International Journal of Interactive Communication Systems and Technologies
Volume 9 • Issue 1 • January-June 2019

Scalability and Performance Analysis of SIP based Multimedia Services over Mission Critical Communication Systems

Ashraf A. Ali, University of South Wales, Pontypridd, UK; Hashemite University, Az-Zarqa, Jordan

Khalid Al-Begain, Kuwait College of Science and Technology, Kuwait City, Kuwait

Andrew Ware, Faculty of Computing, Engineering and Science, University of South Wales, Pontypridd, UK

ABSTRACT

Various studies have suggested enhancing the performance of large-scale systems, such as mission critical communication systems (MCCSs). However, few have modelled and evaluated the performance of such systems in a way that targets overall system performance in real time. Moreover, it is not enough to define the Key Performance Indicators (KPIs) for a system without using them for system performance measurement and performance evaluation. The Session Initiation Protocol (SIP) and IP Multimedia Subsystem (IMS) both have a set of KPIs, such as the registration process delay, that can be used to measure and thus optimize overall system performance. This article articulates different options for system simulation and evaluation. The registration process affects performance and reflects the overall system performance. The article shows how the registration process is delayed and how the overall system scalability are negatively impacted by system overload.

KEYWORDS

IMS, IP Multimedia Subsystem, Long Term Evolution, LTE, Modelling, Performance Evaluation, Session Initiation Protocol, SIP

INTRODUCTION

IP Multimedia Subsystems (IMS) (3GPP, 2006) and Session Initiation Protocol (SIP) (Rosenberg, 2002) performance play a major role in multimedia communication networks by altering the Key Performance Indicators (KPIs) related to the Quality of Experience (QoE) metrics of the end-to-end service. Registration Request Delay (RRD) is one of the SIP KPIs that also influence both IMS KPIs and end user QoE. Therefore, it is crucial to evaluate the performance of both SIP and IMS based on the RRD metric in order to give an indication of the overall system capacity and scalability potential.

5G communications is the new technology that will integrate multiple access technology in to one integrated solution adopted by all vendors and manufacturers. End users and devices will be able to communicate seamlessly with fewer restrictions and more options compared to older technologies Scalability is among the challenges that limit the exploitation of the full capabilities of the current technologies. Many recent studies have tried to overcome the scalability challenge associated with the 5G standards set. A new integrated solution with external SIP application that is accessed over

DOI: 10.4018/IJICST.2019010102

LTE with different voice codecs is introduced in other studies (Haibeh & Hakem, 2017). However, the solution proposed lacks the support of standardised adopted solutions thus introducing complexities in implementations. In a similar SIP performance evaluation an enhancement trial (Subramanian & Dutta, 2009), transaction states of the SIP server are characterised to model the performance of the server. In the trail M/M/c queuing model was used along with a benchmarking performance indicator that reflects system performance. The trial showed that the multi-threaded architecture utilising parallel processing of SIP messages provides a more scalable and efficient solution for larger number of clients. However, the study was based on simple SIP servers that do not reflect more complicated processing required by multimedia services that tend to use more than one SIP server.

Another study (Ono & Schulzrinne, 2008) used the Stream Control Transmission Protocol (SCTP) as a transport protocol for SIP messages instead of TCP and UDP. The implications of using different transport protocol over SIP scalability and performance is presented and it is shown that SCTP has a negative implication over SIP scalability due to the added overhead. Other research (Yavas, Hokelek, & Gunsel, 2016) has focused on the scheduling mechanisms to prevent SIP system overload and to increase its scalability, the proposed solution uses a priority-based mechanism to dynamically estimate the behaviour of the server and provide a more scalable solution compared to the conventional SIP servers. Again, this study evaluates one single SIP server.

This paper analyses how overall system capacity and scalability is affected by additional traffic generated when more users try to access the systems services. This could happen in a mission critical communication system during natural disaster or large-scale attack, resulting in a sudden increase in the number of users.

It was found that, within limits, the system's ability to process the registration requests per time unit increases exponentially when the number of users is increased. Once the limit is reached however, the number of processed requests starts to decrease and eventually degrade leading to system failure. The simulation results show that the system was able to handle a maximum of 7,400 registrations per second, a workload that could occur during a nationwide disaster with many users trying to access the Mission Critical System (MCS).

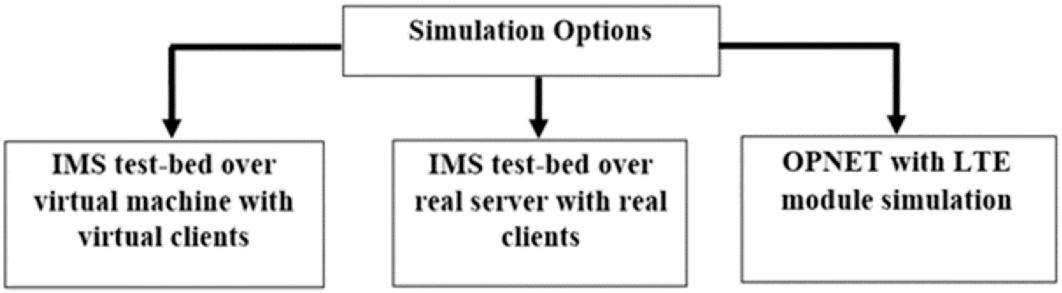
The need for a more detailed study of other SIP and IMS KPIs to provide a better understanding of the overall system performance is thus clear. The study will enable further progress towards system performance enhancement and optimization in order to avoid single point of failure of the system.

RESEARCH METHODOLOGY

A previously developed research methodology (Creswell, 2009) was followed for both the qualitative and quantitative approaches when setting the parameters for all measurements and simulations. The methodology for deciding the qualitative values that need to be investigated can be summarized as follows:

1. Determine the challenges that need to be investigated within the scope of the study. While the project embeds several challenges, the focus was placed on the signalling domain especially between the end user and the core network and the signalling interface between the core network and IMS.
2. Determine the benchmark for what is considered acceptable SIP performance and decide on the metrics that will be measured and used to judge and compare the performance of setup.
3. Decide the appropriate simulation tools to generate the results from multiple sources that meet the appropriate comparison criteria based on the selected tool.
4. Determine the key factors that affect the SIP signalling, in addition to multimedia services operation in LTE and IMS that affect the overall QoS for the Mission Critical system.

Figure 1. Available simulation and testing tools



The Quantitative Methodology to acquire the needed measurements is summarized as follows:

1. Develop a test-bed for the IMS to determine the performance of the system. Then decide the performance metrics that need to be measured in order to facilitate comparison with other implementations and scenarios.
2. Develop a virtual machine to generate virtual clients along with IMS in order to facilitate comparison between the performance of the system with the real running test-bed.
3. Develop a simulation project for both LTE and IMS over OPNET to benchmark their performance against the test-bed implementation performance metrics.
4. Determine the variables, models, scenarios and parameters in OPNET that need to be adjusted and analysed to enable overall system's performance evaluation.

In summary, as shown in figure 1, there are three simulation/testing options.

This section presents the test-beds and simulation results that related to the project. First, IMS test-bed scenario and results are given. Secondly, the OPNET simulation scenarios and results are demonstrated. Thirdly, the limitations and challenges of both experiments are discussed.

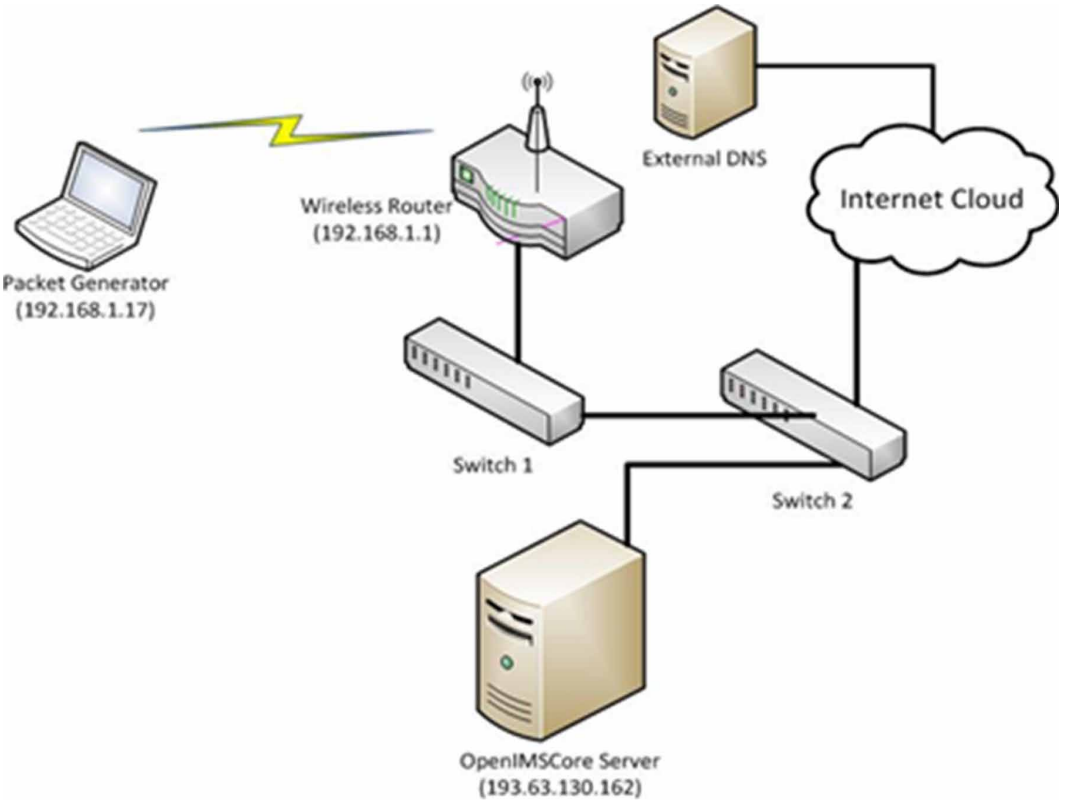
TEST-BED EXPERIMENT

Figure 2 shows the experimental topology of the test-bed. The test-bed is composed of four main parts: the Packet Generator; IMS core which is based on Open-IMS-Core model (Fokus, 2004); Packet Analyser, and Domain Name Server (DNS). The parts function and operation are as follows:

- Packet Generator: The packet generator is responsible for simulating virtual clients that then generate concurrent calls that are transmitted in a serial or parallel manner by a theoretically unlimited number of users. Due to the focus on SIP and IMS performance, the Packet Generator is designed to send SIP Register Message (as defined by RFC 3261) in addition to SIP invite and bye messages. All the messages are transported using the UDP where the sender port address is dynamically allocated so as to avoid using restricted ports at the sender or server sides. The GUI interface of the Packet Generator enables the user to select a predefined set of users and the Proxy IP address. Finally, the SIP request-sending pattern is selected to be either serial or parallel.

- IMS Network: The IMS core is based on Open-IMS-Core (Fokus, 2004) developed by the FOKUS Institute for Open Communication System. The server embed the IMS Call Session Control Functions CSCFs; such as PCSCF, I-CSCF, and S-CSCF, in addition to the home Subscriber

Figure 2. Experiment test-bed



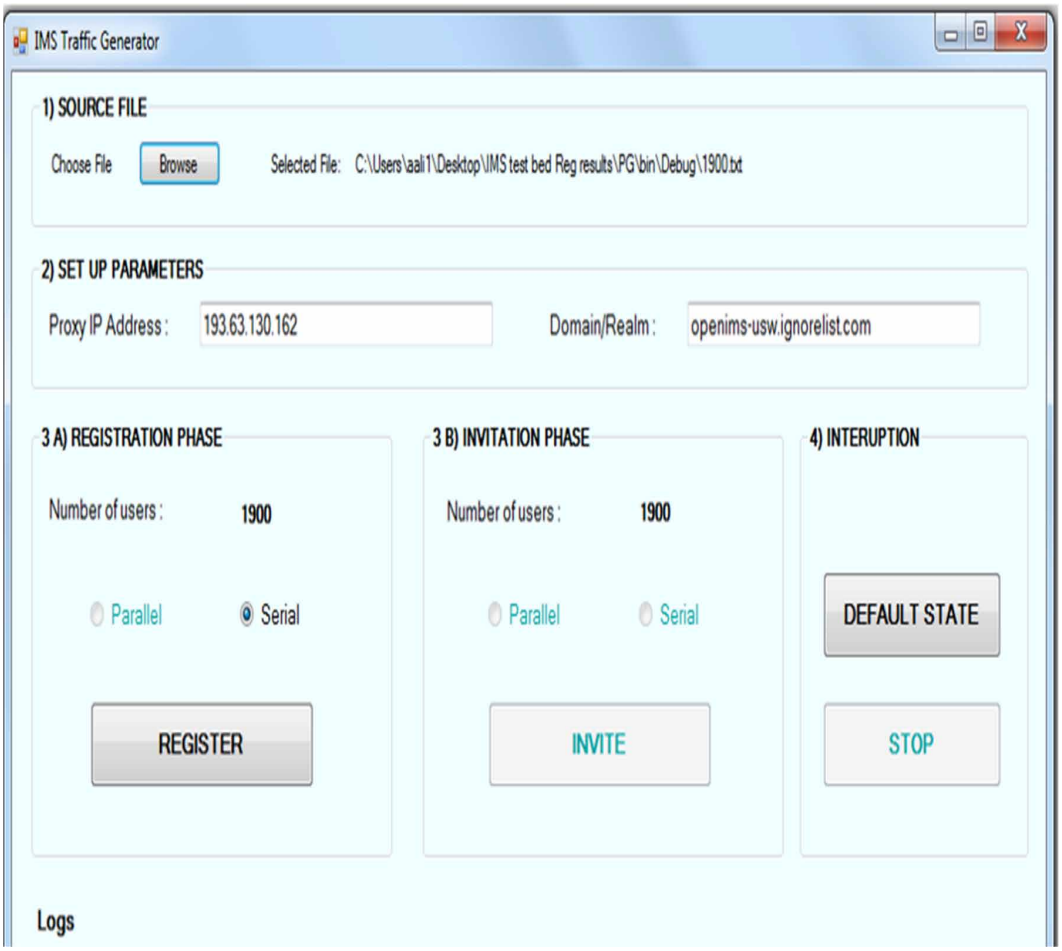
Station HSS. All are considered part of the core architecture for the next generation of networks as specified by 3GPP. The purpose of the experiment is to test the capacity of the IMS in terms of the maximum number of users that can be adopted by the system without causing stability issues.

- Packet Analyser: The packets sent by the Packet Generator are monitored using Wireshark as a packet analyser at the sender side. The trace files extracted from the packet analyzer help in calculating the KPI values for both SIP and IMS.
- DNS: an external Domain Name Server (DNS) to resolve the IP addresses of all servers in the system setup.

Based on the previous setup the experiment aimed to evaluate the SIP performance over the IMS using either a wired or a wireless connectivity with the server. For this purpose, the register message delay was calculated by running Wireshark at the packet generator side and calculating the difference between the sent registration request time and the 2.00OK response reception time. The data was then exported using MATLAB and analysed the Probability Density Function (PDF) and Cumulative Density Function (CDF) curves calculated. These provide a better understanding of the variance in Registration delay within the same scenario and among different running scenarios.

Figure 3 shows the GUI interface for the Packet Generator. The user first selects a predefined set of users' databases and the Proxy IP address (which is the IMS server IP). In addition, the domain name is inserted and the SIP request sending pattern selected (either series or parallel). Pressing REGISTER initiates the sending of the registration requests (one per user) consecutively and dynamically. The packets sent by the Packet Generator are monitored using Wireshark at the sender side while the log

Figure 3. Packet generator GUI



screen at the packet generator records the time stamp of the sent requests and received responses, which helps in calculating the end-to-end application delay (the time between peer application layers).

Based on the previous setup, the experiment aimed to evaluate the SIP performance over the IMS using either wired or wireless connectivity to the server. For this purpose, the register message delay is calculated by running Wireshark at the packet generator side and calculating the difference between the sent registration request time and the 2,00OK response reception time. The data is then, using MATLAB, exported and manipulated in order to generate curves for a better indication of the variance in Registration delay with time. The experiment was repeated multiple times, each time the number of users was incremented in both the wired and wireless scenarios.

RESULTS

In this scenario, the packet generator was wired directly to the router and the number of users sending the registration request were incremented in steps of 200 in the range from 100 to 1,300 users. Figure 4 shows the PDF and CDF of the registration delay for 100 while Figure 5 shows the PDF and CDF for 500 users.

Figure 4. CDF and Density functions of the Registration delay for 100 users

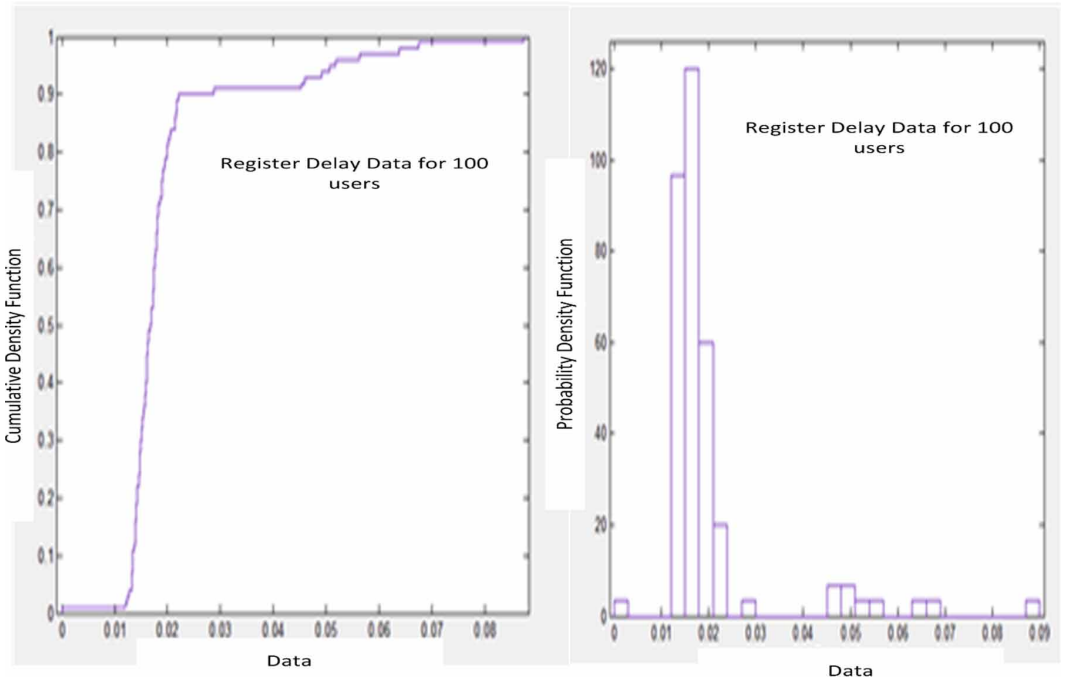


Figure 5. CDF and Density functions of the Registration delay for 500 users

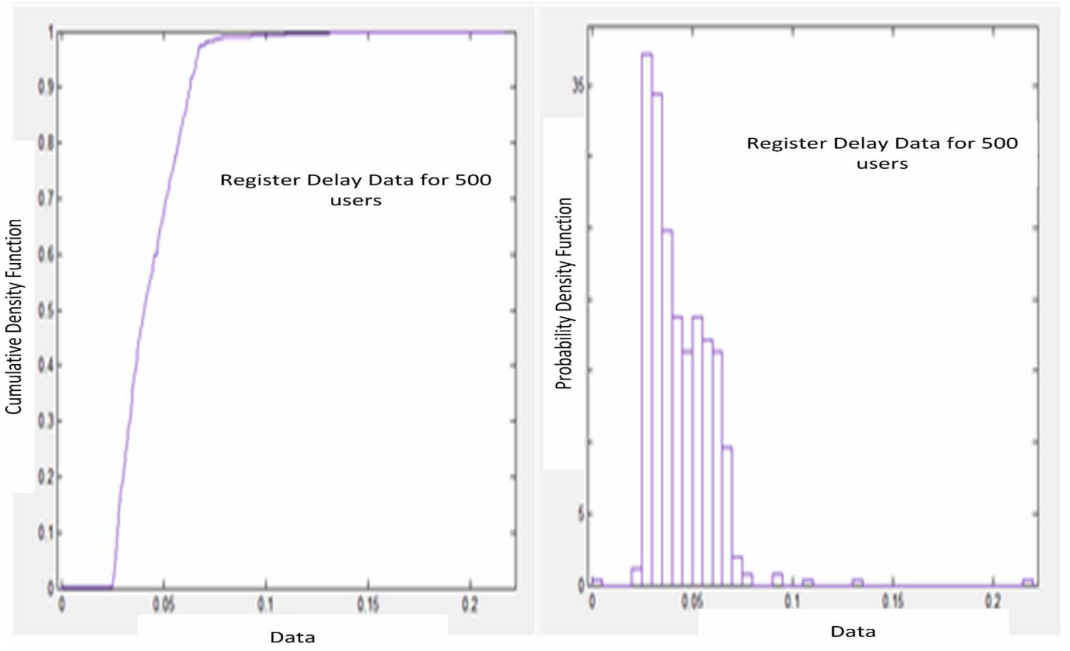
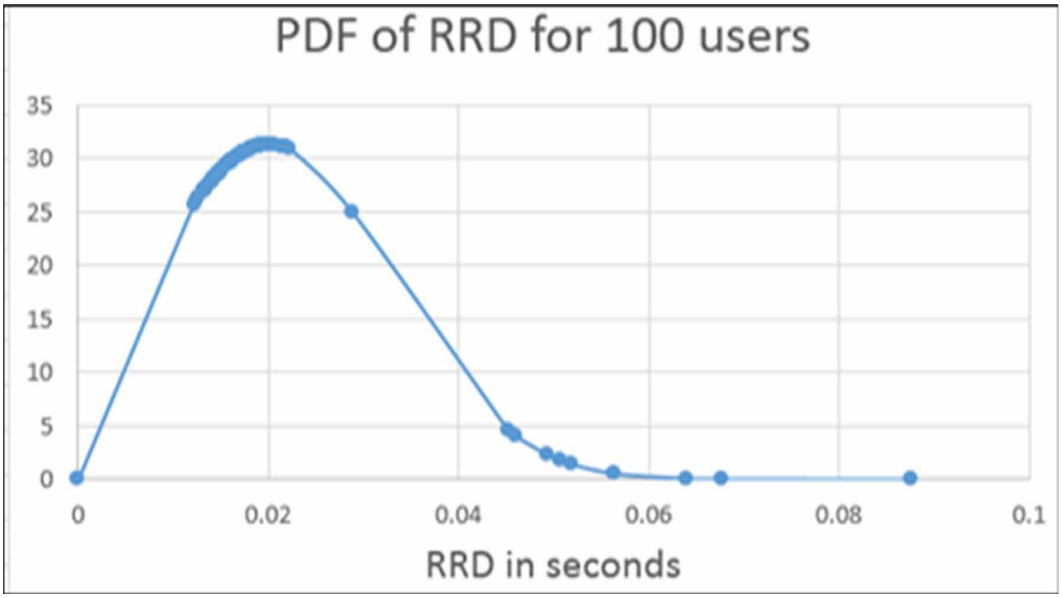


Figure 6. PDF of RRD for 100 users



From figures 4 and 5, it is clear that the registration delay for 500 users more than doubled compared to that achieved for 100 users, and increased even further when increasing the number of users in each step. Similarly, seven varying scenarios were implemented using the test-bed. In the first scenario, 100 user registration requests were transmitted, while in the subsequent scenarios the number of users was incremented in steps of 200 until 1,300 users were considered in the seventh scenario. To test the scalability of the system, the number of users was gradually increased in order to gain a better understanding of the relation between the number of users and the KPI values for both SIP and IMS.

Figure 6 shows the Probability Distribution Function (PDF) and figure 7 the Cumulative Distribution Function (CDF) for the RRD of the first scenario with only 100 users each sending one registration request at a time in sequential order. As shown in figure 7, 90% of Registration requests need less than 40 ms to be completed which meets the requirements of mission critical applications and real-time services. As shown in figure 6, the highest frequency of the registration trials needs on average 20 ms to be completed. This is considered the best-case scenario and was used as a benchmark for the other scenarios in order to enable comparison of both the RRD time and the percentage of trials that finish at certain time threshold.

Similarly, the PDFs and CDFs for all seven scenarios were generated as shown in figure 8 and 9. It is clear that when the number of clients increases, the system needs more time to serve the registration requests. This happens due to accumulation of both SIP and DIAMETER signalling messages in the queues of the CSCFs interfaces (especially in S-CSCF) and the HSS interface. Both S-CSCF and HSS are considered bottleneck points of congestion that are affected significantly as the number of registration requests increase. This leads to a queuing delay that emerges rapidly in the system interfaces, which can eventually cause system failure.

Two performance metrics were used to facilitate comparison of the seven scenarios. The first was the time needed to process successfully 90 percent of requests, referred to as 90% completion time (90CT). While the second was the percentage of successfully completed registration requests within 40 ms seconds (which is the maximum RRD time needed to process 90% of requests in the 100-users scenario), referred to as 40ms Completion Ratio (40msCR).

Figure 7. CDF of RRD for 100 users

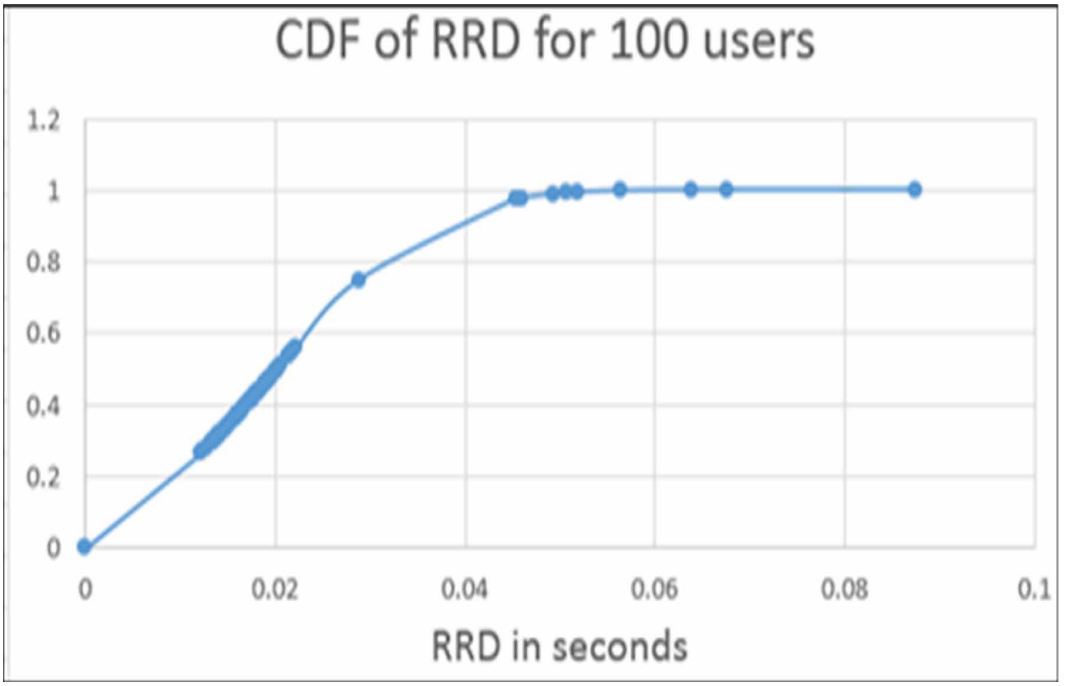


Figure 8. PDF of RRD values for all scenarios

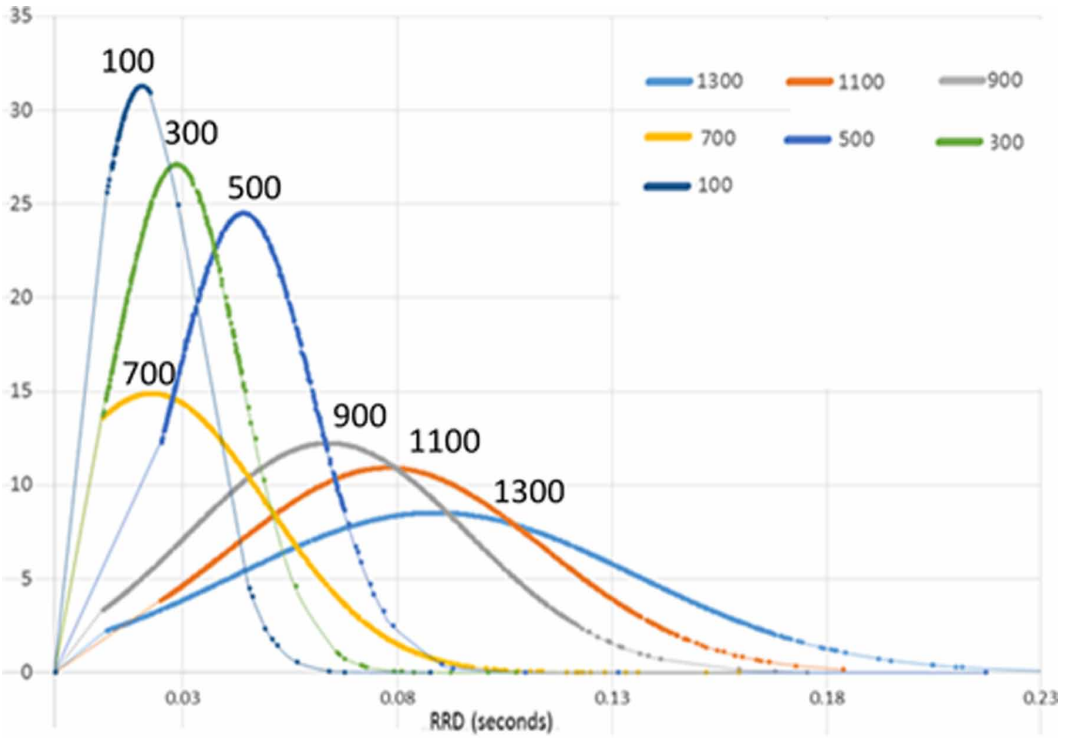


Figure 9. CDF of RRD values for all scenarios

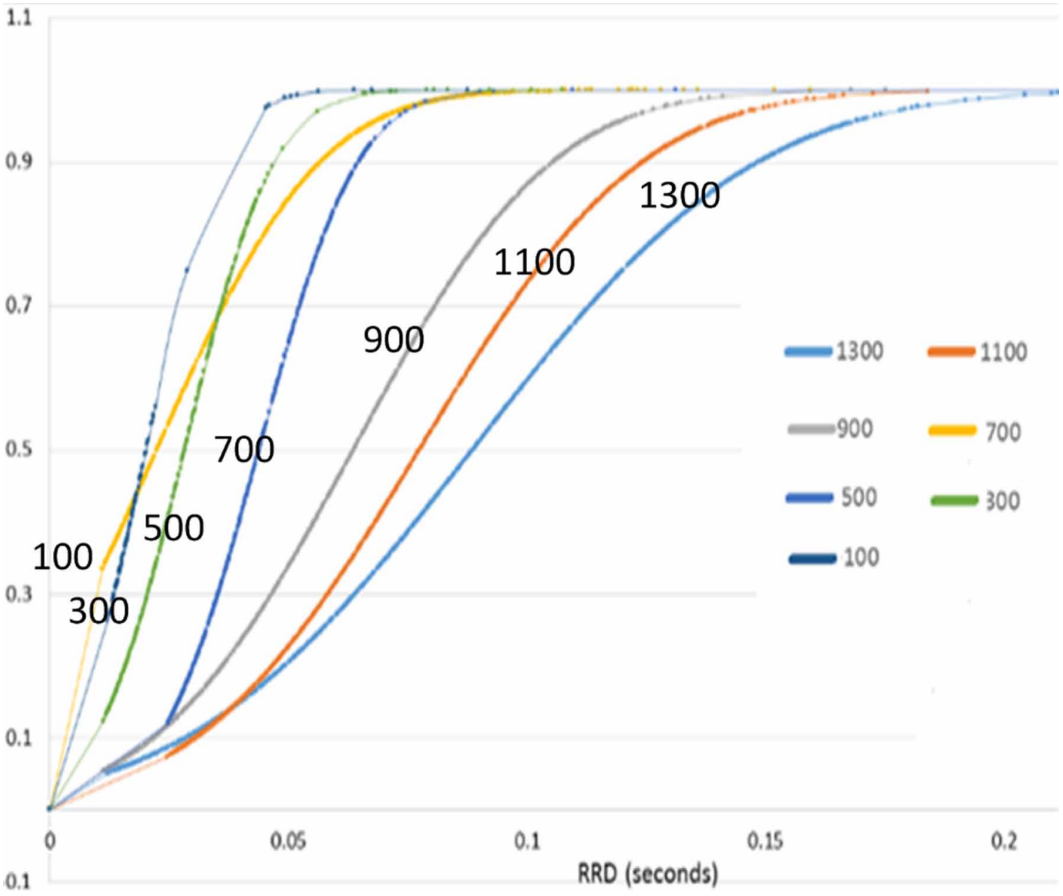


Table 1 shows the average RRD values along with 90CT and 40msCR for each scenario and the registration requests per second (RRps) processing rate (which is the number of registration requests that are successfully processed by the test-bed per unit time). The RRps values replicate a real world disaster scenario, where thousands of users may send registration request to gain access to the system - which is supposed to be scalable and reliable - at the same time.

Based on the results, it can be seen that the RRps increases exponentially as the number of users increases up to a limit (of 1,100 users) before beginning to decrease, leading eventually to system degradation and failure. This is shown clearly by RRps for both scenarios 6 and 7, where the RRps of scenario 7 is much less than the RRps for scenario 6 although the number of users has increased by 200.

As expected, the 90CT increases as the number of users rises, starting from 40ms for scenario 1 through to 150ms for scenario 7. This is to be expected due to the increased processing time needed for the additional received registration request. Moreover, it was found that the 40msCR decreases with an increased number of users. Comparing the values with scenario 1 (the benchmark) shows that only 15% of registration requests needed less than 40 ms RRD value to be completed, which again implies that the system is not able to process the received request within very strict time limit.

Table 1. Calls statistics from simulation results

Scenario no.	RRps	RRD Avg. Value	90CT (ms)	40msCR (%)
Scenario 1 (100 users)	1,800	20	40	100%
Scenario 2 (300 users)	3,600	28	47	80%
Scenario 3 (500 users)	5,900	44	65	40%
Scenario 4 (700 users)	4,900	22	55	73%
Scenario 5 (900 users)	5,500	64	105	23%
Scenario 6 (1100 users)	7,400	77	125	17%
Scenario 7 (1300 users)	5,800	88	150	15%

OPNET SIMULATION

To facilitate investigation of the LTE system, especially the SIP signalling performance over LTE communication network, the OPNET simulator was used to create a scenario with multiple users initiating calls. This enabled the SIP performance metrics to be used to measure the efficiency of the system and its capacity tolerance. Figure 10 shows the created setup.

Simulation Setup and Scenarios

In this research study, OPNET Modeller provides the required level of simulation capabilities to implement and model different multimedia applications over LTE. The system design that was implemented and investigated is shown in figure 10 and is based on the configuration parameters shown in Table 2. The implementation of the LTE network system is based on a single Evolved Packet Core (EPC) that serves two eNBs, each with four clients. The clients in eNB1 make SIP-based VoIP calls to the clients in eNB2 through the EPC in a Normal distribution call generation system, using a fixed-length call. The EPC is then connected to the SIP server, which reflects the performance of the P-CSCF in the IMS, that manage the registration, call initiation and call termination processes using the SIP signaling system using the IP cloud. In this reseach, the simulations were performed without any background traffic in the LTE system and the IP network.This enables us to study the actual performance level for SIP-based VoIP applications within a best effort environment which helps with the results accuracy. It should be noted that this research has not considered any clients mobility performance implication over signalling delays, it is left as a future work to discuss it further.

The simulation implementations has considered four scenarios based on the design shown in figure 10 and the simulation parameters shown in Table 2. The first scenario represents the basic implementation for VoIP applications over LTE using a single pair of UEs between client A-1 in eNB 1 and client B-1 in eNB 2. This scenario examines the best-case implementation of the assigned network system with only one single call at a time. The second scenario has an additional connection with multiple calls with another pair of UEs (client A-2 and client B-2) added to the first scenario. A similar thing is true of the third scenario, where additional pairs of calls are added (client A-3 and client B-3). Finally, the fourth scenario has yet another additional pair between client A-4 and B-4. This gradual increase in the pairs of SIP-based VoIP calls allowed the performance of the SIP signalling system over LTE based communications with additional VoIP calls between different clients to be checked. The highest load of VoIP calls is represented in the fourth scenario that consumes higher bandwidth over LTE where all clients in each eNB are calling one single client in the other eNB. Therefore, the results of these implemented scenarios can be compared and studied throughout the research study in terms of the performance for SIP signalling and efficiency for LTE system.

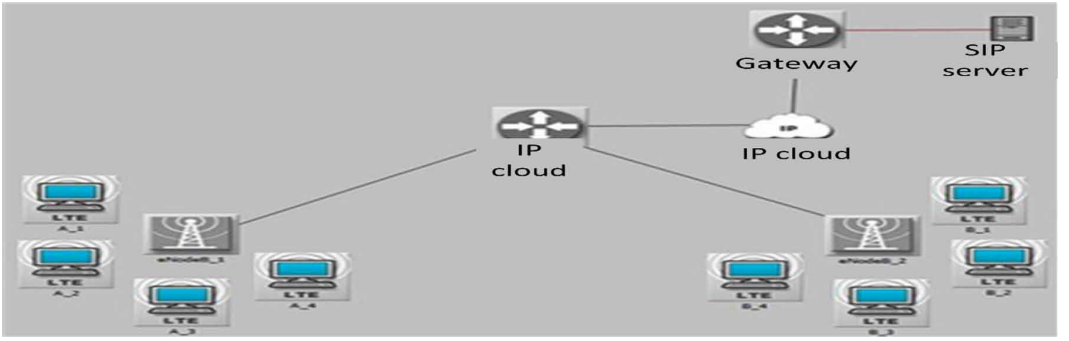
Table 2. Simulation parameters in OPNET

A. LTE Network System					
Number of Simulations	4		Simulation Seed Number		128
Simulation Duration:			30 Minutes = 1800 Seconds		
Number of EPC:		1	Background Traffic	0%	
Number of eNB:		2	Number of nodes for each eNB:	4	
Antenna Gain for eNB:		15dBi	eNB Maximum Transmission Power:	0.5 W	
eNB Receiver Sensitivity:		- 200 dBm	eNB Selection Threshold:	- 110 dBm	
B. Applications: SIP Based VoIP					
VoIP Calls (Unlimited)	Call Duration		Caller	Callee	
	10 Sec		Node A	Node B	
Maximum Simultaneous Calls	SIP Server		User Agent (Caller/Callee)	Voice Codec:	
	Unlimited Call/ Second		1 call at time between each pair	GSM 13 Kbps	
Calls Start Time Offset:			Normal (150 sec, 100 sec)		
Calls Inter-repetition Time:			Normal (20 sec, 5 sec)		

Simulation Results

As the main considerations in this study are SIP signalling and LTE performance for mission critical

Figure 10. System design and implementation for SIP-based VoIP applications over LTE network system in OPNET



systems, the results focus on the call setup time and the related LTE performance metrics. The optimum number of initiated calls for each pair of calls falls between 150 and 180 for 30 minutes of simulation time with a uniform based distribution system for calls initiation. Table 3 shows the number of rejected calls in the overall system for the four scenarios with the implemented normal based system. The number of rejected calls has increased with the increased number of initiated call pairs. For calls implemented from Caller A-1, the number of failed calls *initiation* processes had been increased with the increased number of call pairs with scenarios S2, S3, and S4, where the total initiated calls over all scenarios is 56. This increased fail rate during the call initiation stage is mainly related to the inferior processing performance of the SIP servers' and LTE system performance.

Table 3. Calls statistics from simulation results

SIP calls statistics for the Implemented Scenarios				
Scenario	**S1: 1Pair**	**S2: 2Pairs**	**S3: 3Pairs**	**S4: 4Pairs**
Number of Calls Rejected in the overall system	45	95	152	218
Number of Calls Initiated from Caller A-1	56	56	56	56
Number of failed calls initiation for calls from Caller A-1	27	30	38	34

Call Setup Performance

The purpose of studying the call setup time is to facilitate analysis of the SIP signalling performance during the main SIP signalling stage over different call sessions. As long as the call setup time for the majority of initiated SIP-based calls were in the acceptable range, the performance of the SIP signalling system falls in its acceptable level (D. Malas, 2011) (Voznak & Rozhon, 2010). Figure 11 represents the average call setup time for all successful VoIP calls for the four implemented scenarios.

The results show that the scenario with only one pair of calls had the lowest average call setup with times ranging from between 46 and 47 ms, and increased up to 48.5ms when the scenario involved two pairs. With three pairs of VoIP calls, the average call setup time increased from 47ms to 49.5ms. The longest call setup time registered was for the fourth scenario in which four pairs of VoIP connections were active at the same time. These simultaneous calls affected the SIP signalling performance and increased the average delay by up to 50.5ms. In general, the call setup time for successfully initiated calls over all scenarios is still at an acceptable level when considering the performance of the SIP signalling system. This was due to implementing the LTE network system without having extra overloads due to added background traffic.

Figure 11. Call setup delay

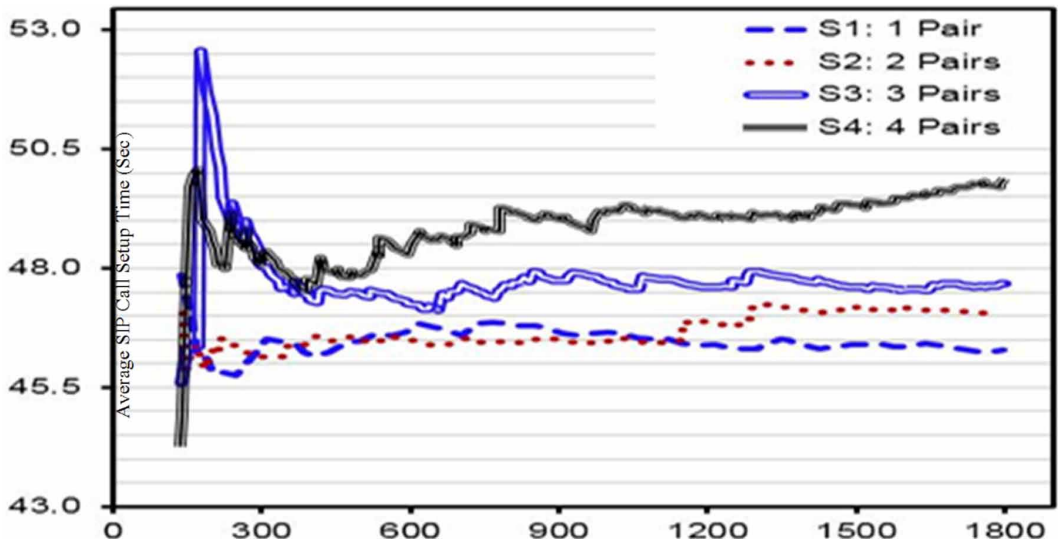
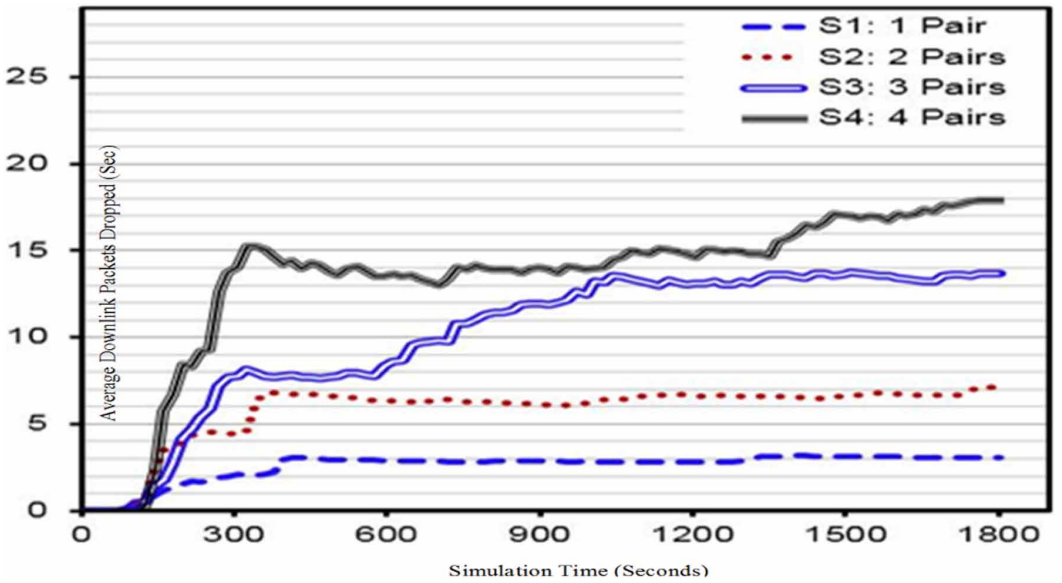


Figure 12. Average packets dropped



LTE Downlink Packets Dropped

The LTE parameters of the implemented system have a direct effect on the performance of the running applications. Real-time applications can be enhanced if the LTE system performance behaviour has been considered. The average number of packets dropped starts between 1 and 3 packets/sec for the single pair scenario and increases to between 6 and 18 packets/sec for the four pairs scenario as shown in Figure 12. The downlink packets dropped of LTE system has a direct link to successful rate of the SIP sessions in which it is directly proportionally increasing.

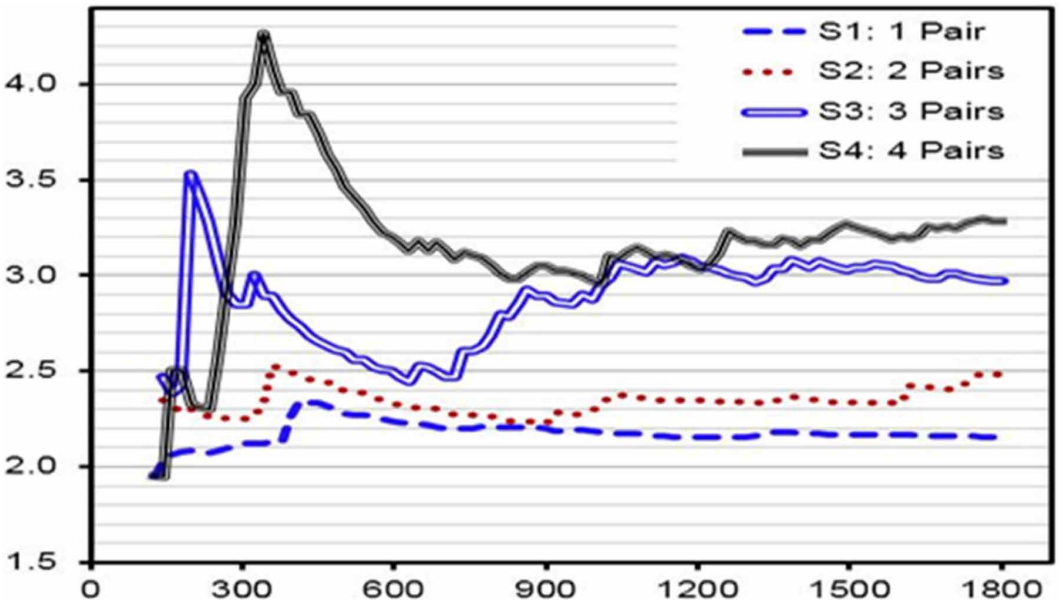
The LTE system delays in the transferred data between LTE components affect the performance for real-time applications. The average LTE delays with one and two pairs of VoIP calls is between 2ms and 2.7ms, as shown in Figure 13. The average LTE delays for three pairs of VoIP calls is from 2.4ms to 3.5ms, and between 2.5ms and 4.3ms with four pairs of calls. The longest delays mostly occur at the system start-up time and stabilise later during the simulation time.

CONCLUSION

Based on the test-bed results, it has been shown that the scalability of the system is negatively affected by the increasing number of registration requests sent to the system. It was found the RRps increases exponentially when the number of users is increased (up to a limit 1,100 users) before the RRps starts to decrease leading eventually to system degradation and failure. This is clearly shown by RRps for both scenarios 6 and 7, where the RRps of scenario 7 is much less than the RRps for scenario 6 (although the number of users increased by 200 users).

Based on the simulation results, it is clear that there is increasing delay in the call setup time when the LTE communication system is used. This delay increases as the number of served client also increases, which indicates that the Delay requirement or the maximum number of users that can be served at a time may not meet the mission critical service requirements. Hence, the need for decreasing the gap of call setup delay for commercial broadband systems compared with other dedicated mission-critical communications systems is of great importance and considered one of the main challenges for mission-critical communications. This means that there is a need for a new

Figure 13. Average LTE delays in second for caller A-1 node



mechanism that minmises access delay overhead by exploring the LTE and IMS domains in addition to the interfaces between LTE and IMS and the interface between LTE and User Element.

The simulation has been implemented using static mobile nodes (that is, the positions of the nodes are fixed). Hence, there is no handoff added complexity for the nodes moving between two cell domains. If mobility were to be considered (that does not imply moving nodes only but rather a dynamic topology) then support for handoff mechanisms between the subscriber stations and different base stations would need to be considered. Therefore, further testing of different communication scenarios for an end-to-end connectivity over LTE communication system is needed. For such dynamic topology, the need for measuring the overall performance of the system in terms of SIP signalling and data streaming delay is crucial.

REFERENCES

Creswell, J. W. (2009). *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. SAGE Publications.

Fokus. (2004). Open IMS core. Retrieved from http://www.openimscore.org

3. GPP. (2006). IP Multimedia Subsystem (IMS), Stage 2.

Haibeh, L. A., & Hakem, N. (2017, October 19-21). A new mobile SIP proxy integration solution to increase scalability in 5G mobile operators. *Paper presented at the 2017 IEEE 8th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCON)*.

Malas, D. A. M. (2011). Basic Telephony SIP End-to-End Performance Metrics. In Request for Comments: 6076: Internet Engineering Task Force (IETF).

Ono, K., & Schulzrinne, H. (2008, Nov. 30-Dec. 4). The Impact of SCTP on SIP Server Scalability and Performance. *Paper presented at the IEEE GLOBECOM 2008 - 2008 IEEE Global Telecommunications Conference*.

Rosenberg, J., Schulzrinne, H., Camarillo, G., Johnston, A., Peterson, J., Sparks, R., . . . Schooler, E. (2002). SIP: Session Initiation Protocol RFC 3261.

Subramanian, S. V., & Dutta, R. (2009, November 10-12). Performance and scalability of M/M/c based queuing model of the SIP Proxy Server - a practical approach. *Paper presented at the 2009 Australasian Telecommunication Networks and Applications Conference (ATNAC)*. doi:10.1109/ATNAC.2009.5464717

Voznak, M., & Rozhon, J. (2010, September 20-25). SIP Back to Back User Benchmarking. *Paper presented at the 2010 6th International Conference on Wireless and Mobile Communications*.

Yavas, D. Y., Hokelek, I., & Gunsel, B. (2016, November 1-3). Analytical Model of Priority Based Request Scheduling Mechanism Preventing SIP Server Overload. *Paper presented at the MILCOM 2016 - 2016 IEEE Military Communications Conference*.