Measurements and Analysis of M/M/1and M/M/c Queuing Delay Models of the Two IP-PBXs in Various Remote Location

Pedram Hajipour*, Nahid Amani*, Alireza Dehestani**, Mojtaba Mazoochi*
** Iran Telecommunication Research Center, North Kargar Street , Tehran, Iran*
***Faculty member of Islamic Azad University, Qom Branch Qom Iran*
hajipour@itrc.ac.ir, n_amani@itrc.ac.ir, ar.dehestani@ie-group.ir, mazoochi @itrc.ac.ir

***Abstract*-Voice over Internet Protocol (VoIP) is a general term for a family of transmission technologies to delivery of voice communications over IP networks such as the Internet or other packet-switched networks. VoIP systems employ session control protocols to control the set-up and tear-down of calls as well as audio codecs which encode speech allowing transmission over an IP network as digital audio via an audio stream. In this paper, Codec in bandwidth management was investigated. Codec is used between different implementations of VoIP variedly; some implementations rely on narrowband and compressed speech, while others support high fidelity stereo codecs.The Session Initiation Protocol(SIP) is becoming a popular signaling protocol for VoIP base applications. The IP-PBX is a software application which provides call routing services by parsing and forwarding all the incoming SIP packets in various remote IP telephony network. Then, the M/M/1 and M/M/c performance model of the IP-PBX was simulated and some of the key performance benchmarks such as average response time to process the SIP calls and mean number of jobs in the system was reviewed .**

Keywords: SIP, IP-PBX, Queuing delay Analysis, Performance Analysis

I. INTRODUCTION

Since the telephone was invented in the late 1800s, telephone communication has not changed substantially. Of course, new technologies like digital circuits, DTMF (or, "touch tone"), and caller ID have improved on this invention, but the basic functionality is still the same. Generally, users do not know that how offered services work, but they know two issues: 1- the same old telephone is used and 2- the service providers charge users for each service. In the 1990s, a number of Researchers in research centers and corporate institutions took a serious interest in carrying voice and video over IP networks, especially corporate intranets and the Internet. This technology is commonly referred to today as VoIP and is, in simple terms, the process of breaking up audio or video into small chunks, transmitting those chunks over an IP network, and reassembling those chunks at the far end so that two people can communicate using audio and video. In recent years, SIP, an Internet Engineering Task Force(IETF) standard, has been considered a promising signaling protocol for the current and future IP telephony services due to its simplicity, flexibility and built in security features [1].There are several ongoing discussions on the Quality of Service (QoS) in SIP base Network within the IETF and other research communities. Quality of Service (QoS) is the ability of a network to provide improved service to selected network traffic. QoS provides control over congestion management, queue management, traffic shaping and policing, and link efficiency. This makes it easier for mission-critical applications to co-exist on a network. key contribution in this paper is obtaining a better understanding of the performance of the IP-PBX system when there is a call flow between two different domains. In this paper, the M/M/1 and M/M/c queuing models without propagation delay proposed in [2] was simulated and was focused on network delay in analytical result. Two IP-PBXs were considered which have proxy server role to process all the incoming SIP calls generated from SIP end points.

II. Related Work

Wu et al. [3] analyze the queuing delay and queuing delay variation using embedded Markov chains in a M/G/1 queuing model. Our work, by contrast, analyzes performance under varying service rates and network delays of an end-to-end native SIP network.

Lipson [4] presents an approach for using model checking of Markov Reward Models to analyze properties of a simple SIP network. The focus is on transient properties related to the number of calls processed prior to system failure or system repair. Rewards are expressed as simple rates of incoming requests for call setups. Our model, in contrast, is a hierarchical model consisting of a high-level Markov Reward Model and a lower-level queuing network model.

V.K.Gurbani, L. Jagadeesan, V.B. Mendiritta, [2] came up with an analytical SIP based performance and reliability model, in which they primarily considered the mean response time and the mean number of calls in the system. They modeled a SIP proxy server as an open feed forward queuing network. analyze the queuing delay and queuing delay variation using embedded Markov chains in a M/M/1 queuing model for Performance and

Reliability in SIP network. Suresh KumarV. Subramanian, Rudra Dutta[1], analyze the queuing delay and queuing delay variation using embedded Markov chains in a M/M /1 queuing model and M/M/c queuing model of the SIP Proxy Server. Raja opal et al.[5] analyzed and proposed the IP Multimedia Services(IMS) network based on the SIP signaling delay predicted performance trends of the network, which allowed them to choose parameter values optimally, the proposed models were based on queuing models for the IMS network that characterizes the sip server workload.

III. SIP Protocol

A. *History*

The Session Initiation Protocol (SIP) is a signaling protocol, widely used for controlling multimedia communication sessions such as voice and video calls over Internet Protocol (IP) [6,7]. Other feasible application examples include video conferencing, streaming multimedia distribution, instant messaging, presence information and online games. The protocol can be used for creating, modifying and terminating two-party (unicast) or multiparty (multicast) sessions consisting of one or several media streams. The modification can involve changing addresses or ports, inviting more participants, adding or deleting media streams and etc. SIP was originally designed by Henning Schulzrinne and Mark Handley starting in 1996. The latest version of the specification is RFC 3261 from the IETF Network Working Group. The SIP protocol is a TCP/IP-based Application Layer protocol. As shown in Figure (1), A SIP User Agent (UA) is a logical network end-point used to create or receive SIP messages and thereby manage a SIP session. A SIP UA can perform the role of a User Agent Client (UAC), which sends SIP requests, and the User Agent Server (UAS), which receives the requests and returns a SIP response [8].

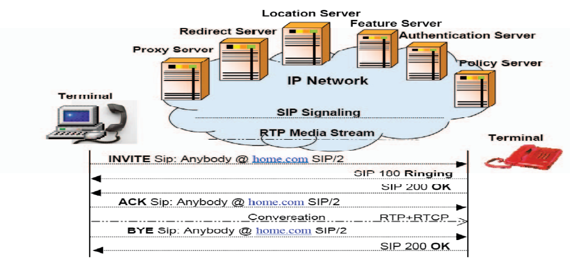

Fig. 1.The SIP Architecture

B. *SIP Messages*

SIP is a text-based protocol with syntax similar to that of HTTP. There are two different types of SIP messages: requests and responses. The first line of a request has a method, defining the nature of the request, and a Request-URI, indicating where the request should be sent[10]. The first line of a response has a response code. For SIP requests, RFC 3261 defines the following methods:[11]

- Provisional (1xx): Request received and being processed.
- Success (2xx): The action was successfully received, understood, and accepted.
- Redirection (3xx): Further action needs to be taken (typically by sender) to complete the request.
- Client Error (4xx): The request contains bad syntax or cannot be fulfilled at the server.

IV. Introduction IP- PBX

An IP PBX or VoIP phone system replace a traditional PBX or phone system and give employees an extension number, the ability to conference, transfer and dial other colleagues. All calls are sent via data packets over a data network instead of the traditional phone network. With the use of a VOIP gateway, you can connect existing phone lines to the IP-PBX and make and receive phone calls via a regular PSTN line. The IP PBX FAQ helps answer common questions about VOIP, SIP, IP PBX / VOIP Phone System hardware & Software, implementation and etc[13].

V. Call flow using a SIP peer/user Trunk pairing

As shown in figure (2), there are a IP-PBX to IP-PBX connection which will be use a user/peer pairing to form a SIP trunk. In this scenario, IP-PBX supports any IP phones and is an integral part of a SIP network. It can play different roles, such as registrar server and B2BUA.The two IP-PBXs are name Rayaphone and ITRC after their IP host address, this will be handy when making outbound routes. For any remote extensions, edit the extension any change the setting for NAT to yes is needed. This tells IP-PBX that the extension is not on the same network as the server. As shown in table (1), local net tells IP-PBX what IP range the server sits on.

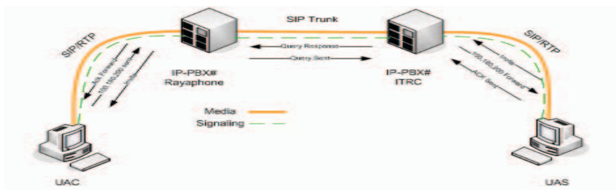

Fig. 2. SIP call in different domain

A. *Capturing VoIP conversation with Wire shark*

The VoIP druid has a quick and dirty web page on capturing VoIP conservations with wire shark.
The best part is that you can actually view the quality and playback the RTP conservations on the capture messages. Wire shark(aka ethereal) is a tool to have in his toolbox when ever working with application using the network, it is simple, efficient and running either on Microsoft windows or linux.wireshark has good plugging targeting the VoIP space(aside many others)[14].this paper focusing on SIP protocol which represents most of nowadays VoIP implementation.

TABLE I
SIP Configuration in IP-PBX

PBX# Rayaphone					PBX # ITRC	
IP address=	170.170.170.80				IP address=	192.168.1.5
Trunk Name =	Rayaphone		*Names are based*		Trunk Name =	ITRC
Username =	UAC		*on the destination*		Username =	UAS
Password =	XXX				Password =	XXX
Outgoing Settings					Outgoing Settings	
Trunk Name:	Rayaphone				Trunk Name:	ITRC
Peer Details:					Peer Details:	
host	192.168.1.5				host	170.170.170.80
username	UAS				username	UAC
from user	UAS				from user	UAC
secret	XXX				secret	XXX
type	peer				type	peer
qualify	yes		Create SIP trunk		qualify	yes
nat	yes		configurations between		nat	yes
Externip	192.168.1.5		PBXs		Externip	170.170.170.80
Localnet	170.170.170.80		using User/Peer pairing		Localnet	192.168.1.5
Incoming Settings					Incoming Settings	
User Context:	UAC				User Context:	UAS
User Details:					User Details:	
secret	XXX				secret	XXX
type	user				type	user
context	from-trunk				context	from-trunk

VI. Bandwidth management in IP- PBX

A. Codec (coder/decoder)

When making a call over SIP, the software or hardware needs to use a codec so as spend/receive information in a certain format and convert it to what you hear. What this means to you and me is that codecs compress data, allow you to transmit the compressed data which is then uncompressed at the receiving end. Different codecs have different compression ratio resulting in different Bandwidth requirements. Generally the higher the compression the more CPU power required to compress and sometimes decompress [13] .

In most cases this means:

Higher compression = More CPU power = Lower Bandwidth

Lower compression = less CPU power = More Bandwidth

As we seen in table (2),there are the approximate Bandwidth requirements for several common VoIP compression algorithms using SIP. They vary slightly depending on the protocol used. We assumed the BW requirements for IP-PBX is equivalent to BWL and BW required for a call(RSVP) per hour is equivalent to BWS and average time duration of a call is equivalent to t_{ave}. in “equation(1),”t here fore we can obtain rate of reservation for each codec. We compare between different codec in figure (3).

TABLE II
codec in VoIP [13,15, 16,17]

Codec	BWL(kbps)	BWS(kbps) per hour
G.711	87.2	39.24
G.729	31.2	14.04
G.723.1	21.9	9.36
G.726	55.2	24.84
G.728	31.5	14.18
GSM	28.7	12.92

$$\text{Rate of Reservation} = \frac{BWL}{(BWS * T_{ave})} \qquad (1)$$

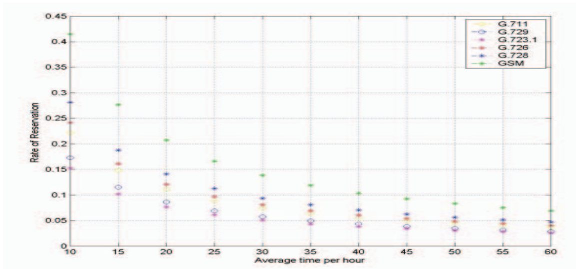

Fig. 3. Codec comparisons in bandwidth management

VII. Call Flow Scenario base IP-PBX

“Fig. (4)” illustrates stages of the proposed scenario for call setup between two different domains by SIP network. The scenario is composed of the following stages:

1.A SIP phone(UAC)/ (170.170.170.47) sends an INVITE request containing standard SDP information to IPPBX# rayaphone(170.170.170.8 0)

2.IP-PBX#rayaphone responses with a 100 Trying message

3.IP-PBX#rayaphone forwards an INVITE request message over SIP trunk for IP-PBX#ITRC (192.168.1.5)

4. IP-PBX#ITRC answers with a 100 Trying message

5. IP-PBX#ITRC sends an INVITE request message to A SIP phone(UAS) / (192.168.1.3)

6. A SIP phone (UAS) responses with a 100 Trying and 180 Ringing message

7. IP-PBX#ITRC forwards a 180 Ringing message to IP-PBX#rayaphone

8. IP-PBX#rayaphone forwards 180 Ringing message to A SIP phone (UAC)

9. A SIP phone (UAS) forwards a 200 OK message to IP-PBX#ITRC

10. IP-PBX#ITRC forwards a 200 OK message to IP-PBX#rayaphone

11. IP-PBX#rayaphone forwards a 200 OK message to A SIP phone(UAC)

12. A SIP phone(UAC) responses with ACK message SIP message to IP-PBX#rayaphone

13. IP-PBX#rayaphone responses with ACK SIP message to IP-PBX#ITRC
14. A SIP phone(UAS) sends a BYE SIP message to IP-PBX#ITRC
15. IP-PBX#ITRC forwards a BYE SIP message to IP-PBX#rayaphone
16. IP-PBX#rayaphone forwards a BYE SIP message to A SIP phone(UAC)
17. A SIP phone(UAC) responses a 200 OK message to IP-PBX#rayaphone
18. IP-PBX#rayaphone forwards 200 OK message to IP-PBX#ITRC
19. IP-PBX#ITRC forwards 200 OK message to A SIP phone(UAS)

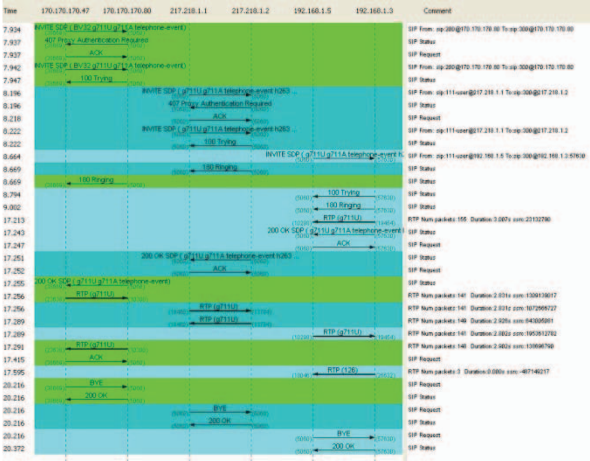

Fig. 4. SIP actual message transaction in various remote location diagrams in wire shark

A. M/M/1 and M/M/c based IP-PBX Model

To assess the usefulness of the queuing model proposed by Gurbani et al[2], we simulated this model(the authors of [2] only provided the analytical model in a network and did not provide a comparison between two different domains while there is varity IP address belong to end user. In this work, the IP-PBX is modeled as open feed-forward queuing network, in which incoming INVITE message from UAC as arrival calls and there are sequence of 14 M/M/1 based queuing station that corresponds to each SIP message in as shown in Figure(5),All the SIP packets are proposed by executing the same IP-PBX software module until each session is established and tore down.IP-PBX model and assumptions of [2] were;

1.The aggregate arrival processes are approximately by a Poisson process. this is due to the well-known property that any process which results from the superposition of a large number of independent point processes approaches a Markov process, if each individual point process" thins out" within the result process.

2The input and output processes traffic in one direction and according to Markov model, their previous values are not dependent. call setup scenario was analyzed by M/M/1and M/M/c queuing model.

3. Performance and reliability of SIP protocol was analyzed by the mean response time and the mean number of jobs in the system. *Mean response time* is the difference between the times it takes for an INVITE sent from UAC to reach IP-PBX until the final response is sent by IP-PBX to UAC. Mean number of calls is defined as the mean number of sessions that are currently in the system [2] .

"Equation (2)",Mean number of jobs N (random variable) in the system at study state is given by the system at study state is given by currently in the system [18,19].

$$N = \sum_{k=1}^{J} \rho_k /(1-\rho_K) \qquad Where \qquad \rho_k = \lambda_k / \mu_k \quad (2)$$

$$\lambda_j = \sum_{k=1}^{j-1} (\lambda_k Q[k,j]) \qquad for \qquad 1 < j \leq J \quad (3)$$

J=14 is the number of stations in the queuing model. Qψ is the one step probability matrix corresponding to the queuing model; that is, Q[i,j]the probability that a job departing station iψ goes to station j. The mean response time for calls is by Little's law R=N/ . They assumed the service rate is fixed at 0.5ms• 1 and the arrival rate at 0.3 ms• 1, Alouf et al [2].

4. Since UAS doesn't parse the SIP packets, the computation will be less, hence the service times are assumed as 0.7/µ for sending the 180 followed by 200 with 0.3 /µ or non 200 responses with 0.5 /µ.

5. Only 80 percent of the INVITE messages will be successful in getting the 180 response and 90 percent of that 180 responses will get the 200 response. Remaining 20 percent of the INVITE messages will get a non-200 response and 10 percent of the 180 responses will receive a non-200 response.

This architecture is a developed model based on the limited capacity of the single server queues for estimating the buffer size and studying the intensity of the generated traffic.

Our key observation was as follows: In earlier realizations of the IP-PBX, it was typical to run each main stage of the processing as a separate standalone concurrent process. In such a realization, each such module is amenable to modeling as a queue, because the coupling between different modules is unsynchronized and through message passing. Such a realization is useful and efficient if the pattern of processing of different packets (i.e. which modules they go through and in what order) varies from packet to packet, and has some statistically (but not deterministically) characterizable pattern. However, it is not any more efficient than a synchronized (function call) processing if almost all packets are expected to go through the same sequence of processing, as is now understood to be the pattern of proxy processing for the INVITE packet which produce most stress on the server (At best, it is equally efficient; at worst, it is significantly less efficient due to messaging and other overhead) .

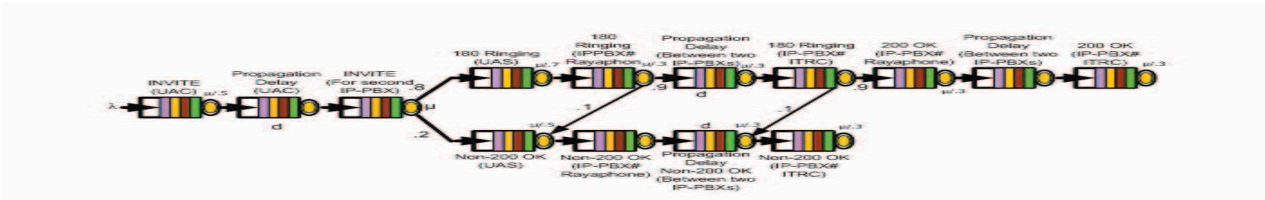

Fig. 5. IP-PBX M/M/1 Queuing delay Model

The M/M/m queuing system is the same as the M/M/1 system except there are m servers (output links). A customer at the head of the queue is served by any server that is available [1].

The new implementation is based on a single queue with multiple servers, and each packet is run through the entire processing before the next one is picked up. Concurrency can still provide some advantage, but now it is used simply to tune the system capability to the load, not linked to software modules. In other words, the entire processing for one packet is performed in one thread of execution. The value of c is determined based on the previous performance study done on the IP-PBX that includes CPU, processing speed and memory characteristics[1].

The mean response time and the mean number of calls for the M/M/c queuing based SIP model can be obtained from any standard work [18,19] and are as follows:

*"Equations (2,3)",*The mean response time and the mean number of calls for the M/M/cψ queuing based SIP model can be obtained from any standard work [18,19] and are as follows:

$$W = \frac{1}{\mu} + [(\lambda/\mu)^{c} \cdot \frac{\mu}{(c-1)!(c\mu-\lambda)^{2}}]P_0 \quad (3)$$

$$L = \frac{\lambda}{\mu} + [(\lambda/\mu)^{c} \cdot \frac{\mu}{(c-1)!(c\mu-\lambda)^{2}}]P_0 \quad (4)$$

$$P_0 = [1 + \sum_{n=1}^{c-1} \frac{(c\rho)^{n}}{n} + \frac{(c\rho)^{c}}{c!} \cdot \frac{1}{1-\rho}]^{-1} \quad (5)$$

The mean response time, mean number of jobs and server utilization data was computed by providing appropriate input values such as , µ and c (only in case of M/M/c) based on the M/M/1 and M/M/c IP-PBX model. We considered c = 3 (number of threads) in our M/M/c model calculations [1].

B. Calculated results in SIP network with propagation delay

In this paper was considered former assumption mentioned and was assumed µ=0.5, in order to calculate system's mean response time with publication delay varying between 0 - 10 ms where the distance between UAC and IP-PBX is 0 - 1000 miles (Figure6,7,8,9).

Each 100 miles is assumed to be equivalent with 1ms delay. As one can see the mean response time with variation of arrival rate is approximately linear. The mean response time increases with publication delay.

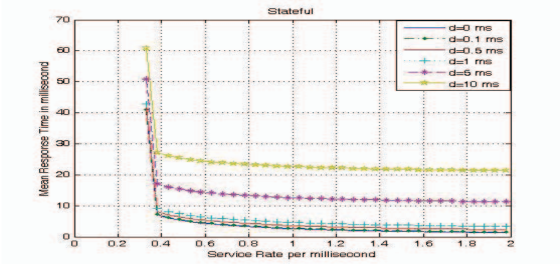

Fig. 6. Mean response time base on service rate changes with different delays

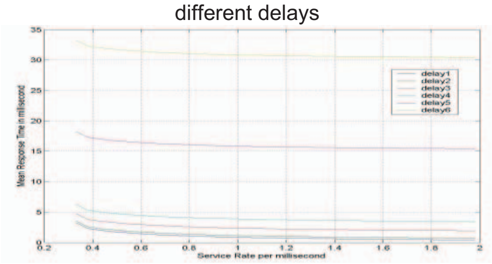

Fig. 7. Mean response time base on service rate changes with different delays (M/M/3 Queuing Model)

Mean number of jobs based on service rate changes with different delays are shown in Figures8,9. As we observe the maximum average number of jobs is less than 10 in maximum distance of 1000 miles. The results of Mean number of jobs and mean response time are obtained by changing base service rate delay while (= 0.3) is supposed constant.

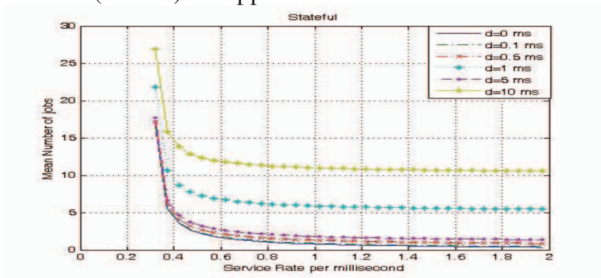

Fig. 8. Mean number of jobs base on Service rate changes with different delays

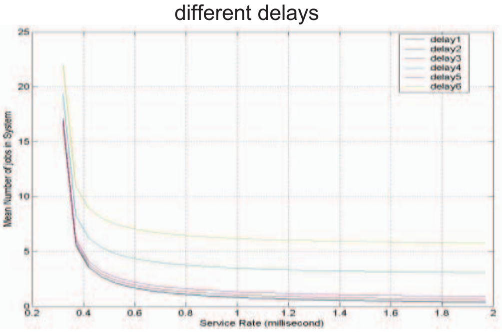

Fig. 9. Mean number of jobs base on Service rate changes with different delays (M/M/c Queuing Model)

C. Comparative between M/M/1 and M/M/C traffic Model in our call flow

1. During the experiments, we observed that the M/M/c

model performed much better when the arrival rate of the incoming packets increases compared to M/M/1 model.

2. In case of simulated M/M/1 model, there are 15 queuing stations processing all the incoming SIP packets, the average response time to establish each SIP session is much longer than the M/M/c model with 3servers per milliseconds. This indicates M/M/c model showed a significant performance improvement compared to M/M/1 model.

3. Average response time and the mean number of calls increases when the call arrival rate increases. This confirms our intuition that the M/M/c model more nearly represents the correct architecture of the SIP proxy server.

VIII. Reliability Analysis

The reliability metrics of interest for the MGCs are the steady-state system availability and the probability of call setup delay (i.e. loss of MEGACO call requests). a standard reliability model was developed for the MGC. The reliability model is then combined with the queuing performance model of Section 7 to predict the probability of call setup delay versus number of subscriber. For this analysis, we used the hierarchical reliability and per formability models and associated closed form expressions for computing availability and loss probability presented in [22].(Figures 10,11)

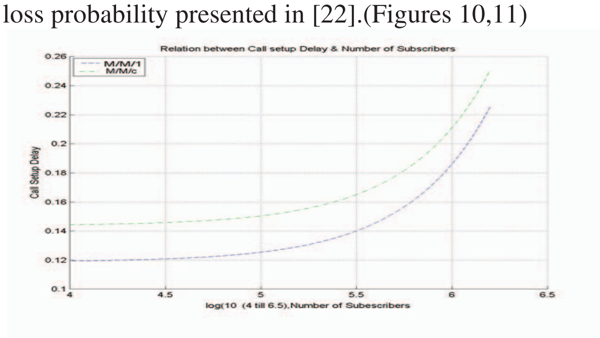

Fig. 10. Call setup delay versus Number of subscriber (Comparative M/M/1&M/M/c)

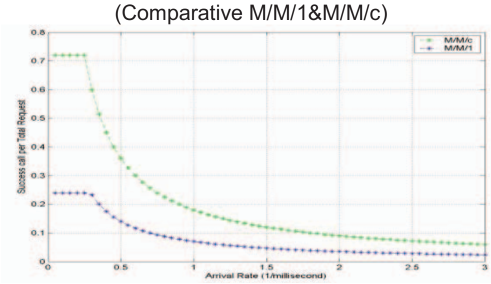

Fig. 11. Success call per Total request versus Arrival rate (Comparative M/M/1&M/M/c)

IX. Conclusions and Future works

Based on the measurements and analysis, SIP network architecture was modeled by Two IP-PBX in various remote locations with presence propagation delay in queuing model. the M/M/cψqueuing model was better than the M/M/1- model with 14 queuing stations, proposed by authors in [2] for prediction of server performance. The average response time, mean number of jobs and server utilization factor of the M/M/c model was a more predictable model with significant performance improvements and also met the ITU-T standards [20,21]. In future, we intend to expand this research, considering MEGACO protocol located in various remote locations and do a comparative study of the performance between SIP and MEGACO, when network delays are introduced into these models.

ACKNOWLEDGMENT

This research was support by Iran Telecommunication Research Center (ITRC). The authors would like to thankthis institute for its financial support.

REFERENCES

[1] Measurements and Analysis of M/M/1 and M/M/c Queuing Models of the SIP Proxy ServerSureshkumar V. Subramanian, Rudra Dutta IP Communication Business Unit, CISCO Systems,Research Triangle Park ,USA,2008

[2] V.K.Gurbani, L. Jagadeesan, V.B. Mendiritta, "Charecterizing the Session Initiation Protocol (SIP) Network Performance and Reliability", ISAS 2005: LNCS 3694, pp. 196-211, April 2005

[3] J-S. Wu and P-Y Wang, "The performance analysis of SIP-T signaling system in carrierclass VoIP network", Proceedings of the 17th IEEE International Conference on AdvancedInformation Networking and Applications (AINA), 2003.

[4] F. Lipson, "Verification of Service Level Agreements with Markov Reward Models,"South African Telecommunications Networks and Applications Conference, September2003.

[5] N. Rajagopal, M. Devetsikiotis, "Modeling and optimization for the design of IMS Networks", *Proceedings of 39th Annu. Simulation Symp.2006 (ANSS06)*, Huntsville, Alabama, USA, April 2006.

[6] J. Rosenberg, J. Schulzrinne, H. Camarillo, P.J. Sparks, R. Handley,E. Schooler, "SIP: Session Initiation Protocol",IETF RFC 3261,http://www.ietf.org/rfc/rfc3261.txt, June 2002.

[7] A. Johnston, "SIP: Understanding Session Initiation Protocol", Artech House Publishers, Volume 1, 2nd edition, Boston, 2004.

[8] RFC 3261, SIP: Session Initiation Protocol

[9] RFC 3986, Uniform Resource Identifiers (URI): Generic Syntax, IETF, The Internet Society (2005)

[10] William Stallings book, p.214

[11] William Stallings book, pp.214-215

[12] William Stallings book, pp.216-217

[13] Ben Sharif "ELASTIX WITHOUT TEARS"Http://www.elastixconnection.com/downloads/elastix_without_tears.pdf,oct 11,2008

[14] "Wireshark Protocol Analyzer" http://www.wireshark.org/ by Wireshark

[15]Http://www.asteriskguru.com/tools/bandwidth_calculator.php

[16] Http://www.VoIP-calculator.com/calculator/lipb

[17]Http://www.connect802.com/VoIP_bandwidth.php

[18]I. Adan, J. Resing, "Queuing Theory", Class notes, Department of ComputerScience and Mathematics, Eindhoven University of Technology, The Netherlands, February 2001.

[19] William J. Stewart, " Probability, Queuing Models and Markov Chains:The tools of System Performance Evaluation", Lecture Notes, Departmentof Computer Science, North Carolina State University, Raleigh, USA, Fall2003.

[20] International Telecommunications Union-Telecommunications StandardizationSector (ITU-T), "Principles of Intelligent network archetecture",Recommendation Q1201, ITU-T,Geneva, Switzerland, October 1992.

[21] International Telecommunications Union-Telecommunications StandardizationSector (ITU-T), "Network grade of service parametersand target values for circuit-switched services in the evolving ISDN", RecommendationE.721, ITU-T, Geneva, Switzerland, 1992