IMPROVING TRAFFIC PERFORMANCE ON IMS NETWORK

BA Alassane[1], OUYA Samuel[2] and FARSSI Sidi Mohamed[3]
[1,2]Laboratory of Computer, Network and Telecommunications, Cheikh Anta Diop University of Dakar, Dakar, Senegal

[3]Laboratory of Medical, Imaging and Bioinformatics Cheikh Anta Diop University of Dakar, Senegal

ABSTRACT

*The IP Multimedia Subsystem (IMS***)** *enables the convergence of voice, data, and multimedia services such as Voice over IP (VoIP), Video over IP, push-to-talk, presence or instant messaging services and so. IMS is well integrated with existing voice and data networks, while adopting many of their key characteristics.*
SIP (Session Initiation Protocol) provides a pathway to build a single unified network, bridging the gap that previously existed between the once-separated Telecom and Internet networks.
The Call Session Control Functions (CSCFs) servers are the key part of the IMS structure. They are the main components responsible for processing and routing signalling messages.
Presence is a service that allows a user to be informed about the reachability, availability, and willingness to communicate of another user. A study shows that presence service can account for 50% or more of the total signalling traffic the IMS core network [1]because of NOTIFY messages flow. So, we try to optimizing their mechanism by giving proposal of reducing SIP latency and waiting time for service in a multi-server environment.
Solutions proposed will ovoid SIP server overload and therefore degradation of QoS (Quality of Service).

KEYWORDS: *IMS, SIP, CSCF, Presence Service, QoS*

I. INTRODUCTION

IMS needs to offer high level of interaction for users. The fact of integrating different networks into one multifunctional IP needs to create a unified communication environment for fixed and mobile users, by offering enriched and integrated services. Those demands, with appropriate levels of quality, are not a simple task.
If the network is not initially properly dimensioned, in short time the nodes will fall into a state of congestion, which will lead to performance degradation of a part or the entire network.
In order to eliminate or reduce the problems of SIP server overload, various approaches have been proposed.
The paper is interested to the traffic flow in the CSCFs and the Presence Server and the proposal to improve its performance for an efficient service.
This paper is organized as follows: section 2 gives the traffic organization in an IMS session. In section3, traffic performance approaches are studied. Mathematical analysis of queue length distribution is also given and proposals of improving traffic. The paper concludes in Section 4.

II. THE IMS TRAFFIC

Call Session Control Functions (CSCFs) are the session routing points in the IMS core network. They distribute incoming calls to the application services.
The CSCFs handle initial subscriber authentication. Application services that receive a message from the CSCFs are defined to permit the processing of that call, and to perform additional service-related checks.

The first point of contact for the user equipment (UE) to IMS network is the Proxy-CSCF. It forwards SIP to and from the home network and may also perform encryption and compression. The Interrogating-CSCF (I-CSCF) is the entry point to the home network. It may function similar to a firewall and hide the internal topology.
At last, the Serving-CSCF (S-CSCF) is the main element in session control. It is fully responsible for registration and controlling of sessions to the UE. It also decides which Application Servers (AS) that needs to be triggered, depending on the Initial Filter Criteria (IFC). The IFC is part of the user profile which is held in the HSS (home subscriber server) and downloaded to the S-SCSF upon registration.
The HSS is a database that contains all subscribers' data, like the services that is allowed to access, the network in which he is granted to roam and the information about the location of the subscriber. Once information about the subscriber has changed, the entire profile is sent to the S-CSCF, making it always synchronized with the HSS. An important function of the HSS is to provide the encryption and authentication keys of the user: when a user registers himself in the network, he must provide the credentials to the S-CSCF and these are checked against the one stored in the HSS [14], [15].
Another important element is the Presence Server (PS), which holds the presence status of each subscriber and a list of 'watchers' that are interested in that information. Presence Service allows a user to be informed about the reachability, availability, and willingness to communicate of another user.

III. PERFORMANCE TRAFFIC APPROACH

For improving IMS traffic, we use two approaches. The first is the flow to and from the Presence service; the second is from the CSCFs. We study SIP overload and quality of service (QoS).
When a presentity's state changes, all its watchers will have to be notified. A recent study shows that presence service (PS) can account for 50% or more of the total signalling traffic the IMS core network handles [1]. From it, the largest portion of the traffic load to the PS is resulted from NOTIFY messages, which are used to notify the watchers of a presentity. This is quite a burden for an IMS network and need to be tackled.

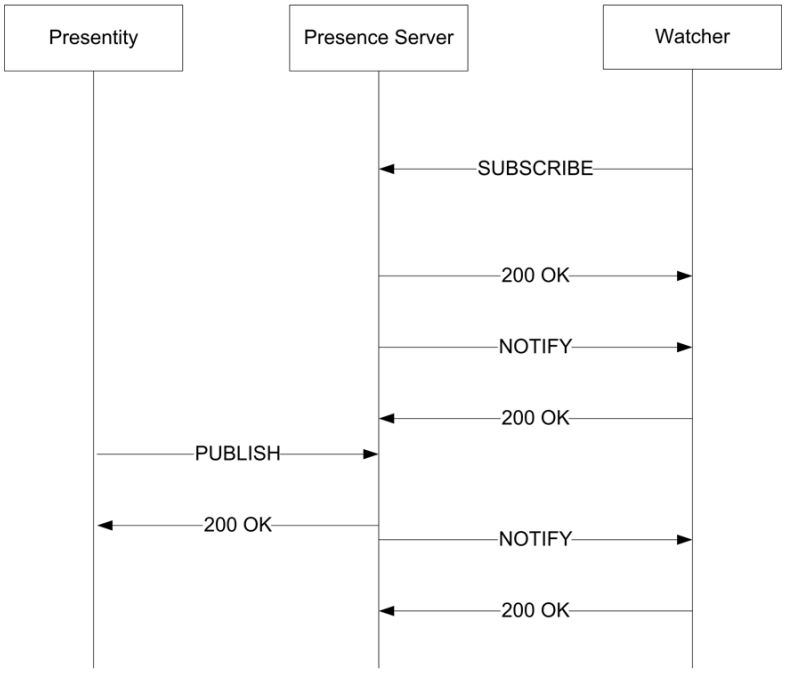

Figure1. SIP presence call flow

NOTIFY messages arrival rate to a PS varies greatly with the online watchers [12].So, optimizing their process mechanism is important to guarantee performance of PSs.

Traffic to a PS is divided into eight types: initial publish, refresh publish, modify publish, terminal publish, initial subscribe, refresh subscribe, terminal subscribe, and notify.
Except for refresh PUBLISH, all the other PUBLISH messages result in NOTIFY messages being sent to all the online watchers. It's a burden.
Suppose transaction generation rate of each type is denoted as r_i , i = 1...8.
Transaction ratio of each type of traffic is given by:

$$\frac{r_i}{\sum_{i=1}^{i=8} r_i} \quad (1)$$

The load incurred by each type of transaction can be denoted as $\rho_{i} = r_i \times t_i$, i = 1,….., 8
Where t_i is the processing time for transaction type i, traffic load ratio resulted by each type of traffic is given by:

$$v_i = \frac{\rho_i}{\sum_{i=1}^{i=8} \rho_i} \quad (2)$$

The transaction rate ratio as well as the traffic load ratio resulted from the different type of messages is calculated based on equations (1) and (2).
Let use data taken from a reference implementation of an IMS presence server when initial publish rate =1/4; that means user logs in once every 4 hours.

Table1: Time for a presence server to process transaction above

Transaction	Time
Initial publish rate	1/4
Refresh publish rate	2
Terminal publish rate	1/4
Modify publish rate	0,5
Initial subscribe rate	0,25
Refresh subscribe rate	2
Terminal subscribe rate	0,25
Notify rate	**n_w online watcher x** (0,5+0,25+0,25)

notify rate = n_w online watcher ×(modify publish rate + initial publish rate + terminal publish rate).
Figure2 indicates that when the number of online watchers is larger than 3, traffic load resulted by the NOTIFY traffic is much higher than the other types of traffic.

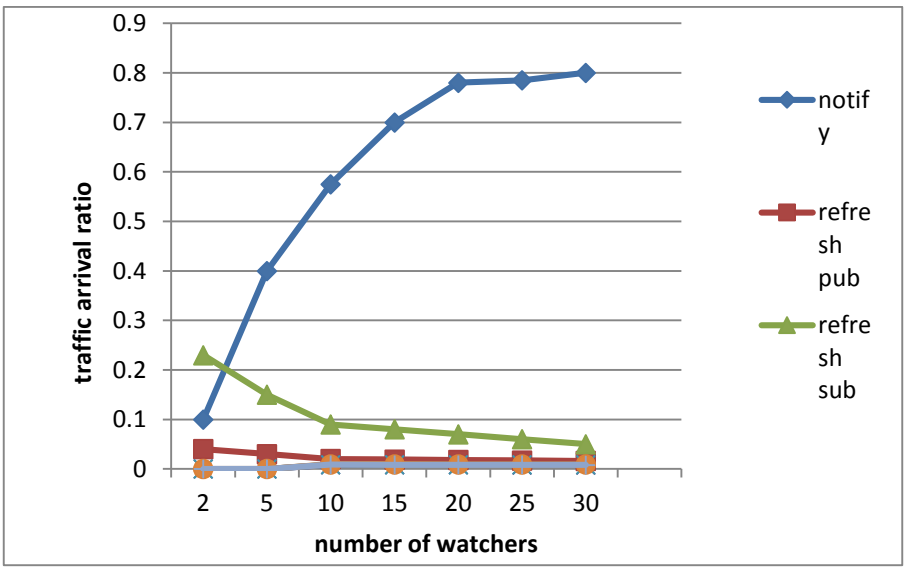

Figure2. Traffic arrival ratio vs Number of watchers

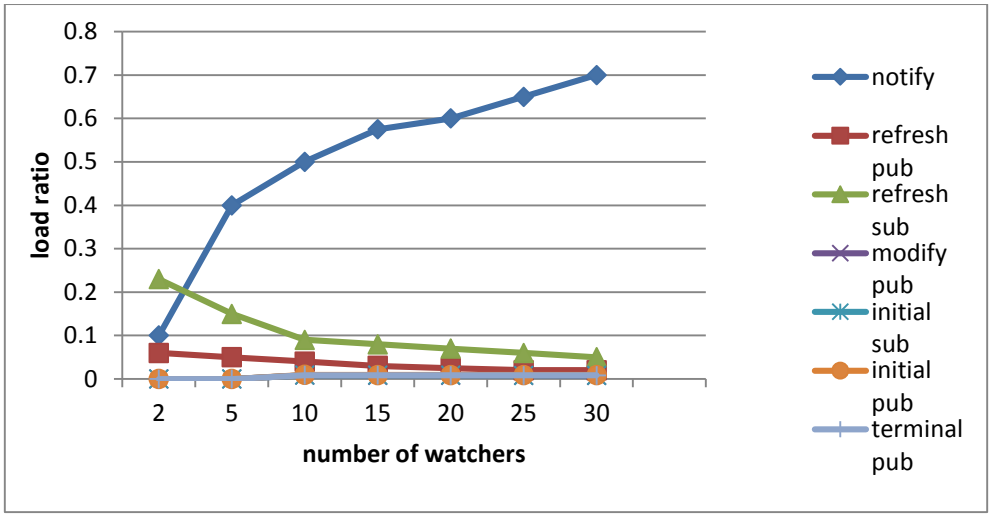

Figure3. Load ratio vs. Number of Watchers

Figure3 indicate that NOTIFY messages ratio is always the largest part the traffic when number of online watchers ≥ 3.
They give us an impression on the amount of traffic load resulted by each type of transactions and help us to identify the potential performance bottleneck when we implement or deploy a PS.
Suppose queue of NOTIFY messages taken as a queuing system with controlled vacation and batch Poisson arrival. That is, NOTIFY messages are sent periodically and when the queue has no NOTIFY messages, the server controlling the NOTIFY queue will be on vacation.
When the vacation is ended after time interval **T=1/θ**, NOTIFY messages in the queue will be sent. Each PUBLISH message results in multiple NOTIFY messages to different watchers. Suppose B the maximum number of messages that can be buffered by NOTIFY queue. When the buffer is full, arriving messages are discarded.
So it is important to control queue length of NOTIFY to control the message loss probability.
Let do mathematical analysis.
When, Q(t) denotes the number of messages in the system at time t and B is the maximum buffer size
S(t) = 0 and S(t) = 1 denote the events that the server is busy and on vacation at epoch t respectively;
Define

$$p_{j,0}(t) = P(Q(t) = j, S(t) = 0), j = 1,2, \ldots, B \quad (3)$$

$$p_{j,1}(t) = P(Q(t) = j, S(t) = 1), j = 0,2, \ldots, B \quad (4)$$

From the theory of Markov chains [7], [8], [9], $\{Q(t), S(t), t \geq 0\}$ has a unique equilibrium distribution

$$p_{j,0} = \lim_{t\to\infty} p_{j,0}(t), j = 1,2, \ldots, B \quad (5)$$

$$p_{j,1} = \lim_{t\to\infty} p_{j,1}(t), j = 0,2, \ldots, B \quad (6)$$

Then, probability that a message arrives and finds there are more than K messages in the buffer is:

$$P_K = \sum_{i=K+1}^{i=B} (p_{i,1} + p_{i,0}) \quad (7)$$

And probability of the buffer being full is:

$$P_B = p_{B,1} + p_{B,0} \quad (8)$$

Mean and variance of queue length are now:

$$\bar{Q} = \sum_{j=0}^{j=B} j(p_{j,0} + p_{j,1}) \quad (9)$$

$$Var(Q) = \sum_{j=0}^{j=B} j^2(p_{j,0} + p_{j,1}) - \bar{Q}^2 \quad (10)$$

$$p_{j0,0} = 0$$

$\bar{Q}$ reflects the average memory space required by the NOTIFY message queue; Var(Q) gives the variance of the memory space requirement which can be used to estimate the peak hour memory space occupation

Suppose NOTIFY messages arrive at the system according to Poisson process with parameter λ and there is one server in the system. The queue discipline is FCFS (first come, first served).
Suppose now service time for a NOTIFY message is assumed to be exponentially distributed with mean **1/μ,** and $\mathbf{n_w}$ the number of online watchers . The traffic intensity $\rho = \boldsymbol{\lambda}\ \mathbf{n_w/\mu}$
Result of P_K calculation with **λ** = 1, $\mathbf{n_w}$ =4, **μ** = 6, **B**=100, **K**=80 give queue length distribution.

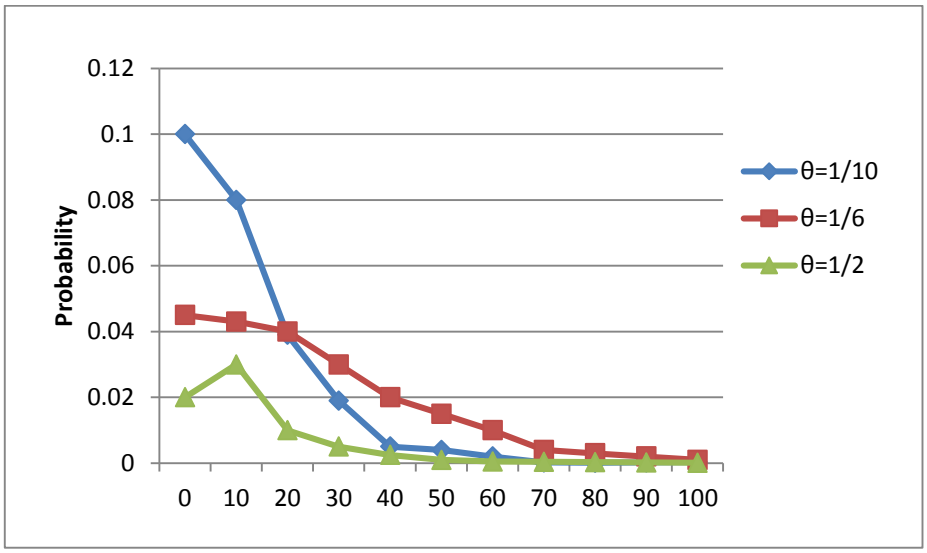

Figure4. Queue Length Distribution with **λ**=1, $\mathbf{n_w}$=4, **μ**=6

Figure4 gives the distribution of queue length with different vacation intervals **T=1/θ** (T in ms)
With the longer vacation time, the variation of queue length becomes larger.
The probability of queue length exceeding 80%B is given on figure 5.

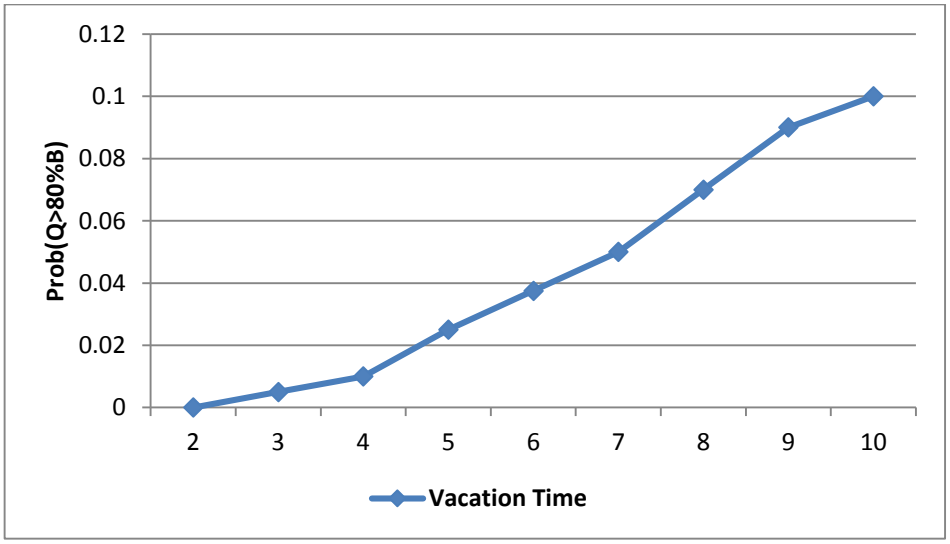

Figure 5. Probability of queue length exceeding 80%B

Note: simulations given using **Matlab**
It indicate the longer vacation time, the larger the probability of queue length exceeding 80%B.
Consider the Erlang C formula:

$$P_W(N) = \frac{\frac{A^N}{N!}\frac{N}{N-A}}{\sum_{i=0}^{i=N-1}\frac{A^i}{i!}+\frac{A^N}{N!}\frac{N}{N-A}} \qquad (11)$$

Where:
A is the total traffic offered in units of erlangs

N is the number of servers
Pw is the probability that a customer has to wait for service.
With A= λ/μ. If:
N=1 Pw = 0, 16
N=2 Pw = 0,012
N=3 Pw= 6, 95 10^{-4}
N=4 Pw= 6.852 10^{-4}

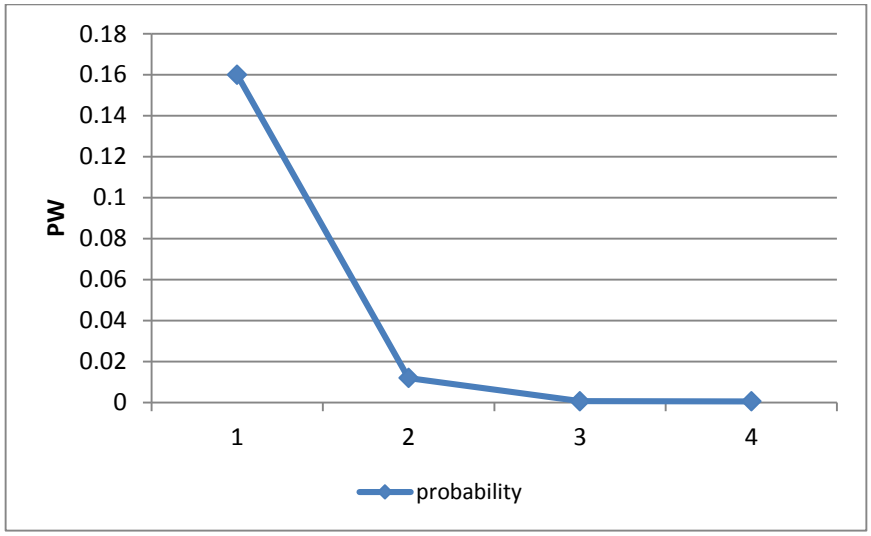

Figure 6. Probability that a customer wait for service

We note lim Pw (n) =0
It is proven the effects of adding servers to the networks assuming M/M/r model with **Erlang C formula** [2]. When adding servers to the network, total waiting time in system is reduced, which has direct impact on system overload as shown when N=1;2;3;4 and in follow figure7.

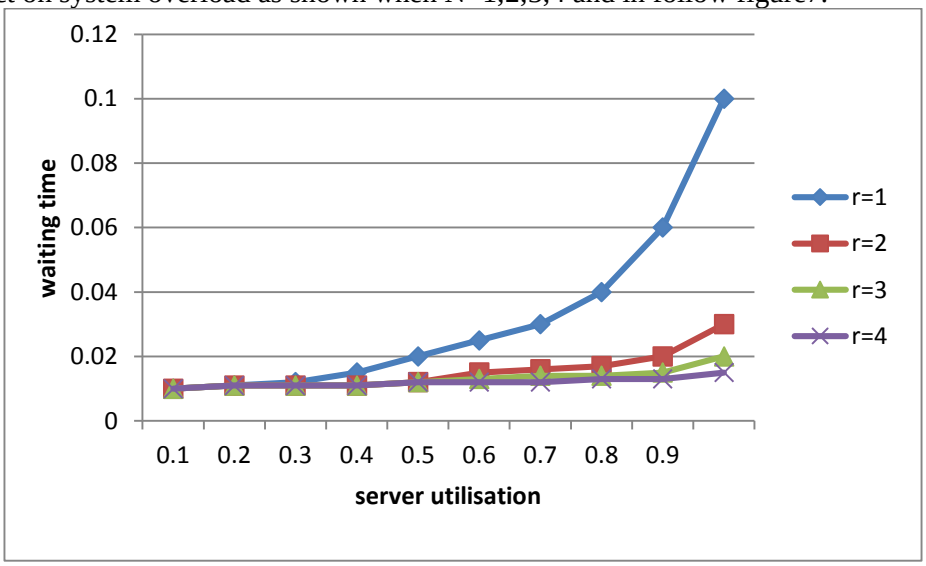

Figure 7. Mean waiting time in multi-server

Let introduce a QoS management on a session basis.
The SIP proxy server is supposed to achieve measures on session durations and data volume exchanged during session initiation phase.
When CSCFs servers (P-CSCF, I-CSCF, S-CSCF) are running on the same host, the SIP message can be internally passed between SIP servers using a single operating system mechanism like a queue. It increases the reliability of the network.

In that case, the SIP messages never live the host, and the IP network is not used at all. This arrangement provides an improvement in performance of IMS network. This performance by co-location IMS servers is measured by SIP latency which is the time spent when a SIP server sends a message to another server to when the second server receives this message. This time will decrease from **25** ms less than **1ms** [3].

IV. CONCLUSION

Using the Erlang C formula, we noted that adding servers increase performance by reducing waiting time for service.
We propose for each type of service (between I-CSCF and S-CSCF (call, data, multimedia.)) to use less than two servers well dimensioned and running on the same operating system.
In all cases, threshold values must be defined to reflect the average memory space required by the NOTIFY message queue , to gives the variance of the memory space requirement which can be used to estimate the peak hour memory space occupation (var(Q)) and reduce waiting time to increase performance. It is good indications for CPUs needed in dimensioning traffic for next studies.

V. FUTURE WORK

Future research needs to work on modeling traffic on IMS nodes, particularly on CSCFs node; it will provide results on separation of incoming flows, and then reduce SIP overload and latency. It also shows dependency between system load and number of servers.

REFERENCES

[1]. C. Chi Bell Laboratories, Alcatel-Lucent chic, Hao Bell Laboratories, Alcatel-Lucent D.Wang Bell Laboratories, Alcatel-Lucent Z. Cao Institute of Information Science Beijing Jiaotong University, Modelling IMS Presence Server: Traffic Analysis &Performance, 2008 IEEE
[2]. Mesud Hadžialić, Mirko Škrbić, Nerma Šečić, Mirza Varatanović, Elvedina Zulić, Nedim Bijedić University of Sarajevo, Problem of IMS modeling – Solving Approaches, CTRQ 2012
[3]. Mlindi Mashologu, Performance optimization of IP Multimedia Subsystem, Dissertation.com, Boca Raton, Florida 2010
[4]. V.S. Abhayawardhana, R. Babbage, A Traffic Model for the IP Multimedia Subsystem (IMS), 2007 IEEE
[5]. Michael T. Hunter Russell J. Clark Frank S. Park Security Issues with the IP Multimedia Subsystem (IMS): A White Paper College of Computing Georgia Institute of Technology {frank,mhunter,russ.clark}@gatech.edu Version 1.0 September 1, 2007
[6]. Pierre COATANEA, IMS : IP Multimedia Subsystem, EFORT
[7]. Nils Berglund, Chaînes de Markov, Université d'Orléans Décembre 2007
[8]. Alexander Klemm, Christoph Lindemann, and Marco Lohmann Modeling IP Traffic Using the Batch Markovian Arrival Process, Modeling IP Traffic Using the Batch Markovian Arrival Process University of Dortmund Department of Computer Science August-Schmidt-Str. 12 44227 Dortmund, Germany
[9]. Modelling Markovian Queues and Similar Processes Winfried Grassmann Department of Computer Science University of Saskatchewan, CORS - SCRO 2000 ANNUAL CONFERENCE Energy, Natural Resources, and the Environment MAY 29-31, 2000 EDMONTON, ALBERTA
[10]. JEREMIAH F. HAYES THIMMA V. J. GANESH BABU Modeling and Analysis of telecommunications network 2004 by John Wiley & Sons, Inc. All rights reserved
[11]. Jeffrey P. Buzen Harvard University and Honeywell Information Systems, Computational Algorithms for Closed Queueing Networks with Exponential Servers Communications September 1973 of Volume 16 the ACM Number 9
[12]. Muhammad T. Alam, Graduate Student Member IEEE, Zheng Da Wu, Member IEEE Admission Control Approaches in the IMS, Presence Service,international journal of computer science volume 1 number 1 2006
[13]. Henning Schulzrinne, Personal Mobility for Multimedia Services in the Internet, IDMS'96 (European Workshop on Interactive Distributed Multimedia Systems and Services), Berlin, Germany, March 4-6, 1996

[14]. Hassan HASSAN Modélisation et analyse de performances du trafic multimédia dans les réseaux hétérogènes, mémoire Doctorat de l'Université Paul Sabatier–Toulouse III 2007
[15]. PETTER LINDGREN, Diameter service creation investigation and HSS, Evaluation, Master's Thesis Supervisor: Stefan Östergaard, Kajsa Goffrich Examiner: Prof. Gerald Q. Maguire Jr, 2010-02-23

AUTHORS

BA Alassane is a PhD student in telecommunications holder of a DEA in Mathematics.

OUYA Samuel is the head of the Laboratory of Computer, Network and Telecommunications, Cheikh Anta Diop University of Dakar, Senegal. Doctor OUYA is a mathematician IT specialist having a long experience in teaching and search. He participates in several IT meets.

FARSSI Sidi Mohamed is the head of the Laboratory of Medical, Imaging and Bioinformatics, Cheikh Anta Diop University of Dakar, Senegal. Doctor FARSSI is a IT specialist who participate in several IT and medical meetings