Problem of IMS modeling – Solving Approaches

Mesud Hadžialić, Mirko Škrbić, Nerma Šečić, Mirza Varatanović, Elvedina Zulić, Nedim Bijedić
University of Sarajevo
e-mail: mhadzialic@etf.unsa.ba, mskrbic@etf.unsa.ba, nsecic@etf.unsa.ba, mvaratanovic@etf.unsa.ba,
elvedina.zulic@bstelecom.ba, nedim.bijedic@bstelecom.ba

***Abstract* – IMS (IP Multimedia Subsystem) network has high demands from perspective of multimedia, flexible and interactive communications. Fulfillment of those demands, with appropriate levels of quality, is not a simple task. The primary objective of this paper is to set emphasis on the proper dimensioning of IMS network, and the need to find a methodology applicable in planning of IMS networks, that will be able to provide quantitative results on basis of the initial request. In addition, the goal is to provide insight to the basic problems of IMS network, and ways of solving them. This primarily refers to SIP (Session Initiation Protocol) servers, which are the key part of the IMS structure, which have to deal with overload due to improper dimensioning of the network. This paper will attempt to present systematically the existing solutions and to provide guidance for the future resolution of this problem.**

Keywords-IMS; modeling; overload; SIP.

I. INTRODUCTION

With the basic idea of integrating different networks into one multifunctional IP - based network, IMS (IP Multimedia Subsystem) aims to create a unified communication environment for fixed and mobile users, by offering enriched and integrated services that were fragmented before. IMS needs to offer a high level of interaction for users - enriched calling and enhanced messaging - sharing videos, images, and other multimedia during a voice call (during one communication session), with high level of service personalization. However, this level of communication flexibility significantly complicates the management, and among other things directly influences the increase of the number of signaling messages that must be exchanged and processed.

SIP (Session Initiation Protocol) represents the core of IMS architecture. SIP servers are the main components responsible for processing and routing all signaling messages. The problem that imposes itself is the problem of dimensioning IMS network, i.e., dimensioning IMS components, which should be able to process large amounts of messages, in order to avoid server overload and therefore degradation of QoS (Quality of Service) and QoE (Quality of Experience).

SIP provides limited built-in overload control mechanism - the 503 Service Unavailable response. However, as the price of rejecting *SIP* sessions typically cannot be ignored, this mechanism cannot prevent server's congestions. When a SIP server rejects a large amount of incoming sessions its performance degrades and, additionally, the impact of overload increases through the network - this is a key observation that distinguishes SIP server overload problem from other overload problems [8]. In order to eliminate or at least partially reduce the problems of SIP server overload, various approaches have been proposed in accordance with the opinion of what is the dominant problem that leads to uncontrolled overload.

On the other hand, if the network is not initially properly dimensioned, the application of congestion control will not be enough. In short time the nodes will fall into a state of congestion, which will lead to performance degradation of the entire network, and at some point, entire network will come to an outage and collapse. Thus, the network dimensioning must be understood as the primary problem, and a congestion control as an additional factor that can improve efficiency of well dimensioned network. Network modeling or a search for a loyal representation which reflects the behavior of IMS nodes and provides required output value in dependency on the input parameters, is a key of good methodologies for planning and dimensioning IMS network.

This paper systematically presents key problems of IMS and provides a review of previous results that deal with overload problem; also provides a review of current achievements in the field of modeling and behavioral analysis of IMS network in order to optimize the same. During the process of result analysis some disadvantages were observed, and every author provided a unique guideline for solving this complex problem. The second section provides an overview of papers dealing with overload problem, while third section gives an overview of previous work in field of modeling, behavioral analysis of the IMS network in order to optimize the same. In last section, we provide guidelines for solving this complex problem.

II. IDENTIFIED CONTRIBUTIONS AND EXPERIENCES IN SOLVING THE PROBLEM OF SIP OVERLOAD

Numerous papers deal with overload problems on SIP servers. Some of the papers include a detailed analysis of possible overload causes, while others contain suggestions and explanations of the different mechanisms and algorithms, which should have the ability to manage and control overload.

Guided by representative papers for the stated problem, authors observed and exposed few potential aspects of SIP overload classification.

- One of the observed aspects involves creation of new protocol or changes on existing protocol. According to existing achievements this aspect could be the most complete solution but IMS, SIP servers and SIP protocol are so widespread in commercial use that there is no sense to try to change main standards in that area.
  For example, Whitehead [9] 2005 described the framework independent of the protocol, GOCAP, but his mapping in SIP has not yet been defined. Even if this framework becomes mapped; questionable is if it will be accepted by manufacturers of equipment.
- Another approach implies the use of new network elements or applications which would predict congestions or overloads. This means additional investments in HW (Hardware) and SW (Software); it demands additional time, resources and efforts on existing applications to send some performance indicators.
  Luca Monacelli [11] describes the overload problems and offers stabilization system, STBZ (STaBiliZer), which protects all network elements of IMS. STBZ is a software application which collect measurements from network, processes it by appropriate stabilization policies and controls traffic shapers in order to avoid congestion on network elements
- One of most commonly used approaches is the creation of new algorithms on existing network elements.
  This approach directly influences source code of network elements, but upgrades and patches are standardized processes and are something that network operators often do which makes this solution acceptable.
  C. Shen and H. Schulzrinne [10] are among the first to deal with overload of TCP - based SIP server. They suggest new mechanisms for SIP overload control - which relies on the existing TCP flow control and congestion control. The algorithm consists of three components: Connection Split, Buffer Minimization and Smart Forwarding. Due to the specific nature of SIP protocol (session based) there is a need for separation of INVITE messages processing - requests which start a session, and other non-INVITE messages, in order to prevent opening of new sessions that could lead to overload and that, on the other hand, will preserve the existing sessions. This part of the algorithm is called Connection Split, which allows that INVITE and non-INVITE messages are treated differently. Smart Forwarding is enforced on an INVITE connection. When an INVITE message arrives, decision about the forwarding of INVITE request is made in relation to the current state of buffer. This way a session that should not be established can be canceled as soon as possible. The paper further examines the impact of buffer size (buffer at the transmitter and the receiver side and applications buffer at the receiver side) to bandwidth and processing time. It was concluded that the best results can be achieved through minimization of the buffer size on the receiving side, which is the third part of the algorithm. This algorithm shows very good results for classical SIP scenarios in core networks, where small number of transmitting servers simultaneously creates overload on the receiving server. But in the case of edge networks where the overload is prevalent, the described algorithm does not provide satisfactory results.
- Whatever solution is used to prevent overload it will not provide satisfactory results if network is not optimized and well dimensioned.
  There is no one disadvantage of this approach and this network optimization step must be applied on any professional network (of any kind).

We concluded that solving SIP overload problem requires combined use of exposed aspects. We propose combination of new algorithm on existing network elements and optimization and well dimensioning of network.

Guidelines stated by C. Shen and H. Schulzrinne [10] will provide a start point in further discussion of SIP overload problems.

III. IDENTIFIED CONTRIBUTIONS AND EXPERIENCES IN SOLVING THE PROBLEMS OF IMS NETWORK OPTIMIZATION

As stated earlier, this chapter provides a review of the previous works in the field of modeling and behavioral analysis of IMS network, in order to provide the guideline for optimization of the same.

The problem identified in IMS networks is the existence of bottleneck nodes at different network layers of IMS architecture. The authors used a variety of scientific methods to investigate this issue of which the largest contributions were provided by analysis and modeling methods.

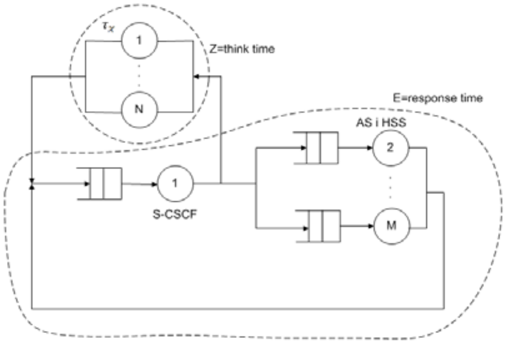

Figure 1. Model of central server for IMS [2]

Mkwawa and Kouvatsos [2], using the modeling method, have proven that exactly S-CSCF (Serving Call Session Control Function) is the throat of the IMS network that processes great number of messages.

They have compared the analytical and experimental results of IMS application modeling, user registration and the establishment of multimedia sessions, and have showed that the registration of users and establishment of multimedia session corresponds to a well-know central server model QNM (Queuing Network Model) Fig. 1.

Using Buzen's algorithm [13] and Little's theorem [12], service rates for S-CSCF, AS (Autonomous System) and HSS (Home Subscriber Server) were calculated, as well as transition probabilities, server utilization and throughput of the model.

Simulation results for server utilization and throughput for the same service speed and probability transitions were proven to coincide with the analytical model Fig. 2. [2] represents one of the ways for modeling the IMS system. This representation only indicates the network issues, but does not provide any kind of solution.

Because of the bottleneck problems additional load on S-CSCF module is not recommended, so overload problems on service layer and dynamic interaction of services are solved using Service Brokers, or SCIM (Service Capability Interaction Management) [4] module Fig. 3.

This module is not an integrated part of S-CSCF, instead it is being realized as an application server. Organization in this way allows better application server utilization in order to shift overload boundaries on the application layer. This opens new issues in the service interaction management field.

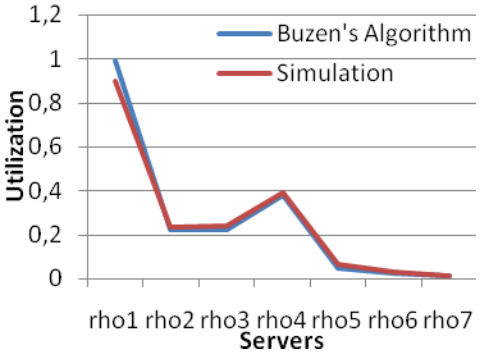

Figure 2. Comparative study of simulations and analytical results of the central server model [2]

The best overview of the results achieved by using modeling techniques are presented in [6] [7]. Both papers indicate the effects of adding servers to the networks assuming M/M/r model with Erlang C formula. It is shown in Fig. 4 and Fig. 5 that, by adding servers to the network, total waiting time in system is reduced, which has direct impact on system overload.

Two approaches were used for the simulation of M/M/r model: hyper-threading and physical server adding. It is shown that prioritizing calls affect waiting time. Fig. 6 and Fig. 7 show that with the increase of the prioritizing calls from 25 % to 50%, waiting time for all calls is increased; so, it is necessary to define a threshold value of the amount of priority calls, which will truly give better QoS and increase the performance.

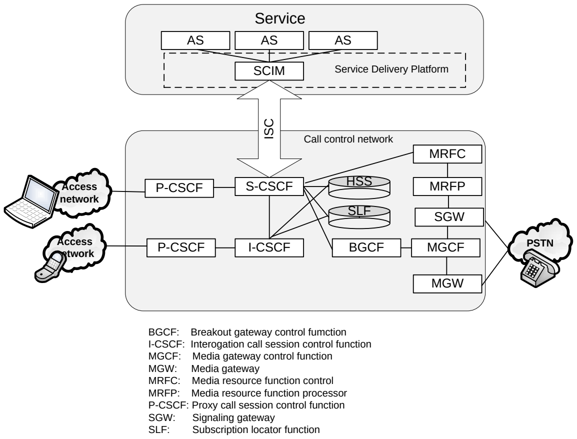
Service
AS
AS
AS
SCIM
Service Delivery Platform
ISC
Call control network
MRFC
MRFP
SGW
MGCF
MGW
Access network
Access network
P-CSCF
S-CSCF
HSS
SLF
P-CSCF
I-CSCF
BGCF
PSTN
BGCF: Breakout gateway control fumction
I-CSCF: Interogation call session control function
MGCF: Media gateway control function
MGW: Media gateway
MRFC: Media resource function control
MRFP: Media resource function processor
P-CSCF: Proxy call session control function
SGW: Signaling gateway
SLF: Subscription locator function

Figure 3. IMS architecture with SCIM [5]

ISBN: 978-1-61208-192-2

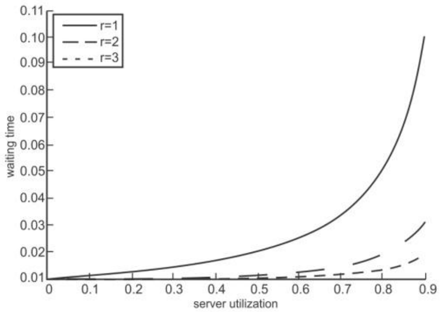

Figure 4. Mean waiting time in multi-server M/M/r system (μ=100) [6]

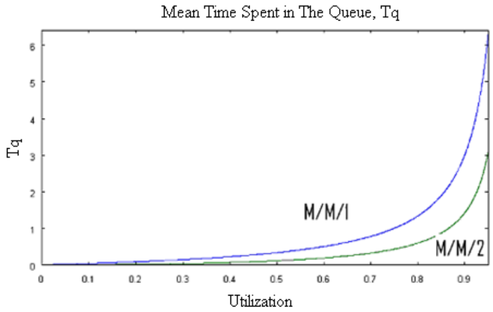

Figure 5. Mean waiting time in queue (Tq) [7]

Analysis was made with several assumptions; CSCF (Call Session Control Function) servers are replaced with one that has unlimited queue, while SIP requests and requested serving time have exponential distributions which considerably simplifies the process computation.

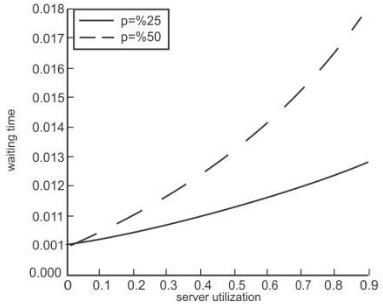

Figure 6. Comparison of prioritizing calls waiting time for two different amount of prioritizing packets [6]

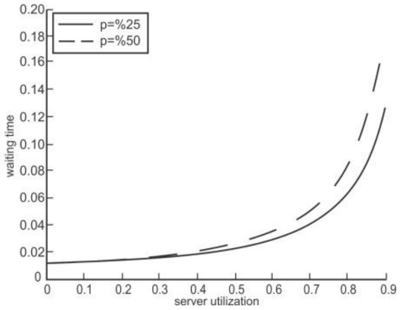

Figure 7. Comparison of non-prioritizing calls waiting time for two different amount of prioritizing packets [6]

IV. PROPOSAL FOR SOLVING OVERLOAD ISSUES AND IMS NETWORK OPTIMIZATION

Analyzing the papers in the domain of SIP servers overload and IMS network optimization can be concluded that there is no unified position on what mechanism successfully prevents overload or which traffic model gives the accurate picture of the actual IMS architecture.

Analyses that have been done so far include certain assumptions, whereby the obtained results do not provide the complete picture of the observed problems.

The complete resolution of defined problems is much more complex and must be considered in environment that is adequate for real systems. This environment assumes the use of multiple SIP servers, limited waiting lines and service times which usually do not have exponential character.

Our proposal for solving the problem of overload and optimization of IMS network is as follows:

- The prioritization of NON-INVITE over INVITE messages on S-CSCF node. This would prevent the accumulation of new sessions, which leads to efficient management and cleaning of active sessions.
- Message prioritizing implies separation of the incoming flow into two queues on S-CSCF. Service time distribution of processing an INVITE message will no longer have exponential character.
- It is necessary to experimentally determine distribution of service time of an INVITE message in case mentioned above.
- By the determination of service time distribution, preconditions are made for calculation of analytical dependencies between parameters which directly affect overload: queue length for INVITE messages and service time.
- Following step is the analytical and experimental analysis with systems that have more physically separated S-CSCF servers with previously mentioned methods applied. This analysis will provide results which should show dependency between system load and number of servers.

- Additional optimization requires application layer modeling which implies classification of applications, e.g., by real-time, non-real-time characteristics. For each group of applications, traffic management model should be defined with theoretically and experimentally determination of service requests distributions, in order to get minimal response time for every group of services. Request priority and the corresponding service prices will have important role in traffic management model. One of possible approaches is the use of multiple conditional optimal paths.

V. CONCLUSION

In order to ensure that IMS concept has the opportunity to be used in real conditions and to provide enriched services with the promising QoS, it is necessary to continue research and to make new progress towards solving these complex issues. The need to work on modeling is still persistent, and it is needed to pursue the faithful “behavior“ mapping of the IMS architecture into a model which will be able to provide analytical dependence of the output parameters with the input ones.

Only well-modeled systems can provide valid results on which can be based further mechanisms of load control, and then the steps that will contribute to the optimization of the overall architecture.

The paper gives an overview of all current achievements and provides a guideline for future work. All of the proposed methods should be the subject of the future research in order to solve the defined problem.

REFERENCES

[1] 3GPP TS 23.517, “IP Multimedia Subsystem; Functional architecture“, 2008., retrieved: December, 2011.

[2] I. M. Mkwawa and D.D. Kouvatsos, “Performance Modelling and Evaluation of IP Multimedia Subsystems“, HET-NETs08, pp. 67-79, February 2008.

[3] 3GPP TS 23.002, “Network Architecture”, 2009., retrieved: December, 2011.

[4] Nicholas S. Huslak and A.C. McQuaide Jr., “Service Brokering: Opportunities and Challenges“, AT&T Knowledge Ventures, 2007., retrieved: December, 2011.

[5] Kenichi Sakura, Soichiro Tange and Hisayuki Sekine, “Service Delivery Platform Implmenting IP Multimedia Subsystem“, FUJITSU Sci. Tech, Vol. 45, No. 4, pp. 409-414, October 2009.

[6] A.M. Amooee and A. Falahati, “Overcoming Overload in IMS by Employment of Multiserver Nodes and Priority Queues“, 2009 International Conference on Signal Processing Systems, pp. 348-352, May 2009.

[7] Mlindi Mashologu, “Performance Optimization of IP Multimedia Subsystem“, Dissertation.com, 2010., retrieved: December, 2011.

[8] Rosenberg, J., “Requirements for Management of Overload in the Session Initiation Protocol“, IETF (SIPPING, Internet Draft), Decembar 2008., retrieved: December, 2011.

[9] M. Whitehead, “GOCAP - one standardized overload control for next generation networks”, BT Technology Journal, Vol. 23, Issue 1, pp. 147-153, January 2005.

[10] C. Shen and H. Schulzrinne, “On TCP - based SIP server overload control“, IPTComm '10 Principles, Systems and Applications of IP Telecommunications, pp. 71-83, August 2010.

[11] Luca Monacelli, “Including Overload Control in Existing IMS Compilant Networks by Using Traffic Shapers“, Mobimedia '09 Proceedings of the 5th International ICST Mobile Multimedia Communications Conference, Article No. 56, September 2009.

[12] John D.C. Little, “A Proof of the Queuing Formula: L=λW", Operations Research, Vol. 9, No. 3, pp. 383-387, May-June 1961.

[13] Jeffrey P. Buzen, “Computational algorithms for closed queueing networks with exponential servers”, Communications of the ACM, Vol. 16, No. 9, pp. 527-531, September 1973.