

Impact of Signalling Load on Response Times for Signalling over IMS Core

Lubos Nagy, Jiri Hosek, Pavel Vajsar and Vit Novotny
Faculty of Electrical Engineering and Communication
Brno University of Technology
Technicka 12, 616 00 Brno, Czech Republic
Email: lubos.nagy@phd.feec.vutbr.cz

Abstract—This article focuses on the performance evaluation of the response time for signalling through a home Internet Protocol based Multimedia Subsystem (IMS), separately for each of IMS core nodes (Proxy-Call Session Control Function, Interrogating-CSCF, Serving-CSCF and Home Subscriber Server) and then on the investigation of the trend-line functions and their equations to describe these delays for various measured intensity of signalling generated load by high-performance tool – IxLoad. In this article, we have found out the trend-line function of response times for each measured message. Thanks to the showed results, some performance parameters like delay in selected IMS core node and their behaviour can be predicted and evaluated.

Keywords—DIAMETER, IP based Multimedia Subsystem, IxLoad, Response time, SIP.

I. INTRODUCTION

THE current trends in telecommunications lead to the network convergence and to effort the greatest number of telecommunication services through one type of transport networks and for one multifunction terminals. Therefore in the past, the operators and vendors were looking for an IP-based connectivity concept allowing the convergence network technologies and opportunities for optimization at all levels of designed communication system. Nowadays, the IMS (IP Multimedia Subsystem) in role of the IP-based service control architecture represents the standard of fixed-mobile network convergence. However to this optimization, it is necessary to know exactly the behaviour of these systems for various real conditions. One of the possible ways how to determine the behaviour of the whole IMS subsystem or only some IMS nodes is the performance analysis either based on the mathematical modelling using queueing theory or performance benchmarking (see *Section II* of this article).

The methodology of IMS/NGN (Next Generation Networks) performance benchmarking is standardized by European Telecommunications Standards Institute (ETSI) in multi-part deliverable that is divided into four separate parts [1]: *Core Concepts, Subsystem Configurations and Benchmarks, Traffic Sets and Traffic Profiles, and Reference Load network quality parameters*. The overall concept of IMS test-beds including the IMS benchmark information model, test parameters and benchmark metrics examples is defined in the first part of this technical standard. The SUT (*System Under Test*) configuration parameters, use-cases and scenarios with metrics and design objectives are presented in the second specification. In the third part, the traffic set, traffic-time profile and test procedures are

defined. The reference load network quality parameters for use-cases defined in the second part of [1] are presented in the last part of this specification.

II. RELATED WORK

There are various research papers, documents or studies that describe the performance analysis of whole IMS. The OpenIMS Core project (in role of SUT) and IMS Bench SIPp project (in role of TS - *Test System*) are often used tools for the performance analysis of IMS subsystem. The related work concerning the performance evaluation of IMS networks can be divided into three main categories, the performance evaluation of maximum load [2]–[4], the performance evaluation of delays of SIP (Session Initiation Protocol) signalling [5] and evaluation of delays of IMS procedures like IMS registration or IMS session setup procedures [6]–[7].

In our previous works, we were mainly focused on the performance analysis using the IMS queueing network model [8] and on the performance evaluation of maximum load signalling over our laboratory IMS network [9] according to specification [1]. In [9], we investigated that the value of the maximum signalling load for these hardware and software configurations is 500cps for defined IHS threshold 0.025% and the HSS entity was the failure point of simulated IMS network. The same bottleneck was described in [3] for even lower values of load (during execution of registration procedures). The similar test-beds are described by others researchers in [2]–[4] and the results of maximum loads correspond to the results measured in our test-bed which is described in [9] and in this section. In [8], we presented the design of IMS mathematical model based on separated M/M/1 queueing system with feedbacks that consists of the same IMS entities, signalling and services as our laboratory IMS network. In this M/M/1 model, the new load balancing methods, that can be used for a selection of S-CSCF server during the registration procedures of subscribers, were designed and evaluated. The obtained results showed that the service latency of the whole IMS core subsystem can be optimized with the help of implemented methods into mathematical network model based on M/M/1 queueing system. However, the service times are not exponentially distributed in the real networks. Therefore, the *main motivation of this article* is targeted at the measurement of delays for each SIP and DIAMETER signalling of IMS core elements using the performance analysis of IMS core elements and standards [1]. Thanks to the obtained results, we will be able to simulate the behaviour of IMS nodes using M/G/1 queueing systems and

evaluate the designed methods for load balancing under more realistic network conditions.

III. IMS TEST-BED

The experimental topology of the test-bed (see Fig. 1) consisted of IMS core subsystem, VoD Application Server and Media Streaming Server with the same following hardware and software configurations: 4x Intel Core i5-2400S CPU @2.50 GHz with 6144k L3 cache, 8 GM RAM (DIMM 1333 MHz), 82574L Intel Gigabit Network Connection, OS GNU/Linux (Debian distribution, AMD64 architecture, kernel v3.2), the software implementations of CSCF nodes are based on the SER (SIP Express Router) and the HSS based on FHoSS (FOKUS HSS) server created by Fraunhofer FOKUS Institute.

The SIP load signalling of selected multimedia services (Video on Demand, Voice over IP and File transfer) with defined intensity (see λ in Fig. 1) of the Poisson arrival process is generated with the help of the high-performance IxLoad application (see TS in Fig. 1). Each of services consists of three phases: registration procedure with subscription, session establishment and termination procedures (only for registered subscribers), and de-registration procedure. The registration phase consists of the registrar and subscription transactions. The generated signalling flows are created with the help of the standardized document 3GPP TS 24.228.

We can define the test-bed architecture with the help of queuing theory that is often used to evaluate the performance parameters of whole networks or only some network nodes. In the Fig. 1, the IMS core nodes are shown as the M/G/1 queuing system. One of the most important performance parameters is the response time of system (see eq. (1)). The mean value of this parameter can be calculated using the Pollaczek-Khinchine formula (known as P-K mean value formula [10]):

$$\frac{T}{\bar{x}} = 1 + \rho * \frac{1 + C_b^2}{2 * (1 - \rho)} \quad (1)$$

Where the $\rho = \frac{\lambda}{\mu}$, C_b^2 is the coefficient of service time variation and the parameter \bar{x} is the mean service time.

The way to calculate these values is following. In the case of P-CSCF server, the response time is always the time difference between the received and forwarded SIP messages. The response time of SIP signalling through S-CSCF server equals to the signalling through P-CSCF except the registration and de-registration procedures. In the case of these procedures, the response times of signalling through S-CSCF server are

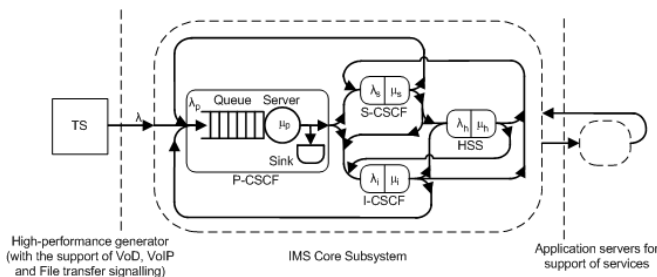


Fig. 1. The test-bed architecture as the queuing system network with feedbacks.

determined as the transactions between SIP and DIAMETER (DIAMETER uses TCP as its transport protocol). It means that the service times are determined as differences between the SIP request received from I-CSCF and DIAMETER request sent to HSS (sending time of DIAMETER MAR - receiving time of the first SIP REGISTER) or the DIAMETER answer received from HSS and SIP response sent to I-CSCF (sending time of SIP 401 - receiving time of DIAMETER MAA). The same way to determine the signalling response times is used for I-CSCF server. In the case of times for DIAMETER signalling through HSS database, the value is calculated as difference between the times of received DIAMETER request and sent DIAMETER answer.

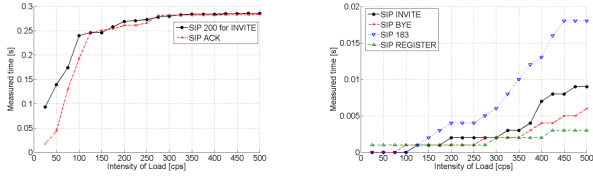
IV. RESULTS AND ANALYSIS

The traffic of three advanced telecommunication services (Video on Demand, Voice over IP and File transfer using SIP/RTSP/MSRP and RTP/RTCP signalling) over IMS experimental network is evaluated for various load intensities separately (from 25cps to 500cps) for each of IMS core nodes (the P-CSCF, I-CSCF, S-CSCF and HSS) and for each of SIP or DIAMETER messages. The most important results are shown in Fig. 2 to Fig. 3(b) and Tab. I to Tab. IV. The maximum value of signalling load (500cps) for used hardware and software configurations was investigated in [9]. The SUT (whole IMS network) was very unstable for the load greater than 500cps. In Tab. I to Tab. IV, the trend-lines of response times for each message through selected IMS servers are shown only for messages with the measured service time greater than 1ms. This limitation, shown in all tables and figures, is the measurement accuracy. The measured messages are displayed in the first column of the shown tables, the formulas of trend-lines of the response time for defined range of signalling load are shown in the second column. The parameter x is the signalling load generated by IxLoad application in role of TS (see λ in Fig. 1). The delay calculation methodology has been described in the previous section of this article. The measured characteristics of the response times into arrival signalling load through SUT (see λ in Fig. 1) are shown in the Fig. 2-4.

The P-CSCF trend-lines of response times (see Tab. I and Fig. 2) are mostly defined with the exponential or logarithmic time complexity. Other measured SIP request or response times (SIP 180, SIP 200s for REGISTER, BYE, SUBSCRIBE, UPDATE and PRACK, SIP 401, PRACK and UPDATE) are set to 1ms (the measured times are less than 1ms). In the graphs (see Fig. 2), the mean values of measured times within the generated load (see λ in Fig.1) for SIP messages with response times greater than 1ms are shown. From these graphs, it can be seen that the ACK and 200 for INVITE messages have the

TABLE I. THE FUNCTIONS OF RESPONSE TIMES FOR EACH MESSAGE THROUGH THE P-CSCF

SIP requests and responses	The trend-line function
SIP Session in Progress	$f(x) = (0.0009) * exp(0.0067 * x)$
SIP OK for INVITE	$f(x) = (-2.346) * x^{(-0.7343)} + 0.3137$
SIP REGISTER	$f(x) = (5.171e - 11) * x^{(2.838)} + 0.0009$
SIP INVITE	$f(x) = (0.0003) * exp(0.0075 * x)$
SIP ACK	$f(x) = (-3.804) * x^{(-0.7923)} + 0.3139$
SIP BYE	$f(x) = (0.0004) * exp(0.0055 * x)$



(a) The messages with higher measured times. (b) The messages with lower measured times.

Fig. 2. The mean values of response times vs. signalling load through the P-CSCF.

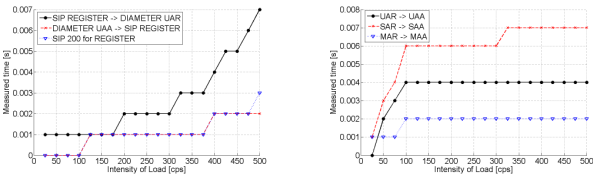
highest response times. However, the SIP Session in Progress and SIP INVITE messages have the greatest increase of the response time within the analysed interval.

The I-CSCF server is the next evaluated IMS node (see Tab. II and Fig. 3(a)). In our test-bed of a home IMS network, this server is active only during the registration or de-registration procedures. Only the SIP 401 response has different time complexity (it has the constant complexity, the measured times are less than $1ms$) than other SIP messages. The SIP REGISTER \rightarrow DIAMETER UAR processing has the shortest rise of measured response times (see Fig. 3(a)).

The last node, which was evaluated within the CSCF core, is the S-CSCF server (see Tab. III). This IMS node presents the central node of the whole IMS network and therefore it can be expected that this node has the greatest response times (see Fig. 4). Actually, there are two interesting facts in this obtained results. First, the highest values of response time are associated with the SIP responses (SIP 200 for REGISTER or SIP 401) created by S-CSCF node during the registration or de-registration procedures when the DIAMETER answers (MAA or SAA) are received from the HSS node. The second interesting result is that the SIP ACK and SIP 200 for INVITE

TABLE II. THE FUNCTIONS OF SERVICE TIMES FOR MESSAGES THROUGH THE I-CSCF

SIP and DIAMETER requests and responses	The trend-line function
SIP 200 for REGISTER	$f(x) = \begin{cases} < 1ms & \text{if } x < 100cps \\ 1ms & \text{if } x \in (100, 375)cps \\ 2ms & \text{if } x > 375cps \end{cases}$
SIP REGISTER \rightarrow DIAMETER UAR	$f(x) = (-6.146e - 11) * x^{(2.956)} + 0.0009$
DIAMETER UAA \rightarrow SIP REGISTER	$f(x) = \begin{cases} < 1ms & \text{if } x < 125cps \\ 1ms & \text{if } x \in (125, 375)cps \\ 2ms & \text{if } x > 375cps \end{cases}$

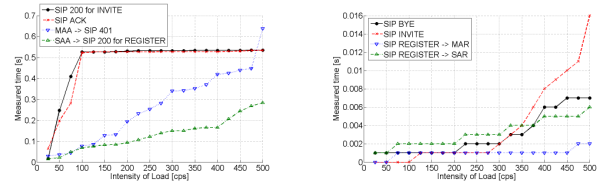


(a) The response times of messages through the I-CSCF. (b) The response times of messages through the HSS.

Fig. 3. The mean values of response times vs. signalling load through the I-CSCF (on the left) and through the HSS (on the right).

TABLE III. THE FUNCTIONS OF RESPONSE TIMES FOR MESSAGES THROUGH THE S-CSCF

SIP and DIAMETER requests and responses	The trend-line function
SIP REGISTER \rightarrow DIAMETER MAR	$f(x) = \begin{cases} < 1ms & \text{if } x < 75cps \\ 1ms & \text{if } x \in (75, 450)cps \\ 2ms & \text{if } x > 450cps \end{cases}$
DIAMETER MAA \rightarrow SIP 401	$f(x) = (0.0016) * x^{(0.936)} - 0.0341$
SIP REGISTER \rightarrow DIAMETER SAR	$f(x) = (0.0013) * exp(0.0032 * x)$
DIAMETER SAA \rightarrow SIP 200 for REGISTER	$f(x) = (0.0042) * exp(0.0038 * x)$
SIP INVITE	$f(x) = (4.528e - 12) * x^{(3.528)}$
SIP ACK	$f(x) = (-115.9) * x^{(-1.711)} - 0.5348$
SIP BYE	$f(x) = (-1.145e - 10) * x^{(2.885)} - 0.0009$
SIP 200 for INVITE	$f(x) = (-277.9) * x^{(-1.952)} + 0.536$



(a) The messages with higher measured times. (b) The messages with lower measured times.

Fig. 4. The mean values of response times vs. signalling load through the S-CSCF.

messages have the highest response time of all SIP messages forwarded by this server. The similar result was measured also in the case of the forwarding this SIP message by P-CSCF node. The measured values of other request or response times (for SIP 180, 183, SIP 200s for BYE, UPDATE and PRACK, SIP PRACK and UPDATE) are not showed in Tab. III because the measured values are less than $1ms$.

The last measured node of IMS network is the HSS database (see Tab. IV or Fig. 3(b)). In our test-bed, the database server is active only during registration and de-registration procedures. It can be seen that the response time of all measured DIAMETER requests/answers has the logarithmic time complexity with relatively low difference between the value of minimum load and the value of maximum load.

In our case, the tested IMS core subsystem is in the role of a home IMS network. The signalling goes through each of evaluated IMS core server (see Fig.1) only the case of the de/registration procedures. The delay of IMS core, which is calculated as formula (2), consist of IMS core element delays and transport delay through network infrastructure.

$$D = \sum D_P + \sum D_I + \sum D_H + \sum D_S + \sum D_T \quad (2)$$

TABLE IV. THE FUNCTIONS OF RESPONSE TIMES FOR MESSAGES THROUGH THE HSS

DIAMETER Command-Codes	The trend-line function
300 (UA{R, A})	$f(x) = (-1.293) * x^{(-1.792)} + 0.004$
301 (SA{R, A})	$f(x) = (-0.046) * x^{(-0.569)} + 0.008$
303 (MA{R, A})	$f(x) = \begin{cases} 1ms & \text{if } x < 100cps \\ 2ms & \text{if } x \geq 100cps \end{cases}$

Where $\sum D_P$, $\sum D_I$, $\sum D_S$ and $\sum D_H$ are the investigated times that the messages spent in CSCFs and HSS and $\sum D_T$ is the time the messages spend within the network infrastructure. Each of core node delays is composited from the queuing and processing delays defined in [7]. The values of $\sum D_I$ and $\sum D_H$ are greater than zero if the signalling is from the registration or de-registration procedures, else the values are equal to zero. The theorem is valid for the home IMS network simulated in this paper.

The successful registration procedure (see eq. (3)) is influenced by three delay parts, thereof two delays are influenced by time the signalling spent in SUT (the tested IMS core) and the response times of TS (IxLoad application).

$$D_{REG} = \underbrace{\sum D_{(REG1 \rightarrow 401)}}_{SUT} + \underbrace{\sum D_{(401 \rightarrow REG2)}}_{TS} + \underbrace{\sum D_{(REG2 \rightarrow 200)}}_{SUT} \quad (3)$$

We do not tie the effect of $\sum D_{(401 \rightarrow REG2)}$ and transmission delay in the following equations. The first of SUT delay is shown in eq. (4). We can define the second one based on assumptions from the first SUT delay.

$$\begin{aligned} \sum D_{(REG1 \rightarrow 401)} &= \underbrace{D_{(REG1 \rightarrow REG1)} + D_{(401 \rightarrow 401)}}_P + \\ &+ \underbrace{D_{(REG1 \rightarrow UAR)} + D_{(UAA \rightarrow REG1)} + D_{(401 \rightarrow 401)}}_I + \\ &+ \underbrace{D_{(UAR \rightarrow UAA)} + D_{(MAR \rightarrow MAA)}}_H + \\ &+ \underbrace{D_{(REG1 \rightarrow MAR)} + D_{(MAA \rightarrow 401)}}_S \end{aligned} \quad (4)$$

If we neglect the effects of lower signalling delays and the impact of delays outside IMS core elements (see eq. (2)–(4)) then we can define for conditions of our test-bed the delay of successful registration procedures as:

$$D_{REG} \approx \underbrace{D_{(MAA \rightarrow 401)}}_{\text{from first SUT delay}} + \underbrace{D_{(SAA \rightarrow 200)}}_{\text{from second SUT delay}} \quad (5)$$

The percentage ratio of derived D_{REG} (see eq. (5)) is 94.4% of measured D_{REG} (see Fig. 2–Fig. 4). From the characteristics and eq. (5), it can be seen that the delay of IMS procedures is mainly influenced by measured delays of S-CSCF server that is in role of IMS networks as IMS central core element.

V. CONCLUSION

This paper deals with the evaluation of response times for signalling through the experimental IMS core subsystem, separately for each IMS core node and message, and for various values of network load. Three advanced telecommunication services were generated by the high-performance IxLoad application. All selected IMS core nodes were situated in the servers with the same hardware configurations.

From the showed characteristics (see Fig. 2–Fig. 4), it can be seen that the central node of the whole IMS network (the S-CSCF server) has the highest values of response time and its

influence on delays of signalling through whole home IMS network from eq. (5). Based on assumption from eq. (2)–(5), we can obtain the similar results for other tested IMS procedures like session establishment and that the S-CSCF server has the highest impact on delay of signalling within a home IMS network. However, the influence of S-CSCF server is not very high in the case of session termination procedure. Therefore, our future work will focus on the problem how to optimize the latency of the whole IMS network e.g. during registration procedures using load-balancing of S-CSCF servers.

Also, we have found out the trend-lines with the correlation (the lowest R-squared index was approximately 0.95, the most commonly value of R-squared was 0.98) that are described by the help of the exponential or logarithmic functions for each evaluated message and IMS core node. In the case of HSS node, only logarithmic function is used to define trend-lines of DIAMETER signalling. This functions could be used to predict the delay either in the node of IMS network or within the whole IMS network.

ACKNOWLEDGMENT

This research work is funded by projects SIX CZ.1.05/2.1.00/03.0072, CZ.1.07/2.3.00/30.0005, EU ECOP EE.2.3.20.0094 and CZ.1.07/2.2.00/28.0062.

REFERENCES

- [1] ETSI European Telecommunications Standards Institute, IMS Network Testing (INT): IMS/NGN Performance Benchmark. ETSI TS 186 008. November 2012.
- [2] G. Din, R. Petre, I. Schieferdecker, "A Workload Model for Benchmarking IMS Core Networks," in *IEEE Global Telecommunications Conference GLOBECOM 2007*, 26.–30. Nov. 2007, pp. 2623–2627. ISBN: 978-1-4244-1043-9.
- [3] R. Herpertz, J. M. E. Carlin, "A Performance Benchmark of a Multimedia Service Delivery Framework," in *IEEE Mexican International Conference on Computer Science ENC 2009*, 21.–25. Sept. 2009, pp. 137–141. ISBN: 978-1-4244-5258-3.
- [4] D. Thissen, J. M. E. Carlin, R. Herpertz, "Evaluating the Performance of an IMS/NGN Deployment," in *Proceedings of the 2nd Workshop on Services, Platforms, Innovations and Research for new Infrastructures in Telecommunications. SPIRIT 2009*, Germany, Oct. 2009.
- [5] S. Pandey, V. Jain, D. Das, V. Planat, R. Periannan, "Performance Study of IMS Signaling Plane," in *IEEE International Conference on IP Multimedia Subsystem Architecture and Applications - 2007*, 2007, pp. 1–5. ISBN: 978-1-4244-2671-3.
- [6] Y. He, J. Veerkamp, A. Bilgic, A. Bilgic, "Analyzing the Internal Processing of IMS-based and traditional VoIP systems" in *IEEE International Conference on Telecommunications: The Infrastructure for the 21st Century (WTC), 2010*, Sept. 2010, pp. 1–6. ISBN: 978-3-8007-3303-3.
- [7] A. Munir, A. Gordon-Ross, "SIP-Based IMS Signaling Analysis for WiMax-3G Interworking Architectures" in *IEEE Transactions on Mobile Computing*, May 2010. Volume 9, Issue 5. pp. 733-750. ISSN: 1536-1233.
- [8] L. Nagy, J. Tombal, V. Novotny, "Proposal of a Queueing Model for Simulation of Advanced Telecommunication Services over IMS Architecture," in *IEEE International Conference on Telecommunications and Signal Processing - 2013*, 2013, ISBN: 978-1-4799-0402-0.
- [9] L. Nagy, R. Krkos, V. Novotny, "Performance Analysis of IMS Network," in *the 14th International Conference on Research in Telecommunication Technologies - RTT 2012*, 2012, pp. 55.–60. ISBN: 978-80-554-0570-4.
- [10] L. Kleinrock, "Queueing Systems, Vol. 1: Theory" Ed. New York: Wiley Interscience, 1975. 417 pages. ISBN: 0-471-49110-1.