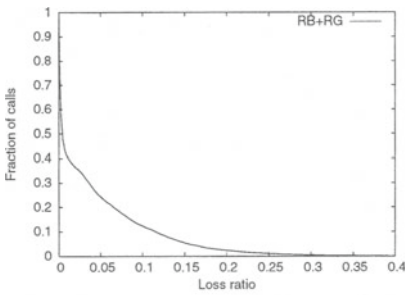

Fig. 14. Blocking probability as a function of the packet loss rate admission threshold.

As was previously noted, the accuracy of an admission control policy can be measured by counting the correct and incorrect decisions. Figure 15 shows the incorrect decisions for a packet loss rate target of 2%. The plot shows both kinds (AB and RG) as well as their sum.

The plot illustrates how the number of incorrectly admitted calls increases as the admission threshold is relaxed, while the incorrectly rejected calls decreases. The sum of the two functions has a minimum for a particular admission threshold at 1.8%, which can be considered as an optimum operating point. That is, a minimal number of incorrect decisions were made at this threshold.

Finally, Figure 16 illustrates the sum of incorrect decisions for different target loss rates as the acceptance threshold is varied. The results show a minimum close to the value of the target loss rate, as was intuitively expected.

The choice of an operating point for the admission control has to take into account many parameters. We can always increase the accuracy by measuring for a longer period. However, increasing the probing period reduces the advantages, since we are extending the period in which a bad call is disturbing the ongoing calls, reducing the overall quality in the process. Also, longer probing times increase the frustration in the case of a rejection.

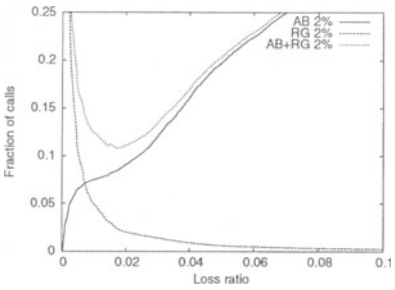

Fig. 15. Erroneous decisions as a function of the admission threshold for a 2% target loss rate

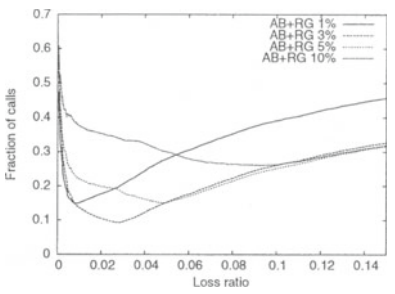

Fig. 16. Erroneous decisions as a function of the admission threshold for different target loss rates.

To summarize, if we use our measurements, we would probe for four seconds and use an admission threshold close to the targeted value. Assume that 2% packet loss is acceptable. In this case, the admission threshold should also be around 2%, which would give a blocking probability of 36%. The admission control decision would then have failed 11% of the time, the majority would be calls that were admitted although they turned out to be bad (9% of the total calls), a smaller fraction would be calls that were rejected but turned out to be good (2%).

4 Conclusions and Future Work

This paper proposes a quality differentiation scheme based on self-admission control without the need of infrastructure changes. The admission control is performed at the application layer and can provide statistical bounds on the packet loss rate that stream flows will experience in the network. We have shown how the admission control mechanism can be devised by blocking calls experiencing an initial loss rate exceeding an admission threshold. An initial admission threshold is motivated by two factors: (1) it makes sense to drop calls that will experience bad quality and thus reduce congestion in the network so that other calls may experience better quality; (2) an audio coder may have an upper bound on quality: exceeding a drop rate will result in unacceptable audio quality.

We have evaluated the admission control scheme by analyzing a large number of IP telephony calls that were made over the Internet. Based on this empirical data, we have shown that it is possible to predict the quality of a call by making an early measurement of the packet loss. From our particular data, we have shown that it is sufficient to make an estimation after four seconds. The analysis we have performed offers thresholds for call blocking probability and failure rates of the scheme.

From a practical point of view, the admission control scheme shown in the paper could be implemented using standard RTCP [11] receiver reports. A small adjustment of the rate that the receiver generates the reports would be enough for our probe-based admission control scheme.

One limitation with our method is that all calls in the experimental data are in fact admitted. The effects of dropping calls to the network as a whole has not been assessed. We claim that this observation is irrelevant in this study for two reasons: (1) all of the calls in the study were disjoint in time; (2) the effect could only be positive, thus our results can be seen as worst-case.

An interesting point is whether the results based on the measured data [10] are generally valid. This is a difficult question, and we cannot claim that the results hold for all network conditions. For example, one could claim differences in timescales (the measurements were made in 2001), networks (most data were made on academic networks), link technologies (no wireless access were available, etc). We hope that future work can help to get a larger understanding of such conditions.

References

1. S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss, "An architecture for differentiated services," RFC 2475, IETF, December 1998.
2. J. H. Saltzer, D. P. Reed, and D. D. Clark, "End-to-end arguments in system design," *ACM Transactions on Computer Systems*, vol. 2, pp. 277–288, November 1984.
3. L. Breslau, E. W. Knightly, S. Shenker, I. Stoica, and H. Zhang, "Endpoint admission control: Architectural issues and performance," in *Computer Communication Review – Proc. of Sigcomm 2000*, vol. 30, (Stockholm, Sweden), pp. 57–69, ACM, August/September 2000.
4. G. Karlsson, "Providing quality for internet video services," in *Proc. of CNIT/IEEE ITWoDC 98*, (Ischia, Italy), pp. 133–146, September 1998.
5. V. Fodor (née Elek), G. Karlsson, and R. Rönngren, "Admission control based on end–to–end measurements," in *Proc. of the 19th Infocom*, (Tel Aviv, Israel), pp. 623–630, IEEE, March 2000.
6. I. Más Ivars and G. Karlsson, "PBAC: Probe–based admission control," in *Proc. of QoFIS 2001*, vol. 2156 of *LNCS*, (Coimbra, Portugal), pp. 97–109, Springer, September 2001.
7. I. Más, V. Fodor, and G. Karlsson, "The performance of endpoint admission control based on packet loss," in *Proc. of QoFIS 2003* [12].
8. O. Hagsand, I. Marsh, and K. Hansson, "Sicsophone: A low-delay internet telephony tool," in *Proc. of the 29th Euromicro Conference*, (Belek-Anatalya, Turkey), pp. 189–197, September 2003.
9. P. Biyani, O. Hagsand, G. Karlsson, I. Marsh, and I. Más, "Early estimation of voice over ip quality," in *Proc. of the 21st Nordunet network conference*, (Reykjavik, Iceland), August 2003.
10. I. Marsh, F. Li, and G. Karlsson, "Wide area measurements of voice over IP quality," in *Proc. of QoFIS 2003* [12].
11. H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson, "RTP: A transport protocol for real–time applications," RFC 1889, IETF, January 1996.
12. *Procedings of the 4th COST 263 International Workshop on Quality of Future Internet Services*, vol. 2856 of *LNCS*, (Stockholm, Sweden), Springer, October 2003.