QoS and Admission Probability Study for a SIP-based Central Managed IP Telephony System

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

***Abstract*- This work presents a study of a SIP-based IP telephony system in terms of Quality of Service (QoS) and admission probability. The system is designed for an enterprise with offices in different countries. A software IP PBX maintains the dial plan, and SIP proxies are used in order to implement Call Admission Control (CAC) and to allow the sharing of the gateways' lines. The system is implemented in a testbed where QoS parameters like One Way Delay (OWD), packet loss, jitter and ITU's R-factor are measured. Simulation is also used in order to build a scenario with a big number of offices in which establishment delays and admission probability are measured.**

***Keywords*- VoIP, QoS, CAC, software PBX, SIP proxy**

I. INTRODUCTION

In the last years, VoIP is becoming an interesting option for telephony systems, and many enterprises are using it instead of Public Switched Telephone Network (PSTN). The deployment of software-based solutions in which a simple PC assumes the role of the old PBX has given a boost to this technological change. Small and Medium Enterprises (SME) which want to avoid the costs of proprietary systems have specially adopted General Public License (GPL) solutions.

VoIP is a real-time service that uses a network 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. Over provisioning is often used as a way to solve this problem, but the continuous growth of traffic and the spread of new services have led to the development of more intelligent solutions in both control and data planes [1]. Call Admission Control (CAC), which accepts or rejects new calls in order to avoid service degradation, is widely used.

An enterprise can avoid dedicated links by using IP telephony to connect different offices. Building a central managed telephony system brings the possibility of sharing lines between offices. This system can control all the telephony resources of the enterprise, achieving some advantages, e.g. cost savings in international calls, as they can be established from a gateway in the destination country.

Some proprietary systems use a control element in each location in order to manage calls [2]. In [3] our group presented a system that adds QoS to a software IP PBX, working with SIP. The offices are grouped in countries and geographical zones, depending on telephone tariffs (Fig. 1).

The use of SIP proxies allows the CAC to be seamlessly integrated into the telephony system, as neither the PBX nor the VoIP terminals need any modification. The inclusion of a new agent in each office can be seen as a drawback, but, as we will see, it is a very simple element that does not require a big processing capacity [4], so it could be easily included as a process in another existing server in the office. SIP allows the sharing or gateways' lines by the use of *redirect* messages, decreasing blocking probability.

This work presents a study of the system in terms of QoS parameters like One Way Delay (OWD), packet loss, jitter and ITU G.107 R-factor [5]. Establishment delay and admission probability are also measured. For the first studies a testbed platform has been used, and for establishment delay and admission probability we have used simulation in order to test a big number of offices. Some of the emulation results have been used as simulation parameters.

This paper is organized as follows: section 2 discusses the related works. System architecture is presented in section 3. The next section covers the system implementation. Section 5 presents the tests that have been carried out, and the results. The last section details the conclusions.

II. RELATED WORKS

As said in the introduction, VoIP is a real-time service that uses best effort networks, so QoS can be added by means of different methods, such as CAC. Two of the most used protocols for VoIP signaling are H.323 and SIP (Session Initiation Protocol). Many proprietary solutions typically use H.323, but SIP is becoming popular because it is an open protocol and some open-source PBX solutions and commercial VoIP systems use it [6]. SIP proxies are elements used to add scalability, as they transfer workload from the network core to the borders.

CAC systems have been classified [7] in two main categories: measurement based, which use the state of the network to take admission decisions, and parameter based which need some measurements during the system's set up, to obtain the parameters that will manage system behavior, as the maximum number of simultaneous calls. In [8] a parameter based CAC system for VoIP was presented. It was mainly centered in adding QoS to Cisco environments, so H.323 was the initial protocol, and an *Integration Component* was included in order to achieve signaling interception. In our

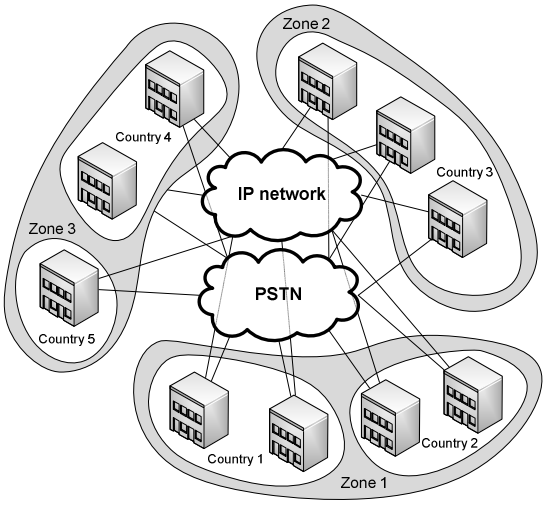

Figure 1. IP telephony scheme.

system this component is not necessary, as we use SIP proxies to intercept the call's signaling in a natural way.

One of the most popular software PBX solutions is Asterisk [9], developed by Digium. It acts as a *back to back user agent*, binding two calls: one from the origin to the PBX, and another from the PBX to the destination. It supports SIP and many other signaling protocols.

As the central PBX has no information of traffic in each office, the addition of a local SIP proxy able to control the calls which are generated at the office, allows the CAC to be seamlessly integrated with the PBX and the phones, as they do not need any modification. Cisco's solutions use an element called *CallManager* that is placed in each location, and interoperates with the *GateKeeper*, which is in the central node [2]. The system evaluated in this work is similar to proprietary schemes, but different parameters can be modified and studied for getting comparatives: codec, CAC schemes, router's buffers etc.

The influence of router buffer policies in the system has to be studied too. In the recent years the traditional "rule of the thumb" of the bandwidth-delay product [10] used to calculate the buffer size has been questioned by the "Stanford model" [11]. Ref. [12] presents a comparative and suggests the use of time-limited buffers, which are good to maintain network delays below an upper bound, but can increase packet loss. We will make some buffer comparisons in this work.

We can obtain OWD and packet loss from network measurements. ITU G.107 proposes R-factor as an estimator of conversation quality. It is based on the different impairments that affect the signal when it travels from mouth to ear. In [13] a simplified expression of delay impairment I_d was obtained, and the conclusion is that R-factor has two roughly linear regions. If delays are bigger than 177 ms, the impairment grows more quickly. R-factor rates calls from 0 to 100. Acceptable values are considered for R > 70. It can be translated to Mean Opinion Score (MOS), which ranges from 0 (bad) to 5 (good) using a simple expression [13].

III. System Architecture

We will now summarize the architecture of the telephony system we are studying, which is designed to work in an enterprise with offices in different countries, which are grouped in zones. The dial plan is only kept at the PBX, thus avoiding management expenses. Internet is used for telephone traffic, saving the costs derived from dedicated lines. We have made two assumptions: the system does not use any reservation protocol, and VoIP is the only real-time traffic we are going to take care of in a special way.

A parameter based CAC is used. It is a simple scheme which makes some measurements during configuration time, and then it assigns a maximum number of calls to each office. The main objective is to assure a minimum QoS for VoIP calls, at the cost of rejecting some of them.

As all signaling messages pass through the local agent, it is able to take decisions about connection requests and it can also keep count of the number of established calls in the gateway. In case no call rejections are decided, the agent only retransmits signaling messages. Internal office calls are directly managed by the local agent and do not require PBX functions to be established, and they are not affected by CAC. For rejected calls (Fig. 2), a *480 Temporarily Unavailable* SIP message is sent to the PBX.

We have used two different algorithms in order to make a comparative. First, there is an isolated mode in which nothing is shared between offices. Each one has its gateway and there are no redirected calls. The other algorithm, noted as sharing mode, is the central managed mode, in which the gateways are shared. Thus, calls can be established from a gateway in the destination country, avoiding the cost of an international call. There is another advantage: the blocking probability is reduced, as more lines are shared. We will see the results in next sections.

The system allows the establishment of different call types. We classify them in 6 types (Fig 3):

- 1: Internal call in the same office.
- 2: Internal call between different offices.
- 3: PSTN call to a country which has no office. It is managed by the local gateway.
- 4: PSTN call established by means of the gateway of the office.
- 5: PSTN call established by means of the gateway of another office.
- 6: PSTN call from an external user. It arrives to the gateway, and its destination is a phone of the office.

The calls that go to PSTN via a gateway can also be redirected if there are no available lines. In this case the agent acts as a SIP redirect server, re-routing the call to another central office (if possible, in the same country) which may have available lines to establish the connection. The redirect server sends a 3XX message reporting about an alternative route. The agent who acts as redirect server will not take part in this call again. This behavior can save international call's

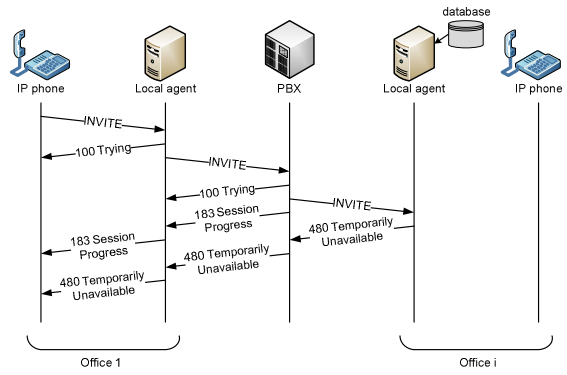

Figure 2. Call rejected by the system.

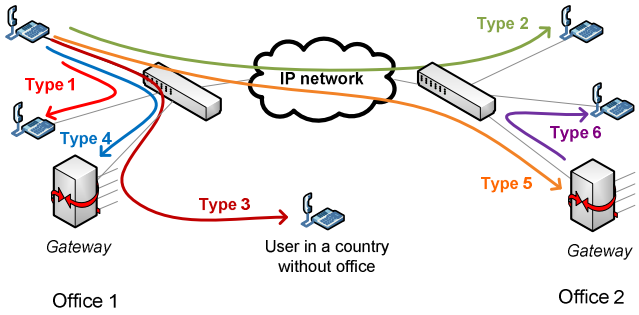

Figure 3. Call types scheme.

costs, as they can be established in two steps: one via Internet to an office in the destination country, and another one to the destination, using the PSTN connections of that country.

IV. System Implementation

Once a system has been designed, a battery of tests has to be deployed in order to measure different parameters, assuring that they are in acceptable ranges. Different tools can be used for this. First, we have implemented the system with real software in a testbed based on virtualization which was presented in [14]. Each computer of the system is translated into a virtual machine

A scenario with four central offices has been built at the testbed. We have implemented offices' routers buffers with Linux *tc* (Traffic Control), using a token bucket queue which limits the bandwidth at level 2.

The used codec is G.729a with two samples per packet. This implies a packetization delay of 25 ms, including 5 ms of look-ahead delay. Each RTP flow has a bandwidth of 24 kbps at IP level, and it is directly sent from the origin application to the destination.

Regarding the applications, some off-the-shelf solutions have been used. First, OpenSIPS 1.4 [15], a project derived from OpenSER, has been selected as SIP proxy. It provides *register server, location server, proxy server* and *redirect server* functionalities. Low computational load and the possibility to add and delete functionalities in a modular way are also interesting features. For the PBX we have used Asterisk 1.6.0.1. It represents an interesting solution because of its flexibility, updates and GNU-GPL license distribution. Finally, we have selected a command line soft phone called PJSUA 1.0. for the tests. It has also been used to emulate the gateways, as we have used no real PSTN connections.

But the testbed imposes some size limitations due to the number of machines it can support. If we want to study admission probability, we need a bigger scenario. An analytical approach would be very complicated, as we have different traffic sources, and many redirections, which have no Poisson distribution [16]. So another useful approach is the use of simulation, which can be complementary to the testbed. A Matlab application has been implemented to simulate the system with different variable parameters: number of offices, users per office, gateways' lines, call arrival rates, areas, countries, network and processing delays, etc.

Some of the parameters, as e.g. network delays, are added after obtaining them from the emulation testbed. The application does not simulate individual packets, but it is able to calculate the QoS parameters of each part of every conference. A part of a conference is defined as the time in which it shares the access network with the same number of simultaneous conferences.

V. Tests and Results

A. QoS Measurements

The first parameter we have considered is the quality of the conversation, which will be measured using R-factor. It depends on OWD and packet loss for VoIP packets.

The distribution of background traffic at IP level is: 50% of the packets are of 40 bytes, 10% of 576 bytes and 40% of 1500 bytes [17]. This traffic shares the access link with the RTP traffic. UDP has been used in order to avoid the flow control of TCP. The traffic sent is the same during the entire test, so we are in the worst case. Each test lasts 400 seconds, in which RTP and background traffic share the same link.

Network delays are added once the RTP packets have arrived to the destination router. The statistical distribution is the one proposed in [18]: first, a fixed delay is added depending on the distance between the two nodes. Another log-normal distributed delay is added, with average delay of 20 ms with a variance of 5. These values have been taken from [19], as typical internet delays.

As we wanted to obtain general results, we have not used the soft phone to calculate the QoS parameters of RTP traffic. Each application has a concrete implementation of de-jitter buffer, which adds a delay and also considers as packets lost the ones that arrive too late. We have used an approximation suggested in [13] to estimate packet loss:

$$loss_{\text{de-jitter buffer}} \sim \mathrm{P}\{\ l > bg\} \qquad (1)$$

Where l is the difference in OWD of consecutive packets, b is the delay of the buffer size, and g is inter-packet generation

time (20 ms in our case). A fixed value of *b*=2 has been used. By the use of adaptive schemes, we could obtain better results.

As we have previously said, we have included in the offices' routers two different disciplines: first, a high-capacity buffer which has a big queue, and second, a buffer with a queuing policy that discards packets that spend more than 60 ms in it. The bandwidth at the uplink is 1 Mbps.

The difference between the high-capacity and the time-limited buffers is that in the first case, the quality of the calls falls very quickly when the bandwidth limit is reached (Fig. 4), whereas the time-limited buffer makes R-factor go down slowly (Fig. 5). We see that acceptable R values are obtained with more background traffic.

On one hand, the time-limited buffer limits the maximum delay of voice packets, but on the other hand it increases packet losses for background traffic. Fig. 6 (a) shows the packet loss with 800 kbps of background traffic. It can be seen that when the traffic is close to the bandwidth limit (1 Mbps), packets start to be discarded, and big ones are discarded in a very big percentage. This occurs when the number of RTP flows is above 8, as we have more than 200 kbps of RTP traffic. Fig. 6 (b) shows that in this situation small packets maintain their bandwidth, whereas big packets are dropped in a bigger percentage, as they frequently do not have place in the queue.

This result is relevant as it shows that RTP traffic is protected because of its small size. So limiting the maximum number of calls is also important to avoid the degradation of the rest of the traffic of the office. This means that CAC can be useful not only to protect VoIP traffic, but also for the objective of avoiding the degradation of background traffic.

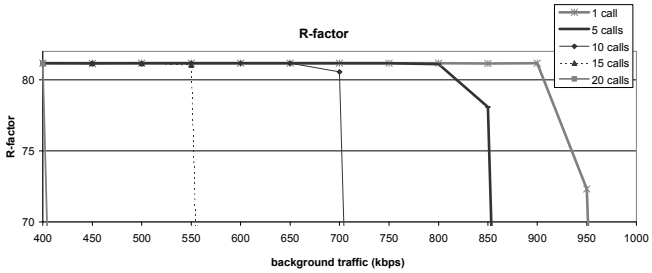

Figure 4. R-factor with high-capacity buffer.

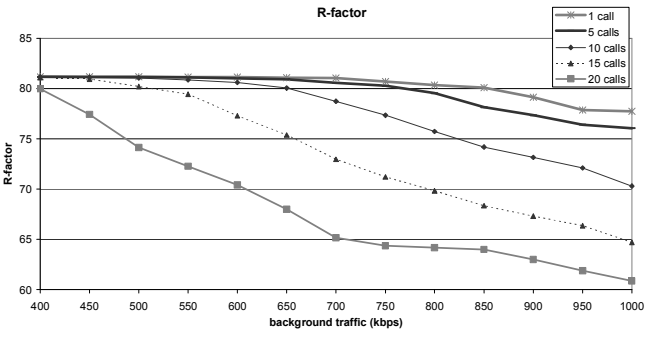

Figure 5. R-factor with time-limited buffer.

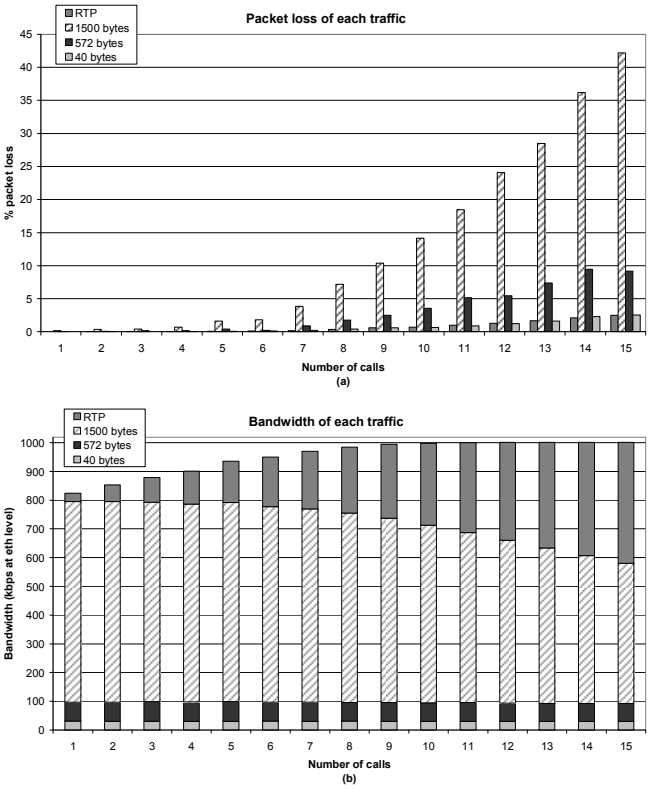

Figure 6. System behavior: 800 kbps background traffic, time-limited buffer.

Fig. 7 (a) and 7 (b) show the performance of the system in terms of OWD and packet loss, when using a time-limited buffer. They are the upper bound for these values in case the maximum number of calls of the CAC is set to each value. The values are represented as a function of background traffic, so each graph reaches congestion in a different moment, when the behavior gets worse.

In the graphs it can be seen that before congestion, OWD has more influence than packet loss in R-factor. Once congestion is reached, both parameters have the same influence. But the queue size limit makes OWD have an upper bound below 175 ms, whereas packet loss grows indefinitely. This result suggests the possibility of adding to the system a buffer that changes its size depending on the number of calls.

The system Inter Packet Delay Variation (IPDV) in Fig. 7 (c) presents a maximum and starts to decrease as the access is near to saturation, because big packets start to be discarded in a high percentage, and they are the main cause of jitter.

B. Establishment Delay

As we have seen, our system has the possibility of redirecting calls. This is good for admission probability and cost savings, but the impact on establishment delay has to be measured. Fig. 8 illustrates the different delays which have to be considered. We will measure the time since the INVITE message is sent by the origin until it arrives to the destination phone. So we can summarize the components of establishment

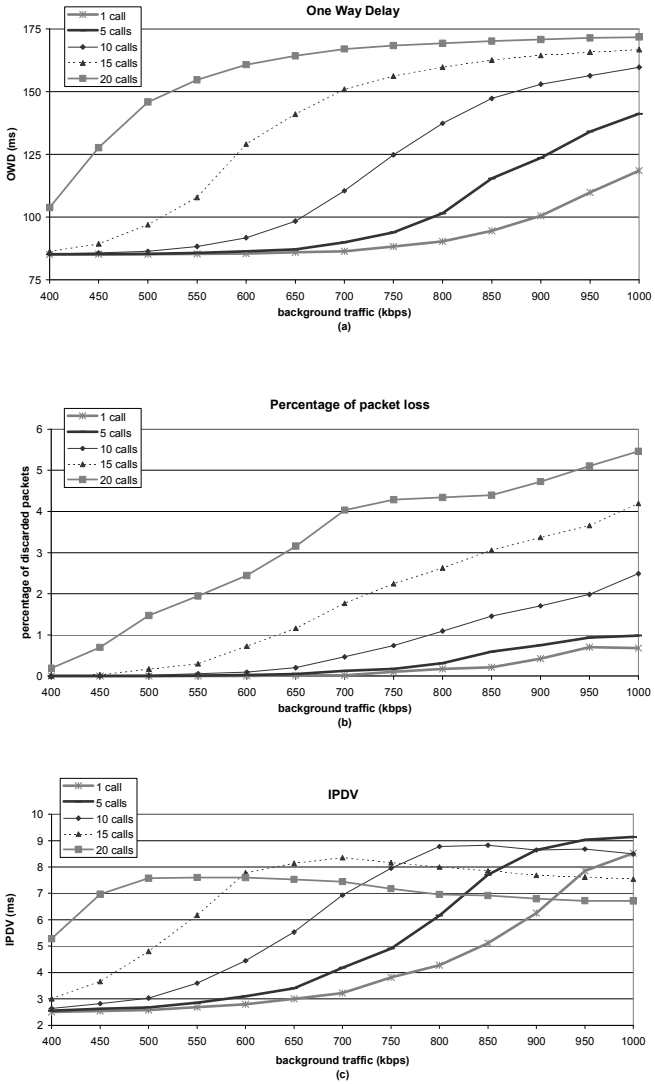

Figure 7. OWD, packet loss and IPDV with time-limited buffer.

delays, as can be seen in Table I. The delays we have introduced in the simulations are:

- Network delay at the origin and destination LANs: We will consider it negligible, as local networks are supposed to be fast.

- Processing time at the SIP proxy and the PBX: They were measured in [3], so we will use 5 ms for the proxy and 3 ms for the PBX.

- Queuing delay at the origin router: It has been estimated to be similar to the queuing delay experienced by RTP packets, as they share the same buffer. So we have used the values obtained from the testbed. We have not considered this delay at the destination router, as the packet goes from a slow network to a faster one. As the PBX is supposed to be at the data center, which has a broadband connection to the Internet, this delay is not considered for it.

- Network delay at the WAN: We have applied the same method used to calculate network delays of RTP packets.

Fig. 9 shows the establishment delay as a function of the RTT of the network, and with different number of offices. Network RTT ranges from 25 to 125 ms, which are typical Internet values [19]. It can be seen that the delay does not change with the number of offices.

Call arrivals follow a Poisson distribution with different values. Call duration has been modeled with a Normal distribution of 180 sec. average. The scenario includes 25 users per office. Every office has a gateway with 6 lines, and CAC limit is 6 calls. λ represents the call arrival rate, in number of conferences per user per hour. We have used $\lambda = 4$.

C. Admission Probability

In this subsection we present some simulation results in order to show the benefits of sharing gateways' lines between different offices.

Fig. 10 represents admission probability as a function of λ, with different number of offices. Each office has 25 users with the same call arrival rate; the gateways have 6 lines, and CAC limit is set to 6 calls. It can be seen that the bigger the number

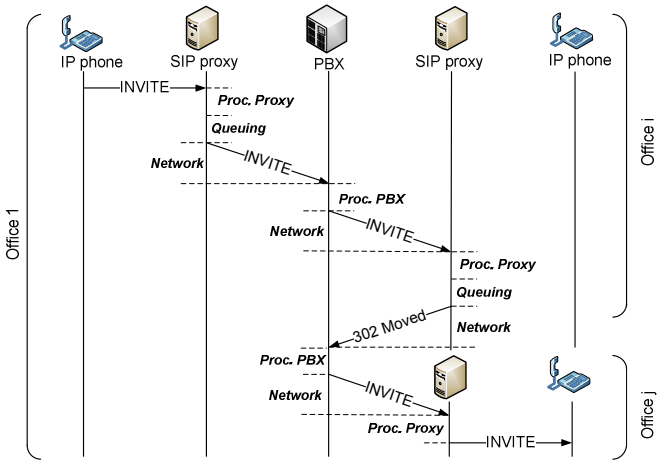

Figure 8. Delays that affect the establishment delay.

TABLE I

COMPONENTS OF ESTABLISHMENT DELAY

	Internal Call	External Call	Each redirection
Network Delay	0	2	2
Processing Proxy	1	2	1
Processing PBX	0	1	1
Queuing Delay	0	1	1

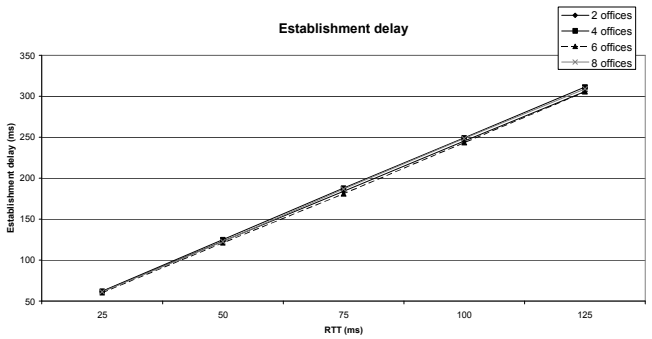

Figure 9. Establishment delay as a function of network RTT.

of offices, the bigger the admission probability, so sharing all the gateways' lines to carry the sum of the traffic is beneficial. On the other hand, when λ is very big, the admission probability decreases.

The influence of CAC limitation has also been tested. We have seen that sharing the gateways we can improve admission probability. But if too many conferences are sharing the same access network, QoS can get unacceptable. In Fig. 11 we have represented admission probability as a function of CAC limit, using 6 lines on each gateway. λ is set to 3 calls per user per hour, and there are 25 users on each office. It can be seen that in *isolated mode*, the values are always the same, as calls are never redirected. In *sharing mode* the admission probability gets better as the CAC limit is bigger. We see again that the system behaves better with a big number of offices, as more resources are shared.

VI. Conclusions

A SIP-based IP telephony system designed for an enterprise with offices in different countries has been tested. A software PBX maintains the dial plan, and SIP proxies are used in order to implement CAC and to allow the sharing of the gateways' lines using SIP *redirect* messages. The system uses a SIP proxy to accept or reject calls depending on parameters established at configuration time.

The tests of the system have been carried out in a virtualization based platform. Each computer of the system is translated into a virtual machine. Network delays have been added, and a queuing discipline has been integrated in each office router. Simulations have also been used in order to test the system with a big number of offices.

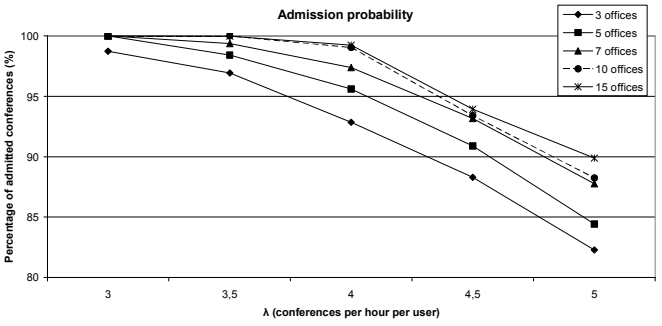

Figure 10. Admission probability as a function of λ with *sharing mode.*

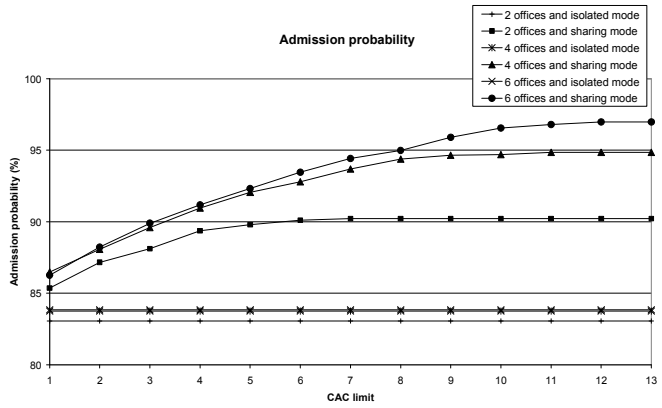

Figure 11. Admission probability as a function of CAC limit.

QoS measurements have been done with different number of calls. R-factor has also been calculated. Establishment delays have been measured as well. The results show that the system does not introduce delays that could impair the quality experienced by users. A study of admission probability has been carried out, in order to show that the sharing of the gateways' lines between different offices achieves an improvement of admission probability.

Acknowledgment

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 the Catedra Telefonica of the Univ. of Zaragoza.

References

[1] X. Chen, C. Wang, D. Xuan, Z. Li, Y. Min, W. Zhao, "Survey on QoS Management of VoIP", In *Proc. of the 2003 International Conference on Computer Networks and Mobile Computing*, IEEE Computer Society.

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

[3] J. Saldaña, 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), Nov. 2009.

[4] S. Wanke, M. Scharf, S. Kiesel and S. Wahl, "Measurement of the SIP Parsing Performance in the SIP Express Router", Proc. *13th Open Eur. Summer School and IFIP TC6.6Workshop*, Enschede, Neth., 2007.

[5] "The E-model, a computational model for use in transmission planning", ITU-T Recommendation G.107. Mar. 2003.

[6] SIP: Measurement-Based Call Admission Control for SIP, http://www.cisco.com/en/US/docs/ios/12_2t/12_2t15/feature/guide/ftcacsip.pdf.

[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, Apr. 2007.

[8] S. Wang, Z. Mai, D. Xuan, W. Zhao, "Design and implementation of QoS-provisioning system for voice over IP", *Parallel and Distributed Systems, IEEE Transactions on*, vol.17, no.3, pp. 276-288 (2006).

[9] Asterisk, The Open Source Telephony Projects, www.asterisk.org

[10] C. Villamizar, C. Song. "High performance TCP in ANSNET". *ACM Computer Communication Review*, Oct. 1994.

[11] G. Appenzeller, I. Keslassy, and N. McKeown. "Sizing router buffers", In *SIGCOMM '04*, pages 281–292, New York, USA, 2004. ACM Press.

[12] A. Dhamdhere and C. Dovrolis, "Open issues in router buffer sizing", *Comput. Commun*. Rev., vol. 36, no. 1, pp. 87-92, Jan. 2006.

[13] R.G. Cole, J.H. Rosenbluth. "Voice over IP performance monitoring". *SIGCOMM Comput. Commun*. Rev. 31, 2 (Apr. 2001), pp. 9-24.

[14] J. Saldaña, E. Viruete, J. Fernández-Navajas, J. Ruiz-Mas, J. I. Aznar, "Hybrid Testbed for Network Scenarios", in *Proc. SIMUTools 2010, the Third International Conference on Simulation Tools and Techniques*. Torremolinos, Malaga, Spain (2010).

[15] OpenSIPS, Open SIP Server, www.opensips.org

[16] T. Venkatesh and C. Siva Ram Murthy, "*An Analytical Approach to Optical Burst Switched Networks"*, Springer, New York, 2010.

[17] Cooperative Association for Internet Data Analysis: NASA Ames Internet Exchange Packet Length Distributions.

[18] S. Kaune, K. Pussep, C. Leng, A. Kovacevic, G. Tyson, R. Steinmetz. "Modelling the internet delay space based on geographical locations". In *17th Euromicro International Conference on Parallel, Distributed, and Network-Based Processing (PDP 2009),* Feb 2009.

[19] AT&T Global IP Network, http://ipnetwork.bgtmo.ip.att.net/pws