ON ANALYTICAL MODELING OF IMS CONFERENCING SERVER

Pavel Abaev
Vitaly Beschastny
Alexey Tsarev
Department of Applied Probability and Informatics
Peoples' Friendship University of Russia
Mikluho-Maklaya str., 6
Moscow 117198, Russia
E-mail: {pabaev, vbeschastny, atsarev}@sci.pfu.edu.ru

KEYWORDS
IMS, SIP, conferencing server, exhaustive service, vacations.

ABSTRACT

The IP Multimedia Subsystem is an architectural framework for delivering multimedia services over an Internet Protocol (IP) network. Originally, it was specified for wireless networks, but has since evolved to incorporate fixed line access as well. It forms part of a Next Generation Network (NGN) which is defined as a packet-based network where the service functionality is independent of the underlying transport technologies. This allows new converged services to be implemented on top of an existing packet switched network.

The centralized conferencing framework limits the ability to have large number of participants because of the overhead on the server. One of the possible efficient solution for increasing of server's performance is grouping meeting requests the same video conference and serving them at one time on an edge-proxy. The queuing model with batch exhaustive service and working vacations is constructed and analyzed. The stationary distribution and the formulas for main performance measurements are obtained. In addition, the optimization problem for increasing a server performance is formulated.

INTRODUCTION

The IP Multimedia Subsystem is an architectural framework for delivering multimedia services over an Internet Protocol (IP) network. IMS uses SIP protocol to support communication sessions with multiple participants (Camarillo, 2008; TS 23.228, 2004). SIP is an application-layer signaling protocol for creating, modifying, and terminating sessions with one or more participants. In 1996, Henning Schulzrinne and Mark Handley started working on creating a SIP protocol within the IETF (Internet Engineering Task Force) project for developing a series of protocols for the provision of multimedia services. In November 2000, SIP was accepted as a 3GPP signaling protocol and main protocol of the IMS architecture (TS.24.229, 2015). In 2002, the RFC3261 (Rosenberg, 2002) recommendation which determines the current protocol form was accepted.

SIP supports both centralized and decentralized frameworks of multi-party communication. Overview of the most recent researches done by Mishra in (Mishra, 2014) shows that distributed approach allows to enhance the conferencing server capacity but it has some limitations with full security and control over the complete network due to unavailability of central server; it increases signaling delay due to multiple functional entities participating in conference session setup. In centralized conferencing architecture (Tien, 2010), a single User Agent (UA) or focus has a direct relation with each participant. Therefore, it shares many problems when the number of participants in conference increases. The centralized conferencing framework does not meet the increasing requirements for server's capacity due to growing demand for conferencing service. However, both of these approaches do not take into account SIP server overloading problems (Abaev, 2012; Abaev 2013). The SIP main shortcomings concerning overload prevention described in (RFC 5390, 2008) by Rosenberg.

One of the efficient approach to enhance server capability is to improve the way how the server serves incoming conferencing request. We consider the method of grouping requests related to the same conference and serving them at one time. This approach will reduce session establishment time and will increase the server throughput.

This paper is organized as follows. We analyze 3GPP standards for basic call flow for conference service session establishment. Then we investigate message grouping approach on the edge server and construct a mathematical model in the form of queuing system with the buffer of finite capacity. We obtain formulas for the main performance measures and perform numerical experiments.

Proceedings 30th European Conference on Modelling and
Simulation ©ECMS Thorsten Claus, Frank Herrmann,
Michael Manitz, Oliver Rose (Editors)
ISBN: 978-0-9932440-2-5 / ISBN: 978-0-9932440-3-2 (CD)

IMS CONFERENCING SERVICE CALL FLOW WITH REQUEST/RESPONSE GROUPING APPROACH

SIP is a request/response-based protocol. According to (TS.24.229, 2015), IMS architecture consists of the following components:

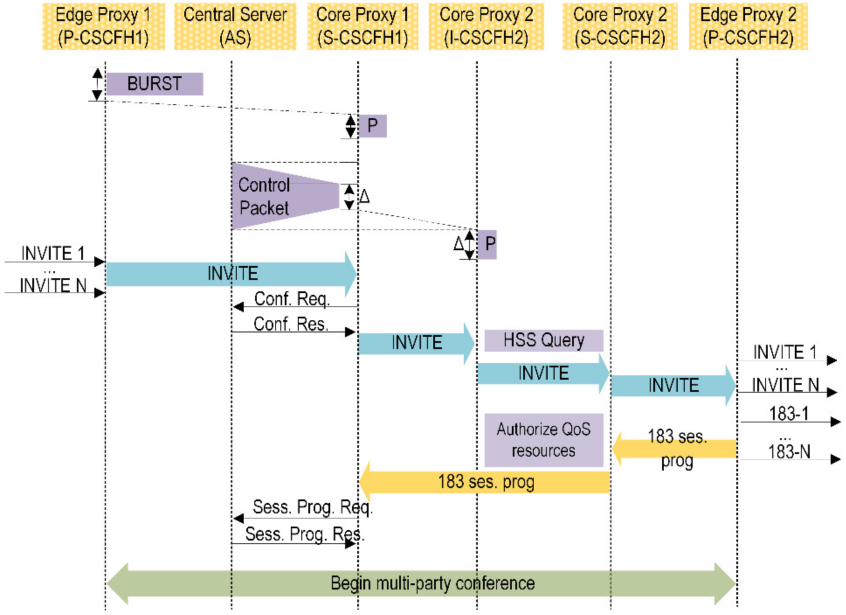
- UE (User Equipment), which takes the role for a request/response pair representing end users.
- central server known as application server (AS),
- SIP proxy servers (CSCFs) which includes edge-proxy (P-CSCF) and core-proxy (I-CSCF and S-CSCF).

Call flow diagram for conferencing session establishment is show in Fig. 1. Let us consider the procedure with more details. A group of UEs initiate a number of requests to try to establish a conferencing service and send INVITE messages towards Edge Proxy 1. All these messages should be delivered to the same P-CSCF server and can be aggregated by Edge Proxy 1 and send towards the target server as a single INVITE request with the list of participants.

When the INVITE message reaches Edge Proxy 2, the server will generate the required number of INVITE requests and send them towards the target UEs. Many 183 Session in Progress responses reach Edge Proxy 2 mean session establishment in progress. When the last response reaches Edge Proxy 2, the server will send a single 183 Session in Progress response towards the Edge Proxy 1.

The implementation of grouping approach on Edge Proxies enhances the bandwidth the whole system but requires to install in the network Edge Servers with higher performance.

A packet can be blocked either at an Edge Proxy or Core Proxy. The Edge Proxy model is used to quantify the probability that a packet is blocked at an Edge Server's buffer. In case of multi-party conferencing, the requests arrive as a batch and the destinations are in the group therefore the Edge Proxy modeled as bulk queue with an exhaustive service with multiple vacations can be used to provide efficient packet aggregation.



Figures 1: Conferencing flow session setup

IMS CONFERENCING SERVER MODEL

Let us assume that customers arrive at a single-server queue with finite capacity buffer R and receive service in accordance with FCFS policy. The customers reach the system in a batch according the Poisson process with rate λ. The bath size is a random variable ξ with the following distribution function

$$P\{\xi = i\} = q_i,\ i = 1,\ldots,K\ .$$

Let $\lambda_i = q_i\lambda$ be the intensity of arriving a batch with i-customers. If there is enough space in the buffer customers stand in a queue, otherwise they will be dropped.

Let n denote the buffer occupancy. Customers are aggregated in the buffer and receive service. Due to server capacity limitations, it processes up to L customers from the buffer at once. Furthermore, server does not wait until the number of customers in the queue exceeds L, but it processes $\min(L,n)$ customers. The processing time is exponentially distributed with the mean μ^{-1}, and it does not depend on the batch size.

The server takes working vacation at the time when the system is empty. During the vacation period new arrived customers are stored in the buffer. The server takes another new vacation if only there is no any new customer in the queue. The vacation time is a random variable which is exponentially distributed with parameter θ. While server is on the vacation, the buffer occupancy starts growing and if it exceeds the value H the vacation starts to process the customers from the buffer. Thus, the server operates in two modes: normal mode $(m=0)$ and vacations mode $(m=1)$, where m is the server operating status. The described system can be denoted as $M^{[X]} \,|\, M^{[Y]} \,|\, 1 \,|\, R \,|\, E, MV$.

The functioning of the system is described by the Markov process $\mathbf{X}(t) = (m(t), n(t))$ over the state space $\mathcal{X} = \{(m,n) : m \in \{0,1\}, 0 \le n \le R\}$.

Clearly $\mathbf{X}(t)$ is ergodic and thus stationary distribution exists. Let $p_{m,n} = \lim\limits_{t\to\infty} P(m(t)=m, n(t)=n)$ be the stationary distribution of the process. The non-diagonal and nonzero elements of infinitesimal operator $A = (a_{m,n})_{\substack{m=0,1,\\ n=0,\dots,R}}$ of the process $\mathbf{X}(t)$ can be written in the following way

$$a_{m,n} = \begin{cases} q_i\lambda,\ (m'',n'') = (m',n') + (0,i), \\ \qquad \begin{pmatrix} (n' \le R-i)\wedge \\ (i \le K)\wedge \\ (m''=m'=0) \end{pmatrix} \vee \begin{pmatrix} (n' \le R\text{-}i)\wedge \\ (i \le K)\wedge \\ (m''=m'=1) \end{pmatrix}; \\ \\ \mu,\ (m'',n'') = (m',n')\text{-}(\min(L,n'),0), \\ \quad (k'+k''=0)\wedge(n'>0), \\ \quad (m'',n'') = (0,0), (m',n') = (1,0); \\ \theta,\ (m'',n'') = (m',n')\text{-}(0,1), (n'=n'')\wedge \\ \qquad\qquad (m'=1)\wedge(m''=0). \end{cases}$$

Performance measures

Let π denote the blocking probability of batch

$$\pi = \frac{1}{\lambda}\sum_{i=0}^{K-1}\sum_{k=i+1}^{K}\lambda_k\left(p_{0,R-i}+p_{1,R-i}\right).$$

The utilization of the server $UTIL$ is given by the following formula

$$UTIL = \frac{1}{\mu}\sum_{k=1}^{K}k\lambda_k\left(1-\sum_{i=R+1-k}^{R}p_{0,i}+p_{1,i}\right)$$

The mean number of customers in the queue can be calculated as

$$Q = \sum_{k=1}^{R}k\left(p_{0,k}+p_{1,k}\right).$$

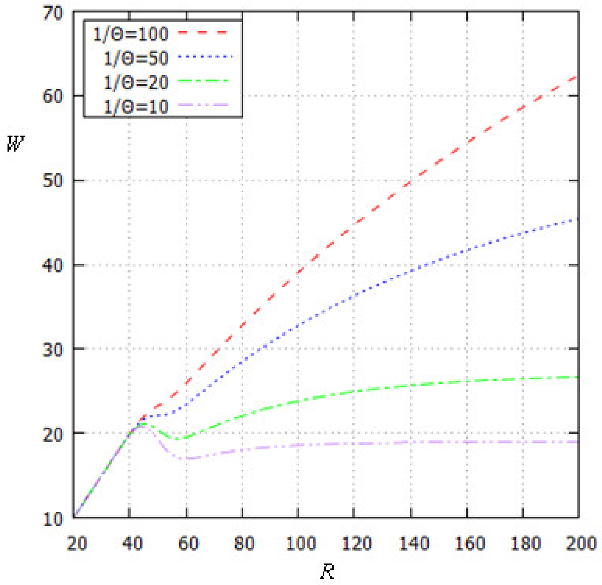
Using Little's formula, the mean waiting time in the queue is expressed as

$$W = \frac{Q}{\sum\limits_{k=1}^{K}k\lambda_k\left(1-\sum\limits_{i=R+1-k}^{R}\left(p_{0,i}+p_{1,i}\right)\right)}$$

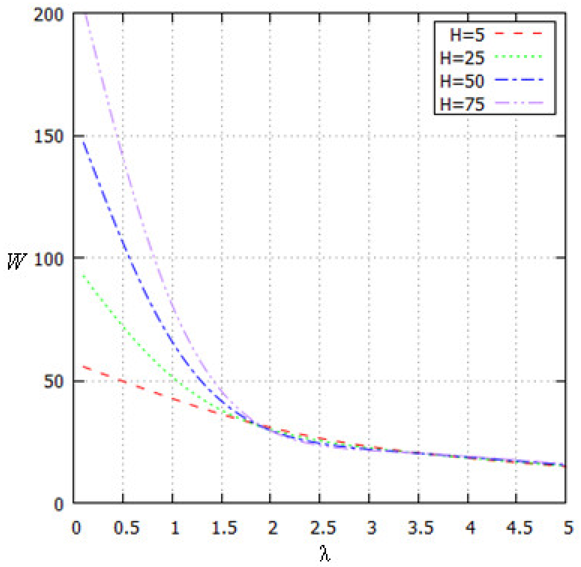
Numerical example

We perform several numerical experiments for a combination of different input parameters: $\lambda \le 5$, θ^{-1} from 10 to 100, $\mu = 6$, $R = 100$, $K = 10$, $L = 5$.

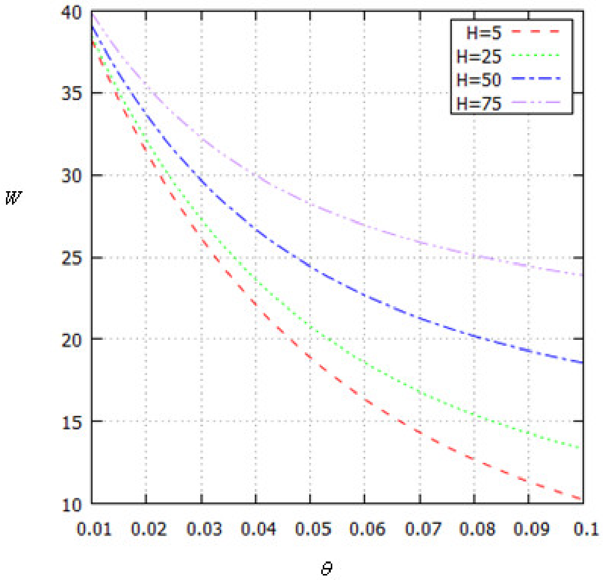
As shown in Fig. 2 as buffer size increases the mean waiting time increase from $R = 40$, but according Fig. 7 blocking probability asymptotically decrease. In observing Fig. 3 we see that increase of intensity rate λ does not make significant influence on mean waiting time for various value of threshold.



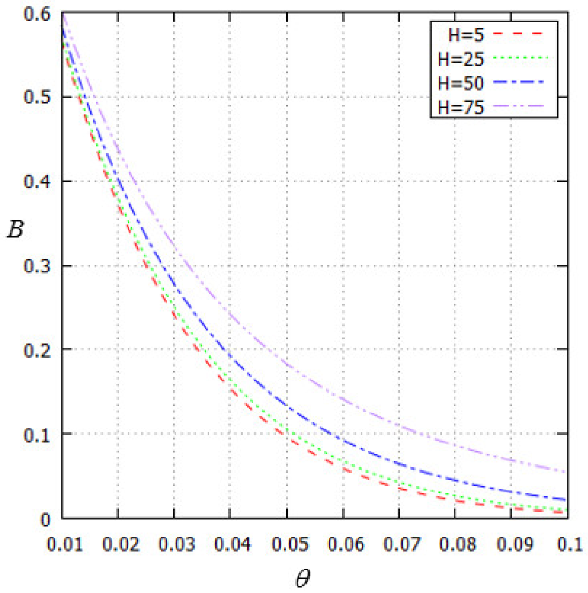
Figures 2: Mean waiting time on varying buffer size for different value of mean vacation time



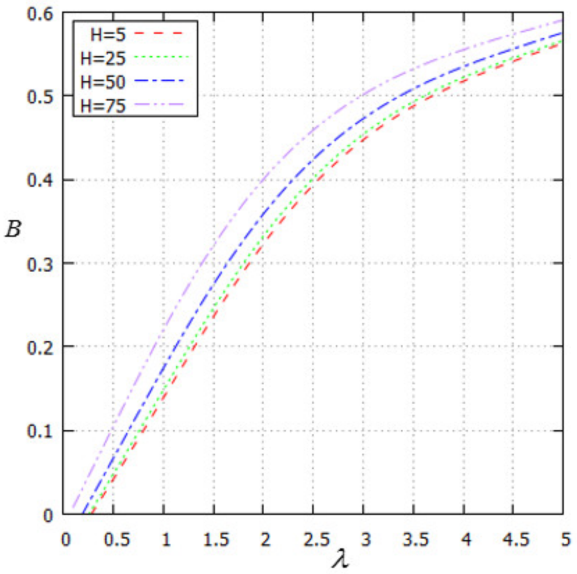
Figures 3: Mean waiting time on varying arrival rate for different value of the threshold



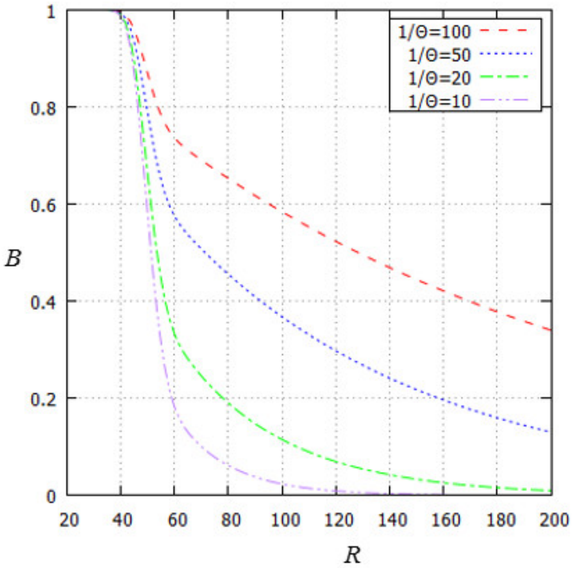
Figures 4: Mean waiting time on varying mean vacation time for different value of threshold



Figures 5: Blocking probability on varying mean vacation time for different value of the threshold



Figures 6: Blocking probability on varying arrival rate for different value of the threshold



Figures 7: Blocking probability on varying buffer size for different value of mean vacation time

SUMMARY AND FUTHER STUDY

In this paper, we give an overview of the IMS conferencing service architecture and a basic call flow. The queuing model with batch arrival batch exhaustive service and working vacations was analyzed.

The effectiveness of the server depends on external factors such as arrival traffic rate and on server policy configuration (value of θ, H) as well. For further study we propose the following optimization problem that will allow to increase IMS server performance

$$F = \begin{cases} \dfrac{\lambda(1 - B(\theta, H))}{L\mu} \to \max, \\ B(\theta, H) \le 10^{-3}, \\ W(\theta, H) \le 10. \end{cases}$$

Notes and Comments. This work was supported in part by the Russian Foundation for Basic Research (grant 15-07-03608).

REFERENCES

Camarillo, G., Garcia-Martin, M. A. 2008. The 3G IP Multimedia Subsystem (IMS): Merging the Internet and the Cellular Worlds, 3rd ed.Wiley.

Rosenberg, J., Schulzrinne, H., Camarillo, G. et al. 2002. SIP: Session Initiation Protocol. RFC 3261.

IP Multimedia Subsystem (IMS), 2004. 3GPP, Stage 2 (Release 5). 3GPP TS 23.228, vol. 5.

IP multimedia call control protocol based on Session Initiation Protocol (SIP) and Session Description Protocol (SDP). 2015. Stage 3 (Release 13), 3GPP TS 24.229.

Tien, A. Le, Nguyen H. 2010. Centralized and distributed architectures of scalable video conferencing services. In the Second International Conference on Ubiquitous and Future Networks, Jeju Island, Korea, pp. 394-399.

Pavel O. Abaev, Yuliya V. Gaidamaka, Konstantin E. Samouylov, Sergey Ya. Shorgin. 2013. Design and Software Architecture of SIP Server for Overload Control Simulation. Proceedings of the 27th European Conference on Modelling and Simulation, ECMS 2013, pp. 580-586.

Pavel Abaev, Yuliya Gaidamaka, and Konstantin E. Samouylov. 2012. Modeling of Hysteretic Signaling Load Control in Next Generation Networks. Lecture Notes in Computer Science. Germany, Heidelberg, Springer-Verlag, 2012, vol. 7469, pp. 440-452.

Mishra G., Dharmaraja S., Kar S. 2014. Performance Analysis of Multi-party Conferencing in IMS using Vacation Queues. IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS), pp. 1-6.

Rosenberg, J. 2008. Requirements for Management of Overload in the Session Initiation Protocol. RFC 5390.

AUTHOR BIOGRAPHIES

PAVEL ABAEV received his Ph.D. in Computer Science from the Peoples' Friendship University of Russia in 2011. He is an Assistant Professor in the Department of Applied Probability and Informatics at Peoples' Friendship University of Russia since 2013. His current research focus is on SDN/NFV, performance analysis of wireless 4G/5G networks and M2M communications, applied probability and queuing theory, and mathematical modeling of communication networks. His email address is: pabaev@sci.pfu.edu.ru.

VITALY BESCHASTNY received a BSc. degree in Applied Mathematics and Informatics in 2014 from People’s Friendship University, Moscow, Russia. Currently he enrolled in MSc program of the Department of Applied Probability and Informatics at the same university. His present research focuses on performance analysis of SDN/NFV, resource management in D2D-enabled cellular networks, mathematical modeling and performance analysis of computer and communication systems. His email address is: vbeschastny@sci.pfu.edu.ru.

ALEXEY TSAREV is currently enrolled in BSc. program of Department of Applied Mathematics and Informatics at People’s Friendship University, Moscow, Russia. His present research focuses on performance analysis of 5G networks, resource management in D2D-enabled cellular networks, mathematical modeling and performance analysis of computer and communication systems. His email address is: atsarev@sci.pfu.edu.ru.