PERFORMANCE ANALYSIS OF IMS NETWORK: THE PROPOSAL OF NEW ALGORITHMS FOR S-CSCF ASSIGNMENT

Lubos NAGY¹, Vit NOVOTNY¹, Jana URAMOVA², Nermin MAKHLOUF¹

¹Department of Telecommunications, Faculty of Electrical Engineering and Communication, Brno University of Technology, Technicka 12, 616 00 Brno, Czech Republic

²Department of Information Networks, Faculty of Management Science and Informatics, University of Zilina, Universitna 8215/1, 010 26 Zilina, Slovak Republic

lubos.nagy@phd.feec.vutbr.cz, novotnyv@feec.vutbr.cz, jana.uramova@fri.uniza.sk, nermin.makhlouf@fri.uniza.sk

Abstract. This article is focused on the proposal of three load balancing methods which can be used for a selection of S-CSCF (Serving-Call Session Control Server) server in IP Multimedia Subsystem (IMS) during the registration procedures of subscribers. All presented methods are implemented and evaluated for various inter-arrival and service times in the mathematical model based on queueing theory. In this article, two methods based on performance parameters (such as utilizations, etc.) and one method based on number of registered subscribers to each of available S-CSCF server are described. The main advantage of third method is that all related information is obtained from traffic analysis through I-CSCF (Interrogating-CSCF) node. Also, the designed methods are compared with other selection algorithms presented in previous research works by others researchers (Hwang et col., Cho et col. or Tirana et col.). The article shows that the implemented methods can optimize the service latency of whole IMS network.

Keywords

File transfer, IP based Multimedia Subsystem, load balancer, S-CSCF selection, service latency, Video on Demand, Voice over IP.

1. Introduction

The current trends in telecommunication lead to the convergence of mobile, wireless and fixed network technologies and the integration of wide variety of services with particular quality of service requirements through common transport network and using a multifunction terminal. At present, the IP Multimedia Subsystem in role of the service control architecture represents a unifying solution for convergence of these technologies. One of the most important QoS-related network parameters whose values can be affected by unbalanced signalling traffic over core network is the service latency. In general, it is necessary to know the behaviour of whole unbalanced systems to perform optimization process. One of possible ways to determine the behaviour of a network is the performance benchmarking or various mathematical tools like a queueing theory.

The main motivation of this article is the proposal of new three algorithms for S-CSCF assignment [1] based on the analysis using queueing theory. One of causes of increased service latency over whole IMS network can be unbalanced signalling traffic over IMS core, mainly over the S-CSCF server as the central service provision node of whole IMS network. One possibility of load balancing is defined in technical specification [2]. The influence of unbalanced traffic on the service latency was written in [3], see second section of this article.

In this article, we will describe new selection algorithms of S-CSCF server to reduce the service latency causing by unbalancing traffic load of IMS core elements. These new algorithms do not use only the performance parameters like server utilization but also the parameter like the number of subscribers registered to S-CSCF servers.

2. Related Work

According to the recommendations defined in the technical specification by 3GPP, the I-CSCF performs the S-CSCF assignment only in three cases and using six main types of parameters (required capabilities for user services, operator preferences on a per-user basic, capabilities of individual S-CSCFs in the home network, topological information of where a subscriber is located and availability of S-CSCF servers). In the first case, the S-CSCF is selected during registration, then during execution services for unregistered subscribers or when the selected S-CSCF server is not responding. All needed information to S-CSCF assignment is transferred using Cx interface within grouped AVPs, and so within the DIAMETER Server-Capabilities AVP [1], [2], [4]. In the case that all transferred information from all available S-CSCF servers are equal, the I-CSCF server performs the selection using the best-fit function.

The proposals of effective algorithms or implementations of management system used to the S-CSCF selection are described in various research studies or papers ([5], [6], [7], [8], [9], [10], [11], [12], [13], [15]). In [5],Oh et at. proposed the data structure for S-CSCF selection (S-CSCF capability and load-balancing information) and three-step selection mechanism performed by I-CSCF. In document [6], the proposed method of S-CSCF assignment is based on periodical update mechanism of load related information of the S-CSCFs using SIP OPTIONS method. In [7], the SIP NOTIFY method is used to transfer the information (the number of registered subscribers, resources usage such as memory, CPU and queueing delays) in load-balancing mechanism designed by authors Abdalla et col. In [8], the authors proposed the re-associated schema of overloaded S-CSCFs based on the server utilization calculation for every 10,000 request arrivals received by each of simulated S-CSCFs.

In next research works [9], [10], [11], [12], [13], a design of a dynamic routing algorithm for managing CSCF servers and its implementation is described by authors. The presented algorithm is based on five criteria with different priorities (system down or up with the highest priority, defined threshold of CPU and memory utilizations, throughput and response time with the lowest priority). However, the evaluation of described algorithm is not shown in [9], [10], [11], [12], [13]. Therefore, this algorithm is one of evaluated algorithms presented in this paper.

Four algorithms (based on uniform random allocation, round robin, lowest response time and shortest expected delay) used to distributed S-CSCF selection were proposed and evaluated for different grids in the next research studies published in [14], [15]. First two of them are well-known algorithms. In the case of *uniform random allocation*, the S-CSCF server was selected from the list of available servers randomly and in the case of *round robin algorithm*, the next available S-CSCF server was always chosen. Third method (*lowest response time*) was based on latency between the P-CSCF and S-CSCF servers measured periodically by load balancer (located in the I-CSCF) using own updated protocol. In the case of the *shortest expected delay* algorithm ([15]), the selected node was chosen with the help of local information (such as queue size) without coordinating with other nodes. In [15], two scenarios were evaluated, in the first one the request message size was one unit and in the second case the message size was greater. The obtained results from first scenario showed that the method of *shortest expected delay* is the best selection algorithm. For second scenario, the results of *shortest expected delay* and *round robin* algorithms were very similar and both were better than *uniform random allocation* and *lowest response time* methods.

In the article [3] we presented the design of home IMS network model based on queueing theory. The IMS network (two Proxy-CSCF, five S-CSCFs, one I-CSCF, one Home Subscriber Server, one Application server and one Media Streaming Server) was described as a single queueing network with feedbacks. Also, we presented the support of three advanced telecommunication services (VoIP - Voice over Internet Protocol, Video on Demand and file transfer using SIP and Diameter for session control messages, and RTP/RTCP/RTSP and MSRP protocols for media transmission and its control) over designed mathematical model in more detail. All simulations presented in [3] were focused on the evaluation and influence of unbalanced signalling load over IMS core, mainly over five S-CSCF nodes (the effect of S-CSCFs on service latency was evaluated using the performance test-bed in [16]) during the registration procedures, on service latency of whole network. For purpose of load balancing, three algorithms of S-CSCF assignment were in more detail designed, described and evaluated. The obtained results of [3] showed that the algorithm based on the combination of two performance parameters, the server utilization of S-CSCF nodes and the number of waiting messages in the FIFO queue of S-CSCF nodes, is better than next two algorithms based on each of selected parameters separately. The algorithm with the best results is used in this article for comparison both with our new algorithms and with algorithms described in the papers of other scientists.

As has been written above, the influence of S-CSCF on service latency during various IMS procedures was evaluated in [16]. In the article, we have neglected the effects of lower signalling delays and the impact of delays outside IMS core elements. The derived service latency of successful registration procedures (D_{REG}) is showed in the Eq. (1)

$$D_{REG} \approx D_{(MAA \to 401)} + D_{(SAA \to 200)}, \qquad (1)$$

where $D_{(MAA \rightarrow 401)}$ is the delay of the DIAMETER MAA \rightarrow SIP 401 processing during the first phases

of registration procedure and $D_{(SAA\to 200)}$ is the delay of the DIAMETER SAA \rightarrow SIP 200 for SIP REGIS-TER processing during the second phases of registration procedure. The percentage ratio of derived D_{REG} is 94,4 % of measured D_{REG} . It can be seen (see the Eq. (1)) that the S-CSCF has the highest impact on delay of signalling within a home IMS network. We obtained the very similar results for session establishment procedures in [16].

Therefore, this article is focused on the problem how to optimize the latency of whole IMS network using load-balancing mechanism of S-CSCF servers.

3. Design of New Algorithms

The designed model of IMS network based on the GI/M/1 queueing systems (with a single server and FIFO queue with infinite capacity) is shown in the Fig. 1. The used queueing model (GI/M/1) is based on generally distributed inter-arrival times and exponentially distributed service times [17]. The most common queueing systems are M/M/1, M/GI/1 or GI/M/1. The designed network model can be defined as an open single class queueing network with feedbacks.

The IMS model (see the Fig. 1) can be divided from the point of view of horizontal layered architecture into four layers:

- end-device layer with load generator for support of VoD, VoIP and File transfer over IMS signalling,
- *transport layer* presents intermediate transport network technologies and routers,
- control layer presents the IMS core elements (an I-CSCF, an HSS, two P-CSCFs and three S-CSCFs),
- *application layer* presents the application (AS) and media streaming servers (MSS) for support of VoD service.

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 [18]. The proposal of queueing network model using Matlab environment was described in more detail in our previous work [3].

3.1. Unbalanced S-CSCFs

The measured latencies of whole designed queueing network for different forms of random user request distributions among S-CSCF servers are shown in Fig. 2. In this case, the service times of all CSCFs, HSS, AS and MSS nodes are set to 0,075 - 0,055 time units (the nodes have the same values for each simulation); others nodes are set to constant values of 0,00001 time units.

Two following random distributions of load were used to simulations of conditions in IMS network during slightly and highly unbalanced distributions of requests among S-CSCF servers:

- defined load of S-CSCFs (no. 1, no. 2, no. 3): [33,33 %, 33,33 %, 33,33 %] (see Fig. 2(a)),
- defined load of S-CSCFs (no. 1, no. 2, no. 3): [95 %, 2,5 %, 2,5 %] (see Fig. 2(b)).



Fig. 2: Service latencies vs. request inter-arrival times for various service times.

It can be seen that the influence of highly unbalanced distributions of requests among S-CSCF on latency of whole IMS subsystem is significantly large. In the case of first S-CSCF load ([33,33 %, 33,33 %, 33,33 %]), the latency of whole queueing system did not have steady state for shorter values of inter-arrival times than value of 0, 1 *time units* and for slow single servers (greater



Fig. 1: Queueing network for simulation of latency in IMS network [3].

service times of request) of IMS nodes (CSCFs, HSS, AS and MSS) than 0,075 time units.

In second scenario ([95 %, 2,5 %, 2,5 %]), it can be seen that the obtained results of service latencies of whole network are higher than the values from the first scenario. The influence of unbalanced load of available S-CSCF servers is showed in Fig. 2. In the next subsections, we designed and described the algorithms for balancing of S-CSCF node to minimize latencies of whole IMS subsystem.

3.2. Algorithm I

First proposed method of S-CSCF selection (see Fig. 3) is based on combination of two performance parameters (see Step 2), the number of waiting messages in FIFO queue and the server utilizations. Firstly, the queue occupation is considered and if there is only one server with lowest queue occupation (see Step 3) then this server is selected (see Step 4). Provided there are more servers with the same lowest level of queue occupation then two of these servers with the lowest ID are chosen for the second round and the server with lower value of utilization is selected (see Step 3b).

This way of selection was designed after evaluation of methods based only on the number of waiting messages in queues and the server utilizations [3].

3.3. Algorithm II

The actual number of waiting messages in queues and the service times of S-CSCF servers are used as next performance parameters to S-CSCF assignment (see Step 2 in Fig. 3). The I-CSCF node performs the selection of S-CSCF using waiting time calculation for incoming request (see PredictedTime - PT in Eq. (2), see Step 3 and Step 4).

PT = ActNum * ServiceTime [time units]. (2)

If there are more servers with the same results then the S-CSCF with the lowest priority ID is chosen (see *Step 3b* in Fig. 3).



Fig. 3: The simplified idea of proposed algorithm I and II.

3.4. Algorithm III

The third algorithm is based on the balancing load of S-CSCF servers using the actual number of registered subscribers to each of simulated S-CSCF servers. This value is in I-CSCF record always incremented at the end of successful registration procedure and decremented at the end of de-registration procedure. If there are more servers with the same number of subscribers then the server with the lowest ID is selected. Moreover, unlike previous methods, this proposed algorithm is independent from the periodic update of the values of performance parameters (server utilization, etc.).

The main benefit of this method, in our designed model of home IMS subsystem, is that any interface for transferring information of S-CSCF is not required because all related information is obtained from traffic analysis through I-CSCF node.

4. Simulation Results

The scenario with different performance conditions of S-CSCFs are simulated and each presented method for latency minimization of whole IMS network model are evaluated for various inter-arrival times (0,095 -0,14 time units, the inter-arrival times of generated SIP signalling by generator to IMS core) and service times. The service times of S-CSCF 1, S-CSCF 2, S-CSCF 3 are set to respectively as the following: 0.055 time units, 0.065 time units and 0,075 time units. The P/I-CSCFs, HSS, AS and MSS nodes are set to 0,065 time units; others nodes are set to values of 0,000001 time units). The length of simulations is set on 5000 time units. Six algorithms of S-CSCF assignment are evaluated and compared with unbalanced scenario. First three of evaluated algorithms (Algorithm 1, Algorithm 2, Algorithm 3, see subsections 3.2, 3.3 and 3.4) were in more detail described in the previous sections. Next three methods are the Lowest response time [14], [15], Hwang's algorithm [9], [10], [11], [12], [13] and also well-known selection method - the *Round robin* algorithm (see section 2) which is based on circular order (if the last selected S-CSCF of last registered subscriber was S-CSCF 1, then S-CSCF 2 will be selected for next subscriber during registration procedure). Also, we can divided these algorithms into two main categories:

- the algorithms based on performance parameters (server utilizations, etc.),
- the algorithms based on analysis of signalling traffic through I-CSCF server (*Algorithm III* and *Round robin*).

In the first case, the best algorithms is Algorithm II and in the second one, the Algorithm III is better mainly for longer inter-arrival times (under value of 0, 125 *time units*) than round robin algorithm. The round robin algorithm is more effective to minimized of latency for shortly inter-arrival times. Also, the obtained results (see Fig. 4 and Fig. 5 or Tab. 1) show that the Algorithm II is most effective to minimize the service latency of whole network.

Table 1 show the overall improvement values compared with the method without load balancing for all simulated inter-arrival times. As has been written

 Tab. 1: The obtained overall improvement for service latency of whole IMS subsystem.

Implemented method	Improvement [%]
Round Robin	$26,067 \pm 6,151$
Hwang's algorithm [9], [10], [11], [12], [13]	$18,433 \pm 11,6$
LRT with feedbacks [14], [15]	$18,574 \pm 13,244$
Algorithm I (see section 3.2)	$20,761 \pm 13,19$
Algorithm II (see section 3.3)	$29,515 \pm 7,709$
Algorithm III (see section 3.4)	$25,274 \pm 3,487$



Fig. 4: Characteristics of latencies vs. request inter-arrival times.

above (see section 2), in the case that all related information from available S-CSCFs are equal, the S-CSCF selection is based on the best-fit function. The most commonly best-fit function for S-CSCF assignment is just the Round robin algorithm. The average improvement value of algorithm II when compared with the round robin as best-fit function is 3,769 %.

5. Conclusion

Three types of S-CSCF selection algorithms were designed and then compared with algorithms presented in the papers of other researchers [9], [10], [11], [12], [13], [14], [15]. In the case of first two designed met-



Fig. 5: Characteristics of latencies vs. request inter-arrival times.

hods, the performance parameters (such as server utilizations, etc.) are used to choose the best of available S-CSCF servers. Third algorithm is based on analysis of signalling traffic through I-CSCF node.

In the graphs (see Fig. 4 and Fig. 5), the average values of service latencies within the inter-arrival rates (rate of generated messages by UEs within IMS network) are shown. It can be seen that the best load balancer methods for all values of inter-arrival and service times are the Algorithm II (based on prediction of waiting times for incoming requests) and Round Robin algorithms. The Algorithm II is the most effective to minimize service latency especially for shortly inter-arrival times. The service latency of whole designed network did not have steady state for shorter values of

inter-arrival times than value of 0,095 time units for settings of service times defined in section *Simulation Results.* This value of inter-arrival time is affected by the relatively high influence of feedbacks in the presented IMS model.

The obtained results from simulation with different performance conditions of S-CSCFs show that designed algorithms are optimizing the service latency compared with the standardized solution and with algorithms presented by others researchers (Hwang et col., Cho et col. or Tirana et col.). The second algorithm is the most effective method to minimize the service latency especially for shortly inter-arrival times. The overall minimum improvement value when compared with the round robin method is under 5 %.

The possible weakness of presented model are the queue with infinite capacity or the service rate of all simulated IMS nodes are exponential distributed. Therefore, our future research work will be focusing on the modelling of the home IMS network with the help of GI/GI/1/k queueing systems. The service times for each of SIP or DIAMETER signalling over IMS core will be defined with the help of performance evaluation of the home IMS network. Thanks to this queueing system, the designed model will more correspond with the real IMS network and real network conditions.

Acknowledgment

This work was supported by Ministry of Education, Youth and Sports of the Czech Republic under projects SIX CZ.1.05/2.1.00/03.0072 and MPO FR-TI4/696.

References

- 3rd Generation Partnership Project. Technical Specification Group Core Network and Terminals: IP Multimedia (IM) Subsystem Cx and Dx interfaces, Signaling flows and message contents (Release 11). TS 29.228 (v. 11.6.0), December 2012.
- [2] 3rd Generation Partnership Project. Technical Specification Group Services and System Aspects. IP Multimedia Subsystem (IMS) – Stage 2. Release 11. 3GPP TS 23.228 (v11.7.0). December 2012.
- [3] NAGY, L., J. TOMBAL and V. NOBOTNY. Proposal of a Queueing Model for Simulation of Advanced Telecommunication Services over IMS Architecture. In: International Conference on Telecommunications and Signal Processing. Rome: IEEE, 2013, pp. 326–330. ISBN 978-1-4799-0402-0. DOI: 10.1109/TSP.2013.6613945.

- [4] 3rd Generation Partnership Project. Technical Specification Group Core Network and Terminals: Cx and Dx interfaces based on the Diameter protocol, protocol details. Release 11. 3GPP TS 29.229 (v11.2.0). December 2012.
- [5] OH, S., Ch. LE and Y. SHIN. The Serving CSCF Assignment Algorithm in Wireless IP Multimedia Networks. In: 46th Midwest Symposium on Circuits and Systems. Cairo: IEEE, 2003, pp. 440– 445. ISBN 0-7803-8294-3. DOI: 10.1109/MWS-CAS.2003.1562313.
- [6] CHATTOPADHAYAY, B. and M. A. MUNOZ DE LA TORRE. S-CSCF load balancing. Motorola, Inc. 2006.
- [7] ABDALLA, I. and S. VENKATEASAN. Notification based S-CSCF load Balancing in IMS Networks. In: Wireless Telecommunications Symposium (WTS). New York City: IEEE, 2011, pp. 1–5. ISBN 978-1-4577-0162-7. DOI: 10.1109/WTS.2011.5960852.
- [8] XU, L., Ch. HUANG, J. YAN and T. DRW-IEGA. De-Registration Based S-CSCF Load Balancing in IMS Core Network. In: *International Conference on Communications*. Dresden: IEEE, 2009, pp. 1–5. ISBN 978-1-4244-3435-0. DOI: 10.1109/ICC.2009.5198895.
- [9] HWANG, J.-H., Y.-G. KOOK and J.-O. LEE. A Dynamic Routing Algorithm for Managing CSCF Agents in the IMS Network. In: 5th International Joint Conference on INC, IMS and IDC. Seoul: IEEE, 2009, pp. 930–935. ISBN 978-1-4244-5209-5. DOI: 10.1109/NCM.2009.120.
- [10] HWANG, J.-H. Routing Algorithm for Dynamic Management of CSCF Agent in the IMS Platform. In: KNOM Review - Network Operations and Management. 2008, vol. 11, iss. 2, pp. 14–27. ISSN 2287-1543.
- [11] HWANG, J.-H., J.-H. CHO and J.-O. LEE. A Routing Management among CSCFs Using Management Technology. In: 12th Asia-Pacific Network Operations and Management Conference on Management enabling the future internet for changing business and new computing services. Jeju: Springer Verlag, 2009, pp. 502–506. ISBN 978-3-642-04491-5. DOI: 10.1007/978-3-642-04492-2_63.
- [12] HWANG, J.-H. and J.-O. LEE. An Implementation of Network Management System Using Dynamic Routing in IMS. *International Journal of Software Engineering and Its Applications*. 2012, vol. 6, iss. 4, pp. 201–206. ISSN 1738-9984.

- [13] CHO, J.-H. and J.-O. LEE. A Management for the Deployment of Presence Service Using A Dynamic Routing Algorithm in the IMS nodes. In: 13th Asia-Pacific Network Operations and Management Symposium (APNOMS). Taipei: IEEE, 2011, pp. 1–4. ISBN 978-1-4577-1668-3. DOI: 10.1109/APNOMS.2011.6076957.
- [14] TIRANA, P. and D. MEDHI. The effects of load distribution algorithms in application's response time in the IMS architecture. In: *Proceeding of* 18th ITC Specialist Seminar on Quality of Experience. Karlskrona: Blekinge Institute of Technology, 2008, pp. 173–181. ISSN 1103-1581.
- [15] TIRANA, P. and D. MEDHI. Distributed Approaches to S-CSCF Selection in an IMS Network. In: Network Operations and Management Symposium. Osaka: IEEE, 2010. pp. 224–231. ISBN 978-1-4244-5366-5. DOI: 10.1109/NOMS.2010.5488465.
- [16] NAGY, L., J. HOSEK, P. VAJSAR and V. NOVOTNY. Impact of Signalling Load on Response Times for Signalling over IMS Core. In: Federated Conference on Computer Science and Information Systems. Krakow: IEEE, 2013, pp. 663–666. ISBN 978-83-60810-53-8.
- [17] KLEINROCK, L. Queuing Systems, Vol. 1: Theory. New York: Wiley Interscience, 1975. ISBN 0-471-49110-1.
- [18] 3rd Generation Partnership Project. Technical Specification Group Core Network and Terminals. Signalling flows for the IP multimedia call control based on Session Initiation Protocol (SIP) and Session Description Protocol (SDP); Stage 3. Release 5. 3GPP TS 24.228 (v5.15.0). September 2006.

About Authors

Lubos NAGY was born in Ilava (Slovak Republic) in 1984. He received his M.Sc. in 2009 at the Department of Telecommunications at the Faculty and Communication at Brno University of Technology in Czech Republic. At the present, he is working toward the Ph.D. degree at the Brno University of Technology. The area of his professional interest is the IP Multimedia Subsystem, mainly the performance evaluation of IMS core entities. He is teaching the laboratory exercises: "Mobile Network Communication Systems", "Network Architecture", "Terminal Equipment" at the same department.

Vit NOVOTNY (born in 1969) received M.Sc.

in 1992 at the Faculty of Electrical Engineering and Computer Science in the Brno University of Technology, in 2001 he received the Ph.D. degree and in 2005 he received Associate Professor degree at the same university. He received or participated in anumber of projects focused on mobile networks or on telecommunication network technologies and services. His professional interests include: fixed, wireless and mobile networks, telecommunication services, integrated services networks and Internet technologies.

Jana URAMOVA was born in Zvolen, Slovakia, in 1980. She graduated at Grammar School of L. Stura in Zvolen and she studied at the University of P. J. Safarik in Kosice. She defended dissertation thesis in 2005 and she has been continuing

her Ph.D. study and research of Information models of network throughput. She has been acting as a lecturer at the Department of information networks at University of Zilina from 2003. She is an instructor of Regional Cisco Network Academy in Zilina and she is lecturing the subjects exercises Communication Networks, Theory of Information, Introduction into Engineering and Computer Networks.

Nermin MAKHLOUF was born in Damascus (Syria Arab Republic). She received her Master's Degree at the department of Communication at Damascus University in 2009. At the present, she is studying the Ph.D. at the Brno University of Technology. Her research interests include Prediction of movement of wireless nodes in mobile ad-hoc networks MANETs and the Markov chains.