# Improving Quality in a Distributed IP Telephony System by the use of Multiplexing Techniques

Jenifer Murillo, Jose Saldana, Julián Fernández-Navajas, José Ruiz-Mas, José I. Aznar, Eduardo Viruete Navarro
Communication Technologies Group (GTC) – Aragon Inst. of Engineering Research (I3A)
Dpt. IEC. Ada Byron Building. CPS Univ. Zaragoza
50018 Zaragoza, Spain
{jenifer.murillo, jsaldana, navajas, jruiz, jiaznar, eviruete}@unizar.es

*Abstract*— **Nowadays, many enterprises use Voice over Internet Protocol (VoIP) in their telephony systems. Companies with offices in different countries or geographical areas can build a central managed telephony system sharing the lines of their gateways in order to increase admission probability and to save costs on international calls. So it is convenient to introduce a system to ensure a minimum QoS (Quality of Service) for conferences, and one of these solutions is CAC (Call Admission Control). In this work we study the improvements in terms of admission probability and conversation quality (R-factor) which can be obtained when RTP multiplexing techniques are introduced, as in this scenario there will be multiple conferences with the same origin and destination offices. Simulations have been carried out in order to compare simple RTP and TCRTP (Tunneling Multiplexed Compressed RTP). The results show that these parameters can be improved while maintaining an acceptable quality for the conferences if multiplexing is used.**

***Keywords**- IP telephony; QoS; R-factor; RTP multiplexing; software PBX; TCRTP; VoIP*

## I. Introduction

In the last years many enterprises are replacing their old PSTN telephony systems by new ones that use Voice over Internet Protocol (VoIP) so telephone calls and video conferences can be carried out using Internet. One of their objectives is cost saving, taking advantage of their Internet connections.

But many of these enterprises have their resources decentralized, so each office is independent from the rest and it manages its Internet connections and gateways' lines. By making a global management of the resources, enterprises could save costs and increase admission probability. Many conferences could then be established using Internet connections.

A software PBX can be used in order to join the offices' telephony systems, building a central managed system which allows some advantages, like sharing gateways' lines by offices in different locations. International calls can then be set up in two steps: one using Internet to the destination country, and a local call to the final user.

VoIP is a real-time service that has to work in a network that was initially designed for *best effort* services, but users demand a Quality of Service (QoS) similar to the one they were used to have with traditional telephony systems. This has led researchers to deploy solutions capable to add quality to IP networks. One of them is Call Admission Control (CAC), a well known system that accepts or rejects new conferences in order to avoid service degradation, both for new conferences and for the ones that are already established. More specifically, parameter-based CAC [1], counts the number of conferences that are simultaneously established in each office, and it only accepts new conferences if this number is below a limit which is decided at configuration time. In this work we will obtain QoS with an estimator proposed by ITU G.107 [2], which is R-factor. It rates calls from 0 to 100, and acceptable values are considered from R > 70.

In an enterprise with many offices, each one with a big number of users, a number of conferences can be simultaneously established between the same pair of offices (Fig. 1). In this case, the use of multiplexing techniques could achieve some improvements. If we introduce in the same packet the samples from different conferences, we can save bandwidth while only adding small delays. In this work we use TCRTP (Tunneling Multiplexed Compressed RTP) multiplexing technique, which was approved by IETF in 2005 as RFC 4170 [3], with the category of "Best Current Practice". It will be explained with more detail in next section.

As the use of multiplexing reduces the bandwidth usage and there is a CAC system implemented, the quality of the calls (in terms of R-factor) is expected to be improved. In order to study the benefits of distributed telephony systems, in previous works [4] we built a scenario matching an enterprise with several offices in different areas (maybe different countries). In its first stages it was implemented in a testbed using emulation, but this introduced a size limitation because of the number of machines required.

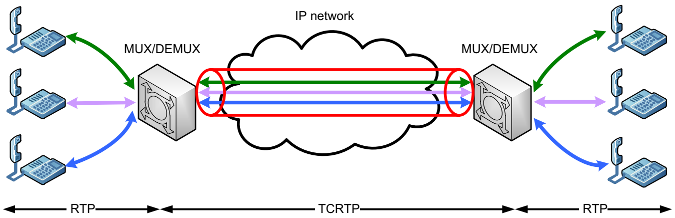

Figure 1. Conferences sharing the same source and destination offices

This work has been partially financed by CPUFLIPI Project (MICINN TIN2010-17298), MBACToIP Project, of Aragon I+D Agency and Ibercaja Obra Social, and NDCIPI-QQoE Project of Catedra Telefonica, Univ. of Zaragoza.

Simulation will help us to study the behavior of the system when gateways' lines are shared and when multiplexing is used, in terms of admission probability and R-factor. In this work we have used the previous results obtained on the testbed, like delay and packet loss, as simulation parameters, thus using a hybrid approach.

This paper is organized as follows: section II discusses the related works. System architecture is presented in section III. The next section covers the test platform. Section V presents the measurements that have been carried out. The last section details the conclusions of the present work.

## II. Related Works

In this work we will show the advantages of using a central managed VoIP system to achieve a better admission probability while maintaining QoS, and how multiplexing techniques are adequate to improve R-factor. Other related issues are CAC systems and queuing policies of the router's buffer, which may have a big influence on QoS parameters.

Nowadays there is a tendency in business environment to try to obtain cost savings by the introduction of new solutions like VoIP. In [5] Intel published the results of a pilot program in which several employees used VoIP based on SIP (Session Initiation Protocol) in their production environment. They checked improvements in terms of cost savings and user productivity, coming to the conclusion that VoIP technology is very beneficial to enterprises. Reference [6] shows another study that illustrates the improvements obtained by the use of VoIP instead of PSTN. They obtain savings on hardware cost, provisioning, billing, maintenance and service. Furthermore, they support VoIP systems as the new telecommunications solution for enterprises.

Regarding to CAC systems, there have appeared several studies and developments [7]. As a summary, it can be said that solutions based on storing information about network resources usually have more scalability and implementation problems than solutions based on measurements of network conditions. Some CAC systems use parameter-based mode, in which initial tests are needed at configuration time. Measurement-based CAC makes active or passive tests when the call is going to be established, but at the cost of adding delays and overhead to the system. In [1] a parameter-based CAC was added to a commercial environment using H.323.

Regarding to buffer policies, different options were studied in [8]. They are mostly centered on TCP connections in core routers. One of these methods is Stanford Model, which recommends the use of small buffers instead of using the "rule of the thumb" based on Bandwidth-Delay Product. In the same work, it is presented a buffer that limits the maximum queuing delay. We have used that approach in this study, as it can be useful to maintain packet delays under a defined limit.

Finally, we will focus on multiplexing techniques. In last years, there have been developed several techniques in order to reduce IP overhead. Some protocols were defined, such as IPHC [9], CRTP [10], ECRTP [11] and ROHC [12]. In this work we have used TCRTP [3], a standard that combines ECRTP with other multiplexing and tunneling protocols. It will be explained in the next section.

## III. System Architecture

We will now make a short summary of the main characteristics of the SIP-based telephony system presented in [4]. We will work with two modes: the *original* one, in which all the RTP flows are independent from the rest, and the *multiplexing* one, in which the traffic of the conferences that share the same source and destination offices is multiplexed using TCRTP. The scenario corresponds to an enterprise with several offices in different areas (maybe different countries), which is equivalent to some commercial solutions [13].

As we can see in Fig. 2, an office has several users, and a gateway connects it with PSTN. Each office has an Internet access with a limited bandwidth, and the voice calls are controlled by a CAC system, as we will see later. A software PBX is located in the Data Center, which is connected to the IP network. To reduce management expenses, the dial plan is kept at the PBX, and not distributed to the offices. Furthermore, Internet is used for telephone traffic delivery among offices, instead of dedicated lines. We assume that the system does not use any bandwidth reservation protocol, and we will also assume that VoIP is the only real-time traffic we are going to manage in a special way. In order to prevent bottlenecks in the Data Center, the PBX only takes part in the conferences' signaling, while RTP traffic is directly sent between offices. Besides, the PBX manages the gateways of all the offices, introducing the possibility of sharing their lines with the others.

On one hand SIP traffic uses a centralized system, so PBX acts as a B2BUA (Back to Back User Agent). On the other hand, RTP uses a star topology, so each pair of offices is directly connected with a tunnel of multiplexed packets. Star topology avoids the delays that would appear if RTP traffic had to go through the PBX.

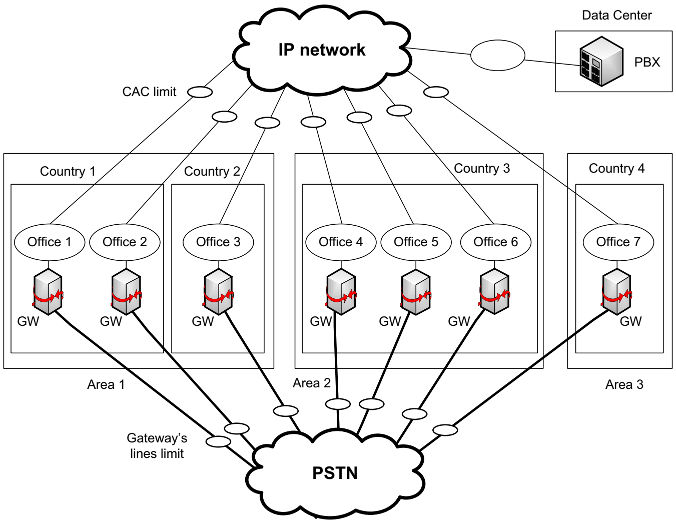

Figure 2. System architecture

A parameter-based CAC system has been set up to assure a minimum quality for VoIP conferences. We will not study CAC systems in this paper, but we will use this particular one because it is enough for the aims of this research. In every office there exists a local agent which includes a SIP proxy in order to implement CAC decisions. The parameter used in the CAC system is the maximum number of conferences allowed in each office's link, which is established after some traffic measurements that are carried out at configuration time.

As we have previously said, we have used TCRTP as multiplexing technique. Fig. 3 presents its protocol stack and Fig. 4 shows the scheme of a multiplexed packet. First, ECRTP header compression scheme compresses IP, UDP and RTP headers into a new one. Next, PPPMux is used and finally the packets are sent using a PPP and L2TP tunneling scheme. ECRTP is used end-to-end, because if a tunnel is used it is not necessary to apply ECRTP on each router of the path. A packet sent from the origin has to be the same at the destination, so the multiplexer-demultiplexer system has to be transparent for the communication. And the demultiplexer needs some information to rebuild the original packet and deliver it to its destination. This information is called *context.*

When *multiplexing mode* is used, we have to realize that a number of conferences in a tunnel occupy less bandwidth than if the conferences are sent separately. So the CAC system has to consider this in order to take the admission decision, as calls can no longer be counted in the original mode. The first idea we thought was to score the bandwidth of each tunnel depending of the number of conferences that contains, so CAC system could limit the bandwidth of each office's access instead of the number of conferences. According to Fig. 4, we can obtain the bandwidth for a tunnel with *k* conferences as:

$$E[BW_{kflows}] = E[PS_{kflows}] / IT = (CH + k\,(MH + E[RH] + S)) / IT \quad (1)$$

Where:

- **PS$_{\mathbf{kflows}}$:** Size of a packet multiplexing *k* flows.
- **CH:** Common Header. It is the size of the header shared by the whole multiplexed packet. It corresponds to IP/L2TP/PPP headers. Its value is 25 bytes.
- **MH:** PPPMux Header. It is included at the beginning of each compressed packet: 2 bytes
- **RH:** Reduced Header. It is the size of the reduced header that precedes the samples of each RTP flow. Compressing protocols generate a number of possible reduced header sizes, so *E[RH]* has to be calculated as the expected value of the header size, taking into account the probability of having each size. We have used the values of [14] for the probabilities of each header size.
- **S:** Samples. It is the size of the voice samples of one RTP packet. We have use G.729a codec with two samples per packet, so S = 20 bytes.
- **IT:** generation time of voice packets. For the used codec it has a value of 20 ms.

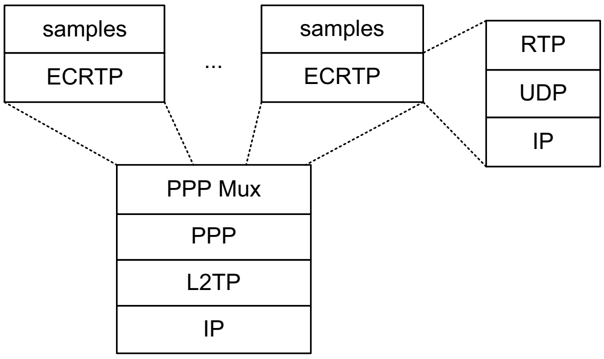

Figure 3. TCRTP protocol stack

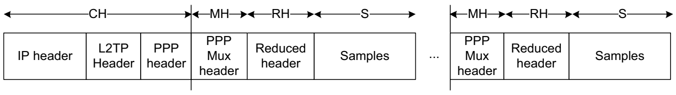

Figure 4. Scheme of a multiplexed packet

We can see that for a fixed codec, (1) grows linearly with *k*, so finally we have decided that CAC system has just to count conferences, independently of the bandwidth of the tunnels.

A CAC is used to assure a minimum quality for the new conferences and the established ones, at the cost of rejecting some of them. The local agent not only controls the number of conferences that use the Internet connection, but it is also in charge of the count of free and engaged lines in its office's gateway, so if the gateway has all the lines engaged, the local agent redirects the conference to another gateway, always trying to find the cheapest route.

In the previous work this system was probed in a testbed using real-time emulation, but in the current work we also use simulation. Besides, an advantage of using simulation is that the number of offices may be high, like it happened in [1]. Furthermore, the real-time emulation system has been used to obtain some values, such as OWD (One Way Delay) and packet loss, which are necessary to obtain the maximum number of conferences that can be simultaneously established with an acceptable quality. With these parameters, we can also calculate R-factor values for the conferences' audio, using E-Model [2].

## IV. Test Platform

We have used the testbed presented in [15] in order to implement the scenario. We use a hybrid approach, combining emulation and simulation. Next, we will explain them with more detail. A test is divided in the next stages, as it can be seen in Fig. 5:

- Emulation, in which there are three machines involved. First, the traffic generator sends RTP and background traffic. Later, these traffics go through the router, which emulates different buffer policies. And finally there is a receiver to capture the traffic.
- Offline processing, where some delays are added to the traffic captured in the previous stage.

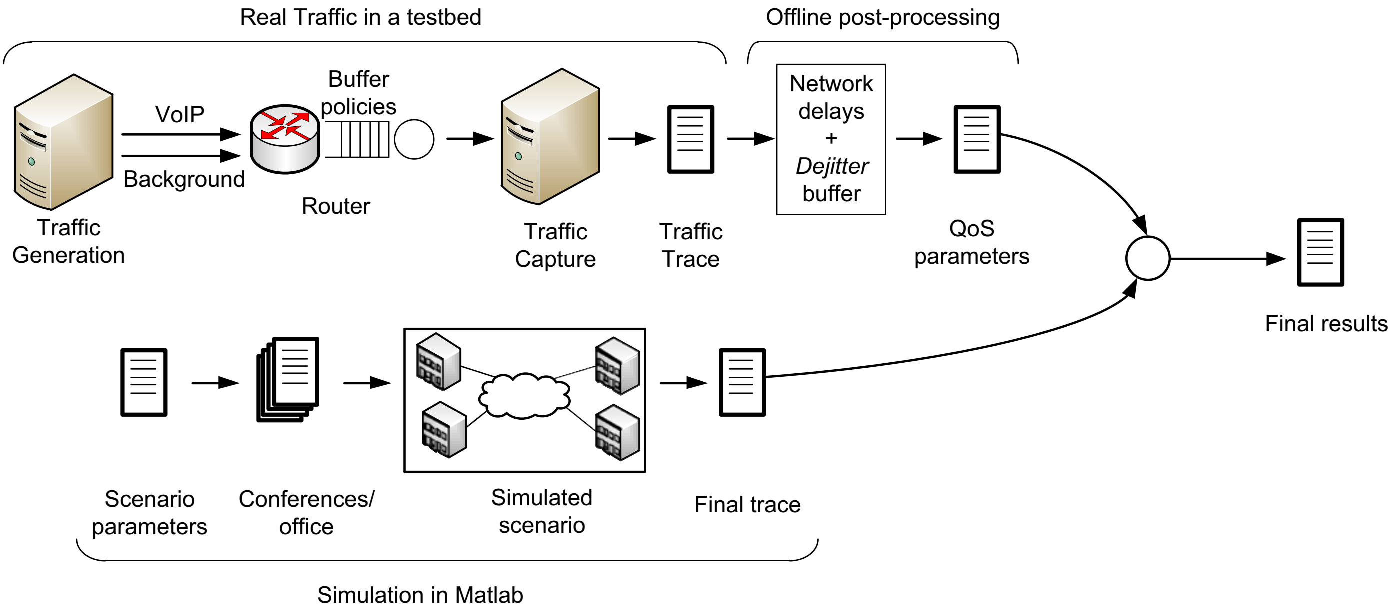

Figure 5. Measurement diagram

- Simulation of the whole scenario and final results. We use Matlab in order to build a bigger scenario, and we use the results of the second stage to obtain final results such as admission probability and R-factor.
- The final results are obtained by merging of the simulation and testbed ones.

### *A. Emulation Network*

We have used the testbed presented in [15] in order to implement a reduced version of the system and to obtain QoS parameters which will later be included in the simulations. For the implementation, some off-the-self software solutions were used. Thus, the PBX is the version 1.6 of Asterisk, the SIP proxy uses the version 1.4 of OpenSIPS and gateways and softphones are PJSUA 1.0. Traffic generator D-ITG is used to generate background traffic and saturate the access's link in each office. Furthermore, in order to limit the bandwidth, there is a queuing policy in each office's router, implemented with a Linux tool called Traffic Control (*tc*). It can set up different buffer sizes with some parameters as bandwidth, latency limit, buffer limit or the size of the bursts. Traffic amounts have to be properly corrected, because *tc* takes into account level-2 headers (Ethernet in our case) to calculate bandwidth limit.

### *B. Offline Processing*

We have considered different delays in our system, as we can see in Fig. 6. Packetization delay depends on the codec, retention delay is related to multiplexer, there is also a process delay introduced by the multiplexer and the demultiplexer, queuing delay depends on the buffer policy of the origin router and there is a dejitter buffer in the destination machine which also adds a delay.

The sum of all these delays should not exceed the value of 150 ms recommended by ITU in G.114 [16]. A deeper study of these delays can be found in [17].

With this testbed we have obtained OWD and packet loss in two situations: in one hand, when different number of RTP flows is sent through the network and in the other hand when different number of flows is multiplexed in tunnels. In the Internet access of each office we will have a tunnel for each of the rest of the offices, so if we denote the maximum CAC limit as *C* and the number of offices as *O*, then we will need a number of tests corresponding to the combinations with repetition of *C+1* values for the conferences (zero conferences is also included) and *O-1* values for the offices, as an office will establish a tunnel with every one of the rest. So the number of required tests is:

$$\mathrm{CR}_{\mathrm{C+1,O\text{-}1}} = \binom{C+O-1}{O-1} \qquad (2)$$

As an example, when having a CAC limit of 15 flows and 4 offices, in the simple RTP situation only 16 tests were necessary, but in the multiplexed one 815 tests have to be carried out, corresponding to a combination without repetitions of 16 and 3 values. We will use these parameters to calculate R-factor in different moments of a call (in our case conference's audio).

### *C. Simulation*

We have used Matlab to generate executions with different parameters, such as number of offices, users, gateways' lines, countries, areas, establishment delays, etc. Each execution contains information of all the conferences, such as start time, duration and extensions involved. Call arrivals follow a Poisson distribution. The conferences' duration has been modeled using a Gaussian distribution, of 180 seconds average and a variance of 30. Finally, different probabilities to generate the destination of the call can be set.

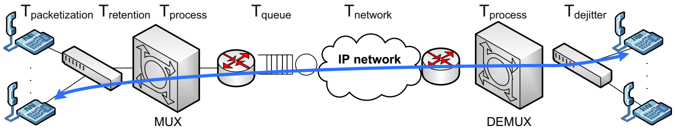

Figure 6. System delays

Later, each execution can be simulated using two different algorithms. The first one, from now called *isolated mode*, is used as a reference. It does not implement a CAC system and each office is independent from the rest, so the system does not share the gateways and there are no redirected conferences. Thus, all the conferences whose destination is PSTN do not go through the IP network. The second one, from now called *sharing mode*, simulates a central managed system, which shares all the gateways with all the offices and implements a CAC system, letting the local agent redirect conferences to other routes, always looking for the cheapest way. With this algorithm, we try to set up most of the conferences by means of Internet.

*Sharing mode* is used to exchange blocking probability in gateways for blocking probability in the IP network, as it is usually cheaper to contract more Internet bandwidth than more gateways' lines. So we expect the admission probability to increase, because of redirecting and we can save costs of international conferences too. But increasing admission probability involves more traffic introduced in the Internet access, making QoS decrease.

When the simulation ends, we use the QoS parameters obtained with the testbed in order to estimate R-factor values. Although our simulation system does not distinguish individual packets, it is able to manage parts of conferences. The value of R in each interval depends on the number of simultaneous conferences in a link. So we can estimate not only the average R-factor for a conference, but values in different conference's moments too. We assume that each interval of the conference has no influence in the next, which means that buffers' occupancies are independent between intervals. We have done this assumption because buffer delay limit is 80 ms, while parts of the conferences are usually tens of seconds long. R-factor is considered acceptable if it is above 70.

## V. Tests and Results

The aim of the tests is to show the improvements in admission probability and R-factor when offices' lines are shared and RTP multiplexing is used. Both parameters have an influence on the QoE (Quality of Experience).

### *A. Admission Probability by sharing gateways*

The first tests calculate admission probability for *isolated* and *sharing mode*. This parameter will be the same when using *original* or *multiplexing mode*, because, as we have said previously, CAC system has just to count conferences. But admission probability has to be calculated first in order to know the values of CAC limit that will be acceptable.

To obtain these results, we simulate several executions, in which all the conferences are addressed to the gateways. The variable parameters are the generation conferences' rate (λ) and the number of offices of each country.

Fig. 7 shows that, when using *sharing mode*, the bigger the number of offices, the bigger the admission probability. This fits with the idea of Erlang formula "the more you share, the more you get". Of course, if λ becomes very big, admission probability decreases.

We have done another test including CAC limitation. In Fig. 8 it can be seen that *sharing mode* provides better values of admission probability than the *isolated* one, so sharing gateways' lines is a good idea to obtain better values of admission probability. On the other hand we can see that increasing the CAC limit in *sharing mode* makes the admission probability increase, while values of *isolated mode* do not vary. Specifically, when there are 2 offices we can see that admission probability increases until CAC limit is 6, and later the value remains constant because the limitation is fixed by the gateways' lines. Furthermore, when the CAC limit increases in *sharing mode* there are more conferences that go through Internet, owing to the fact that the gateways' lines are shared.

### *B. R-factor*

Once we have studied the relationship between CAC limit and admission probability, in this section we will focus on the conferences' audio, to study its behavior in terms of R-factor. We are now interested on conferences that go through the IP network, because they are affected by RTP multiplexing, while the ones that are established inside the office are supposed to have high R values as the bandwidth is very big. We have used a scenario with four areas, with one country and one office in each area. There are 25 users in each office who generate conferences destined to a different gateway from their office. It is worth noting that in this test we have selected background traffic of 80% of the bandwidth limit in the office's routers. We consider this value big enough to test the system in a bad case. The variable parameters are the generation conferences' rate and the CAC limit in the offices.

Fig. 9 shows R-factor average for every conference of same execution with the *original* and the *multiplexing mode*. It can be seen that using *multiplexing mode* the R-factor values obtained are better than using the *original mode*. And it can also be seen that the variability of R-factor of different conferences is more delimited in *multiplexing mode*.

Fig. 10 compares *original* and *multiplexing mode* in terms of R-factor, depending on CAC limit. We can obtain several conclusions from this figure. First, it can be seen that there is an improvement by the use of RTP multiplexing. It is due to the fact that using *multiplexing mode,* the bandwidth of the traffic sent to the network is smaller than the one sent using the *original mode*. So the value of the CAC limit for *multiplexing mode* can be bigger than the one for *original mode*, in order to obtain the same value of R-factor.

On the other hand, focusing on a specific conference's rate value, the figure shows that if CAC limit grows, R-factor decreases. This result means that increasing the number of simultaneous conferences without increasing the bandwidth makes the conversation quality get worse.

As a concrete example, when conference's rate has a value of 10 calls per hour and CAC limit of 10, it can be seen that *multiplexing mode* obtains an R-factor value near 80, while with *original mode* it is below 70, which is an unacceptable value. And it can be seen than with *multiplexing mode* CAC limit could be increased to 25 while maintaining R-factor in acceptable values and increasing admission probability.

With these results and the ones obtained in the previous section, we realized that there is a tradeoff between admission probability and R-factor. So R-factor margin achieved by multiplexing could be sacrificed in order to obtain better values of admission probability. So we can use these results to find the CAC limit for which we will have an acceptable QoS for system conferences.

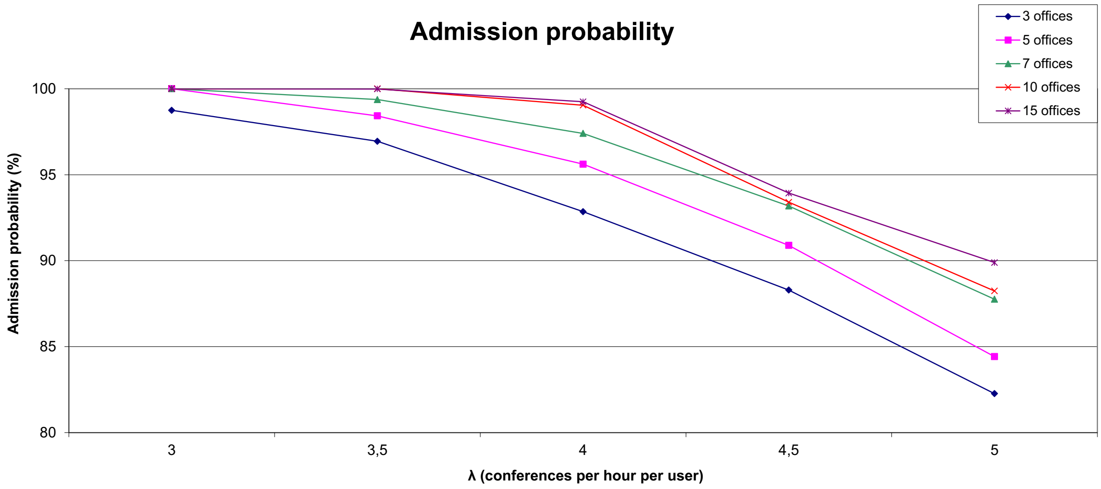

Figure 7. Admission probability with *sharing mode*

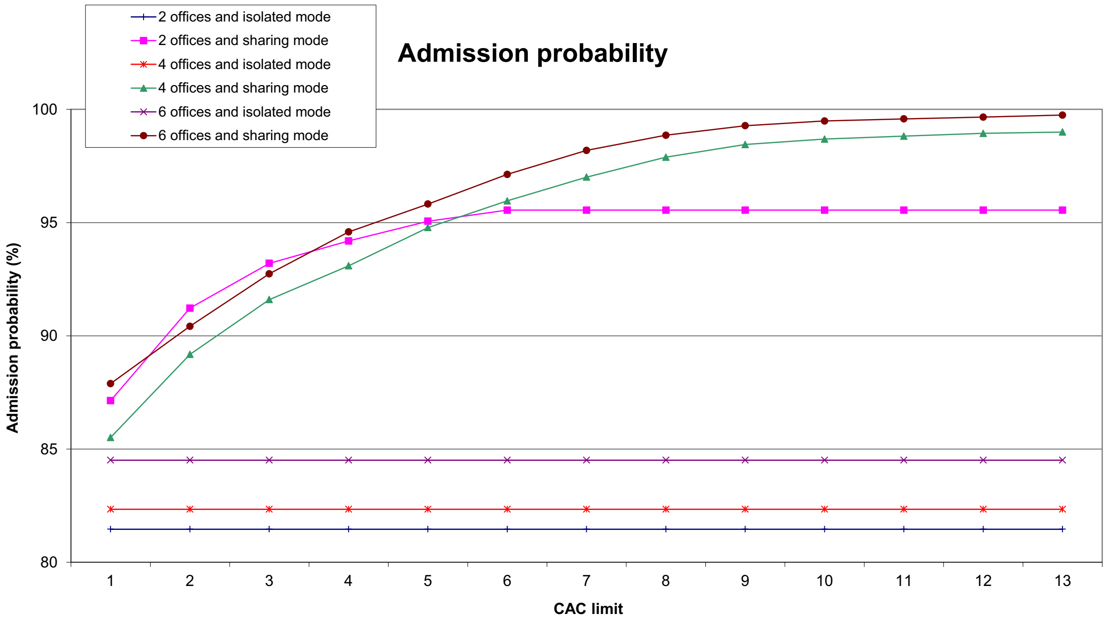

Figure 8. Admission probability with *isolated* and *sharing mode*

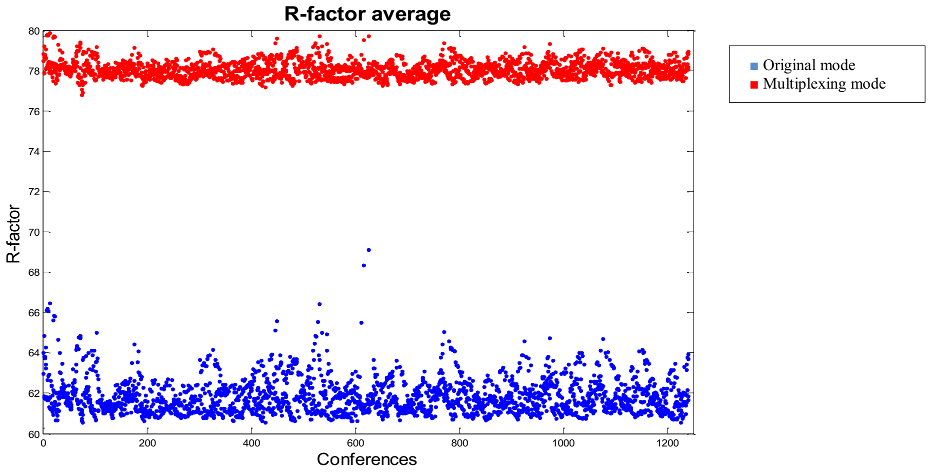

Figure 9. R-factor average for every conference of one execution

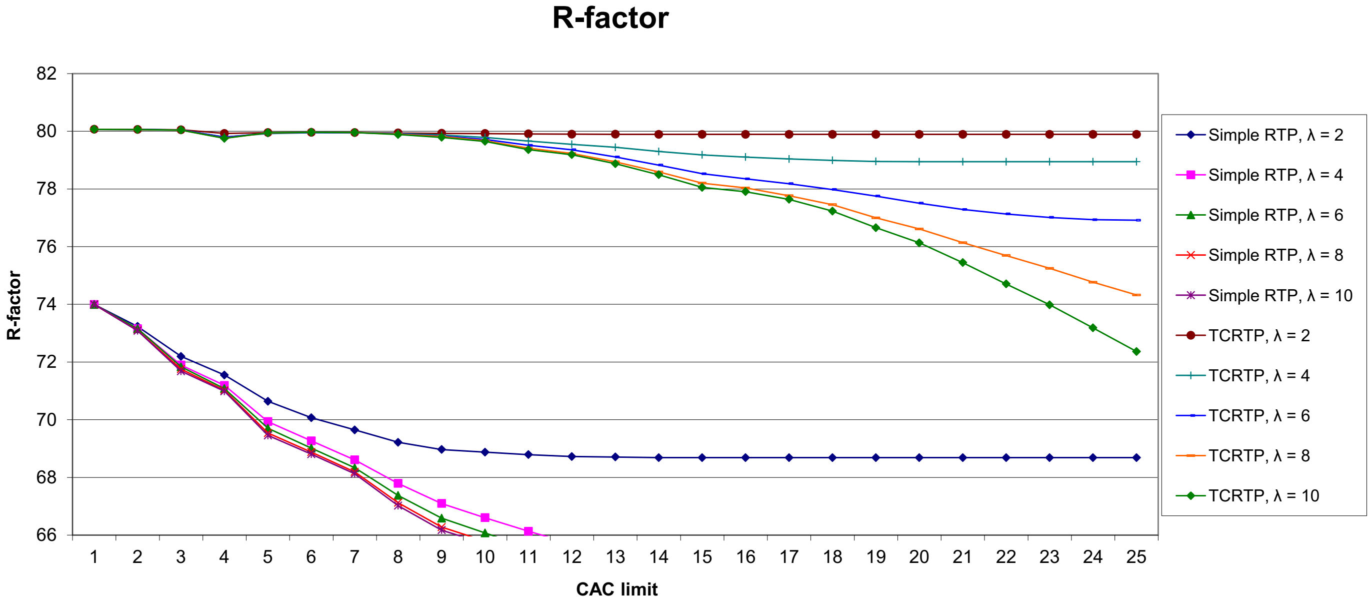

Figure 10. R-factor with *original* and *multiplexing mode* as a function of CAC limit. R-factor is considered unacceptable below R=70

## VI. CONCLUSIONS

In this work we have studied a telephony system based on VoIP and SIP protocol. The scenario matches with an enterprise with several offices in different areas. There is a CAC system to limit the number of conferences that can be established. Using a hybrid approach, QoS parameters which are necessary for the calculations have been obtained using a testbed based on emulation, while simulation has been used to generate executions based in several parameters.

On one hand, we wanted to test if sharing of gateways' lines we could improve the admission probability. On the other hand, we wanted to compare the system behavior when traffic is sent as simple RTP flows and when TCRTP multiplexing is used, in terms of R-factor.

We have simulated the system in different situations. The results obtained show that admission probability increases if gateways' lines are shared. And there are also improvements in terms of R-factor when TCRTP multiplexing is used, which can be translated to improvements in admission probability.

The developed simulations could also be used to plan a system before its deployment, as they help to estimate admission probability, delay, packet loss, R-factor, etc.

## REFERENCES

[1] S. Wang, Z. Mai, D. Xuan, and W. Zhao, "Design and implementation of QoS-provisioning system for voice over IP," Parallel and Distributed Systems, IEEE Transactions on, vol.17, no3, pp. 276--288, 2006

[2] ITU-T Recommendation G.107, “E-model, a computational model for use in transmission planning,” 2003

[3] B. Thompson, T. Koren, and D. Wing, “RFC 4170: Tunneling Multiplexed Compressed RTP (TCRTP),” 2005

[4] J. Saldana, J. Aznar, E. Viruete, J. Fernández-Navajas, and J. Ruiz-Mas, “QoS Measurement-Based CAC for an IP Telephony System,” QShine 2009, The Sixth International ICST Conference on Heterogeneous Networking for Quality, Reliability, Security and Robustness. Las Palmas de Gran Canaria (Spain), 2009

[5] S. Sacker, M. Santaiti, and C. Spence, “The Business Case for Enterprise VoIP,” Intel Corporation, 2006

[6] B. Athawal, “Replacing Centric Voice Services with Hosted VoIP Services: An Application of Real Options Approach”

[7] R. Solange, P. Carvalho, and V. Freitas, “Admission Control in Multiservice IP Networks: Architectural Issues and Trends,” IEEE Communications, vol.45, no. 4, pp. 114--121, 2007

[8] A. Dhamdhere, and C. Dovrolis, “Open issues in router buffer sizing,” Comput. Commun. Rev., vol. 36, no. 1 pp. 89--92, 2006

[9] M. Degermark, B. Nordgren, and D. Pink, “RFC 2507: IP Header Compression,” 1999

[10] S. Casner et al. RFC 2508, “Compressing IP/UDP/RTP Headers for Low-Speed Serial Links,” 1999

[11] T. Koren et al. RFC 3545, “Enhanced Compressed RTP (CRTP) for Links with High Delay, Packet Loss and Reordering,” 2003

[12] C. Bormann, Ed. RFC 3095, “Robust Header Compression (ROHC),” 2001

[13] VoIP Call Admission Control, http://www.cisco.com/en/US/docs/ios/solutions_doc/voip_solutions/ CAC.pdf

[14] G. Dimitriadis, S. Karapantazis, F.-N. Pavlidou, "Comparison of Header Compression Schemes over Satellite Links", In Proc. International Workshop on IP Networking over Next-generation Satellite Systems (INNSS'07), Budapest, Hungary, 2007.

[15] J. Saldana, E. Viruete, J. Fernández-Navajas, J. Ruiz-Mas, and J. Aznar, “Hybrid Testbed for Network Scenarios,” SIMUTools 2010, the Third International Conference on Simulation Tools and Techniques. Torremolinos (Spain), 2010

[16] ITU-T Rec. G.114, “One-way transmission time,” 1996

[17] J. Saldana, J. Murillo, J. Fernández-Navajas, J. Ruiz-Mas, E. Viruete, and J. Aznar, “Evaluation of Multiplexing and Buffer Policies Influence on VoIP Conversation Quality,” CCNC 2011 – 3rd IEEE International Workshop on Digital Entertainment, Networked Virtual Environments, and Creative Technology. Las Vegas, 2011