Paper

Statistical Analysis of Message Delay in SIP Proxy Server

Pavel Abaev[1], Rostislav Razumchik[2], and Ivan Uglov[3]

[1] *Department of Applied Informatics and Probability Theory, Peoples' Friendship University of Russia, Moscow, Russia*
[2] *Institute of Informatics Problems of RAS, Moscow, Russia*
[3] *Moscow Technical University of Radio Engineering, Electronics and Automation, Moscow, Russia*

Abstract—Single hop delay of SIP message going through SIP proxy server operating in carriers backbone network is being analyzed. Results indicate that message sojourn times inside SIP server in most cases do not exceed order of tens of milliseconds (99% of all SIP-I messages experience less than 21 ms of sojourn delay) but there were observed very large delays which can hardly be attributed to message specific processing procedures. It is observed that delays are very variable. Delay components distribution that is to identified are not exponentially distributed or nearly constant even per message type or size. The authors show that measured waiting time and minimum transit time through SIP server can be approximated by acyclic phase-type distributions but accuracy of approximation at very high values of quantiles depends on the number outliers in the data. This finding suggests that modeling of SIP server with queueing system of G|PH|c type may server as an adequate solution.

Keywords—single hop delay, SIP server, statistical analysis, waiting time.

1. Introduction

Collapses of SIP proxy servers in carrier networks as described in [1] influenced the increase of research interests of SIP overload problem. Experimental evaluation of SIP server given in [2] evidenced that its performance greatly depends of server scenario and how SIP protocol is used. A SIP server can become overloaded due to various reasons such as denial of service, flash crowds (e.g., sudden increase in the number of phone calls), unintended traffic (e.g., unnecessary multiple copies of messages due to configuration mistakes), software errors (e.g., memory leaks, infinite loops). Traditional approach – resource overprovisioning – can reduce the overload probability significantly, but such a passive action would cause low average capacity utilization and increase the capital costs. Simple rejection of calls can mitigate the overload quickly, but it would reduce the revenue and decrease user's perceived quality of service. Meanwhile SIP specifications do not provide much guidance on how to react to overload conditions and thus significant efforts have been made to address the SIP overload problem (see, e.g. [3]–[7]). The overview of current state of art in SIP overload control algorithms which aim at preventing server crashes in carrier networks can be found in [8].

According to [9], when designing a mechanism for protecting a SIP server from overload one needs to take into account many issues: how to identify and indicate load status of the server, what is the procedure which allows reduction of overload with respect to quality of service requirements. Another problem is if it requires cooperation between servers and user agents or not and in which way, how this procedure can be implemented, whether it requires changes in the protocol or not. Taking into account all of these issues within single analytical model is a challenging task. Almost surely that this approach will lead to intractable results if one considers close to real-life values of model parameters, i.e., number of servers, clients, and concurrent sessions. Nevertheless analytical results even for single components of SIP network (if one manages to obtain and validate them) are of high importance due to their versatility. Their process validation is either performed by direct comparison with measured data from real-life experiments or performed within simulation environment which allows complex assembly models from simple ones. Leaving aside all the simulation drawbacks of SIP protocol it is a valuable mean for assessing performance of SIP networks and, as stated in [10], event-driven simulation has been widely used for evaluating SIP network performance.

In order to find the root cause of overload in SIP network, different analytical models have been proposed to analyze the statistical characteristics or dynamic behavior of SIP. Analytical model (queueing network composed of six $M|M|1$ queues) for estimation of mean response time for call setup in SIP network is presented in [11]. Analogous model but with constant service time was analyzed in [12]. In [13] and [14] results of analytical modeling of SIP proxy server as $M|M|c$ queue are compared with experimental data. Problem of service times characterization motivated authors of [15] to model SIP server as $M|G|1$ queue with service time distribution with six modes. Such choice of service distribution was justified by measurements of times SIP messages spend inside SIP proxy server. More general model $M^{[X]}|G|1$ which allows two types of batch arrivals is considered in [16]. Markov-modulated queueing model is introduced in [17] to analyze the queuing mechanism of SIP server under two typical service states. Authors state that the model can be used to predict the probability of SIP retransmissions, because the theoretical retransmission probability calculated by Markov Modulated Poisson Process (MMPP) model is located within the confidence interval of the real retransmission probability obtained from nu-

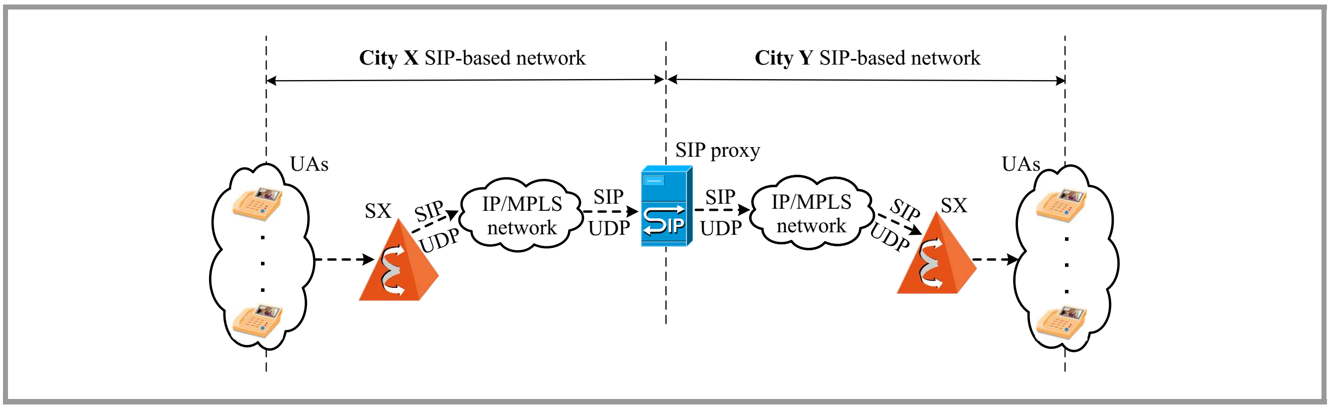

Fig. 1. Fragment of SIP based transit network under study.

merous simulation replications. Having noticed that it may be unnecessary to reject calls upon a short-term overload, authors in [10] developed a fluid model to capture the dynamic behavior of SIP retransmission mechanism of a single server with infinite buffer. A related study of a tandem server gives the guidance on how to extend the innovative approach to model an arbitrary SIP network. Fluid model for an overloaded SIP tandem server with finite buffer was developed in [18].

Both for analytical and simulation modeling it is of major importance to know the characteristics of processes that govern the incoming flow(s) and service of messages: the more close these processes resemble real-life behavior of entities of SIP network the more accurate model is. Although there are some results suggesting different SIP traffic models and service time distributions for messages entering SIP servers, amount of existing results is not enough to elaborate unified point of view. This paper contributes to the understanding of statistical properties of message service process in SIP servers, operating in carrier's backbone network. The SIP traffic circulating between two geographical regions through SIP proxy server is captured just before it enters server and right after it leaves server. Then it is analyzed following methodology proposed by authors of [19] for analysis of single-hop packet delay through operational router in Spring IP backbone network. During analysis there were observed packet processing and queueing delays and no transmission delays at output link (due to high speed of the interface). Additionally there were detected very large message delays that we were unable to explain. Having no information on how messages of different types are processes inside SIP server under consideration, assumption was made that one can model SIP server as single node queue with c processor serving messages according to some phase-type distribution (PH distribution) i.e. $\cdot|PH|c$ queue. Here "$\cdot$" means that any analytical model can be used for input flow or queueing system may be fed with SIP trace instead. Phase-type distribution is expected to embrace possible complex processing that may take place inside SIP server. One of the results is that waiting time of a SIP-I message in the queue is not exponentially distributed. The authors managed to fit measured waiting time in PH distribution using EM algorithm and ProFiDo tool (see [20]) though the fitting is not absolutely accurate at high values of quantiles (greater than 99.9%). From this experimental observation it follows that service time distribution is also of phase-type (with the same number of phases) due to the fact that waiting time in single queue systems of type $G|PH|c$ is of phase-type (see, e.g. [21]). Given experimental conditions, it is suggested that service time of message of given type equals roughly minimum time which message of such type spends in SIP server. Results of service time distribution fitting show that it can be modeled at acceptable level of accuracy with phase-type distribution.

The paper is organized as follows. In the Section 2 a description of traffic collection procedure and some insight into the traffic nature is given. Then in Section 3 delay measurements with step by step analysis of elements which contribute to single hop delay are presented. Conclusion contains short overview of obtained results and plans for further research are shown in Section 4.

2. Traffic Collection and Data Description

We consider the same fragment of telecommunication operator network depicted in Fig. 1 as in paper [22]. Traffic aggregation happens on regional access network using Signalling System No. 7 (SS7). As the caller and callee are located in different regions, the traffic goes through the two transit regional nodes – softswitches of the fourth class. Signalling exchange between them is organized by means of the SIP-I protocol. On the traffic route between softswitches SIP proxy server is set for the purpose of logical separation of regional networks.

All the SIP-I traces were captured on SIP proxy server's network interfaces during one week (7 consecutive days, starting from Wednesday, 24 hours per day) by means of span session created on one L2/L3 switch (see Fig. 2). All the transaction were stateful, underlying transport protocol was UDP. In order to capture traffic going in and out of SIP server, passive traffic monitoring system was implemented

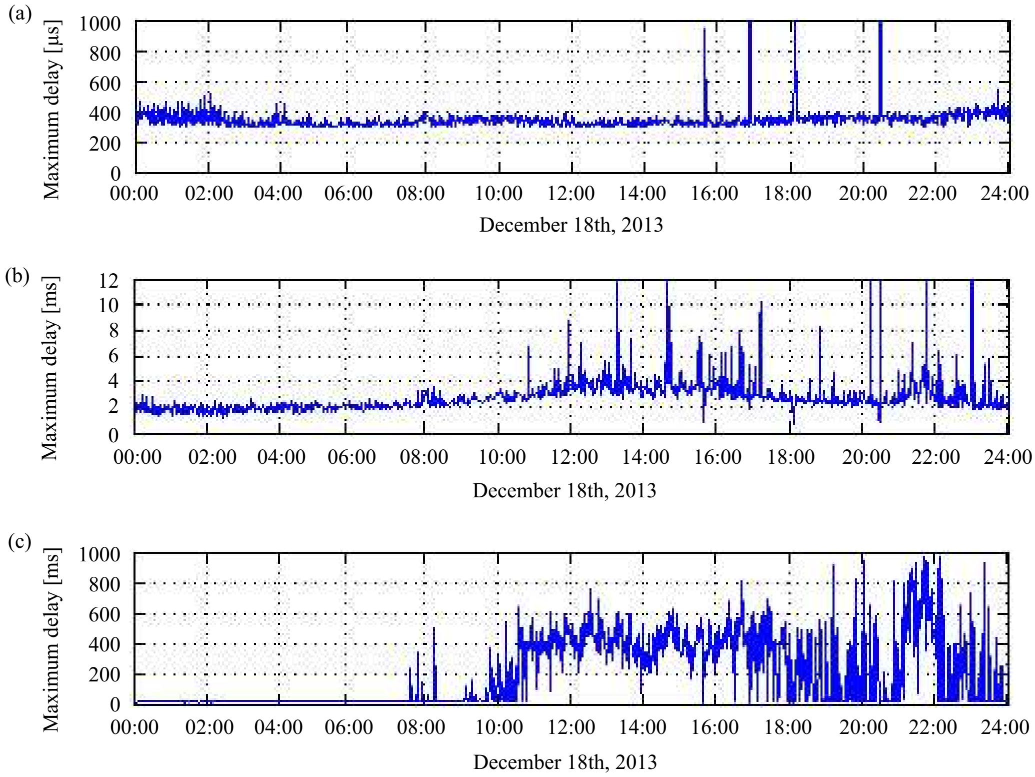

Fig. 2. Minimum (in μs), average and maximum delay (in ms) per minute during workday (24 hours).

which utilized traffic mirroring on nearest to SIP server L3 switches (see Fig. 3 and [23]).

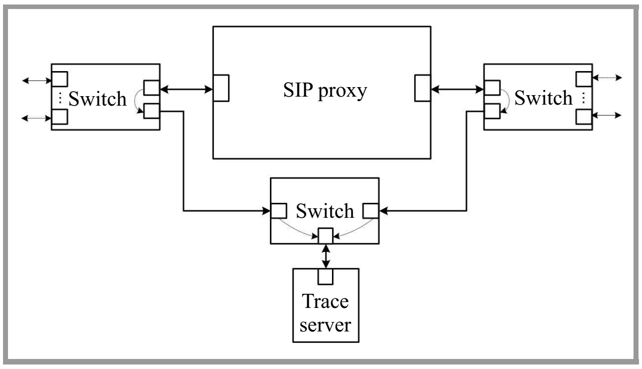

Fig. 3. Layout of monitoring circuit.

The load on the switches during the experiment did not exceed 30–40% of maximum rated capacity, thus providing no significant effect on measurements. The analysis of the flow volume of SIP-I traffic through SIP server has shown that for data capturing it is possible to use standard network interface cards installed on the data collection server. During experiment all types of SIP-I messages, passing through the SIP server for 7 consecutive days were captured.

The problem of time synchronization was solved without special tools – one trace server captured traffic from both SIP server interfaces. Since data volume flowing through SIP server during the measurement period achieved more than 100 gigabytes it was necessary to organize the automatic data backup. It was done using standard Linux operating system tools. The data collection procedure was organized in the same manner as described in [19]. Using Linux utility (tcpdump) libpcap data recording was performed. All information contained in each data packet was recorded.

In order to obtain delay we computed difference between timestamps at the input and output of monitored links. During measurement period no configuration changes, reinstallation of software or replacement of SIP server equipment parts were made. After data collection process had been finished data pre-filtering procedure was carried out. Data arrival, departure stamps and size of each SIP-I message were extracted. One of the major problems encountered in data analysis was unique identification of SIP message after its passage through the SIP server. Unfortunately, the methods of identification by the hash sum based on cyclic redundancy check algorithm (as it is done in [19]), which are widely used to identify the packet after its passage through L2/L3 device, in this case do not work: SIP server changes headers of SIP-I messages.

To solve this problem SIP-I messages are identified using fields which are not subject to change inside SIP server, e.g., session ID, number of messages in transaction, the type of message. By comparing such fields message filtering including deletion of duplicate messages was performed. Technically identification was performed using two tables. The first contained data of SIP-I messages recorded on one interface, the second – on another interface. Each table line values were message fields and arrival timestamps (up to 1 μs). To map messages in the first and the second table the first table was sequentially scanned and checked for matching entries in the second table. As message capturing had been started at both interfaces at the same time it was assumed that the processed message by SIP server would appear at another interface with the serial number value not far from number under which it was registered at the first interface. Thus the search for SIP-I message with given serial number started in the second table from message with the same serial number. The depth of the search had been changed dynamically (depending on the number of messages per time unit) and took at least 20 s before and after the time of messages registration with the serial number from the second table.

Additional difficulty is that SIP-I messages generate additional flows of retransmissions and thus SIP server may resend the same SIP-I message when the corresponding timer fires out. Having matched SIP-I messages we have noticed the following. The organization of the monitoring circuit (see Fig. 3) implies that there are 4 serialization delays that contribute to measured delays and no forwarding/backplane delays in L2/L3 switches because traffic is simply mirrored between interfaces. But in the collected data we have observed delay values much lower than 50 μs. In fact delay values a little higher than 50 μs were present as well and such speed of SIP-I message processing seemed improbable. We found that the border line for delays with improbable values is 300 μs. The percentage of values in measured data below 300 μs is only 0.0006721%. We were unable to explain them and thus excluded them from the data and further analysis. Such delays might have occurred due to errors in identification algorithm or due to incorrect retransmission of SIP-I messages by SIP server.

Notice that SIP-I messages may have size more than 1500 bytes and thus have to fragmented before being processed. Analysis of this issue showed that percentage of fragmented messages in the data does not exceed 0.04%. Only for INVITE messages it is around 2% which is due to the presence of encapsulated big-sized IAM SS7 messages. These message were left out of further analysis as well.

3. Delay Analysis

In this section key observations of measurements of the sojourn (delay) time of SIP-I messages in SIP server are presented. We plot the empirical probability density function of the measured SIP server delay and quantify step-by-step its contributing factors. The outcome of this analysis is the proposal of analytical models for different ingredients of the delay.

3.1. General Observations

Figure 2 shows the minimum, average and maximum values of SIP server delay observed each one minute interval during 24 hours of December 18th 2013. The simplest statistical analysis showed that each day there appear critical outliers (single abnormal high delay values). For example, on the 18th of December 2013 one could observe delay values of 3.231 s, 5.952 s, 6.571 s, 12.102 s for 183 message; on the 19th of December 2013 delay values 35.702 s., 15.213 s, 12.642 s and 10.519 s of 183 message and delay values 6.501 s, 4.499 s of BYE message passing through SIP server were registered. One can see such outliers in Fig. 2. Analysis of two-week data showed that such delays appear more often during workdays than during weekends but total number of outliers per day does not usually exceed 10-30. Identification of exact reasons why it happens, i.e., errors in forwarding algorithms, switching of CPU to some other intensive task, output blocking, memory locks, poor scheduling did not give any results and in further analysis such values were excluded from the data.

Aside from outliers minimum delay is quite stable throughout all the day. The minimum delay corresponds to the minimum amount of time SIP-I message needs to go through SIP server. It is tightly bounded and bounds depend on the time of the day. This suggests that there is at least one packet that experiences no queuing in each one minute interval. The average delay varies more irregularly and is directly related to SIP server utilization. It increases from early morning as call rate increases and drops down in the evening. Maximum delays are much larger and more variable than the average delays. In the middle of the day maximum delay remains above 240 ms and in the evening shows regular spikes of hundreds of milliseconds reaching up 931 ms. Delays during the weekend are much more stable. The average delay is almost uniformly bounded by 4 ms, maximum observed delay is 792 ms.

3.2. Step-by-step Analysis

In Fig. 4 one can see empirical probability density function of SIP-I message delay, measured within monitoring

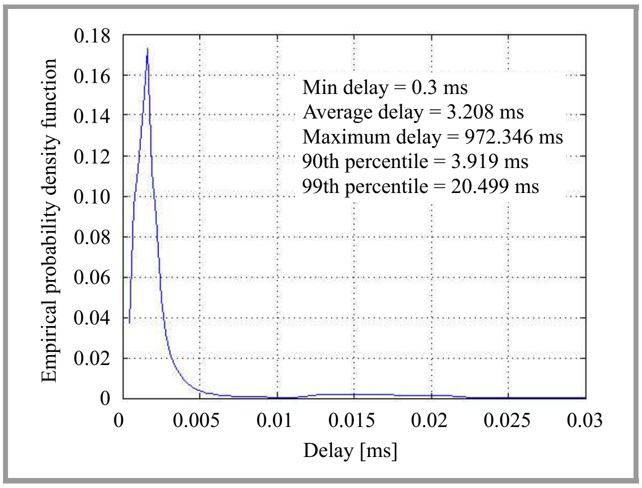

Fig. 4. Empirical probability density function of SIP-I messages measured delay (this delay includes delay in SIP server and other processing delays in switches and network interfaces).

circuit. Average delay value is around 3 ms, whereas 99% of all SIP-I messages experience delay of less that 21 ms and maximum observed delay is almost 1 s. There is only one distinct peak at the beginning of density function and one small but wide peak in the middle. We now try to conjecture what is this small peak related to. In Fig. 5 relative frequency distribution of SIP-I message sizes is plotted. This distribution has many peaks but one can clearly distinguish two groups of message sizes: 375–675 bytes and 1000–1500 bytes (more than 99.5% of SIP-I messages fall in these groups). Now we group SIP-I messages in these two groups and separately plot the empirical density function of the delay experienced by SIP-I messages in the given group (see Fig. 6). One can see that small wide peak is present again for SIP-I messages of size 375–675 bytes but is absent for SIP-I messages of size 1000–1500 bytes. This suggests that the delay peak is related to message type but not size (as it is shown to be in IP backbone router, see [19]) and one should look for corresponding message types only among those messages which have size 375–675 bytes. Thus from this group there were selected those message types which suffer delay more than 10 ms and surprisingly the most frequent message type was 200 OK of size 391–399 bytes. Greater sizes of 200 OK messages do also occur in the whole data set but their delays remain lower. In Fig. 7 one can see empirical probability density function of 200 OK message (size 391–399 bytes) and all other SIP-I messages. It is clear now that small wide peak in Fig. 4 is related to 200 OK message of specified size.

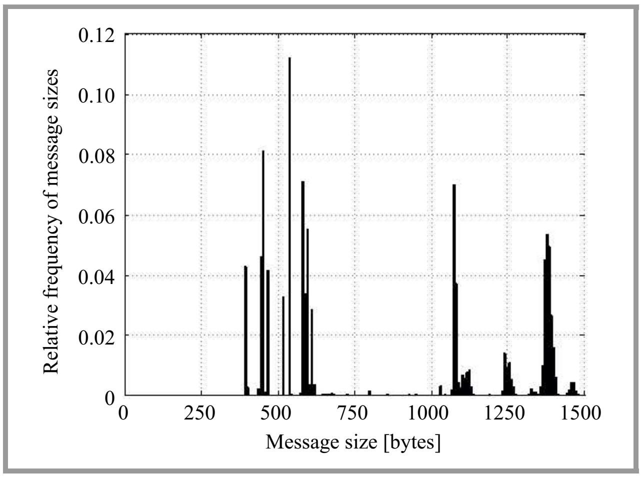

Fig. 5. Relative frequency distribution of SIP-I message sizes.

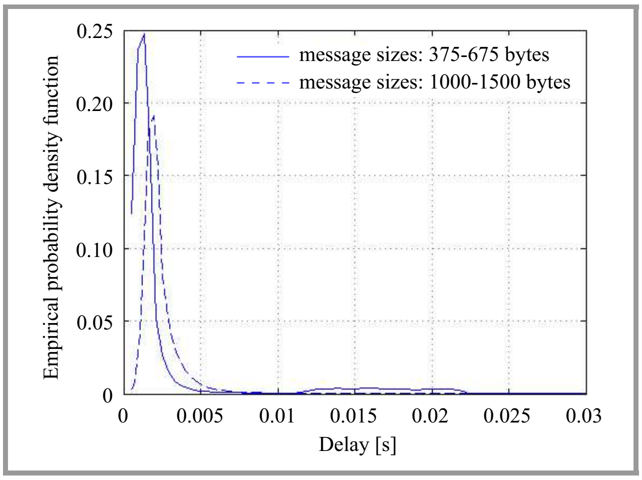

Fig. 6. Empirical probability density function of delay given that size of SIP-I message is from one of two groups: 375–675 bytes and 1000–1500 bytes.

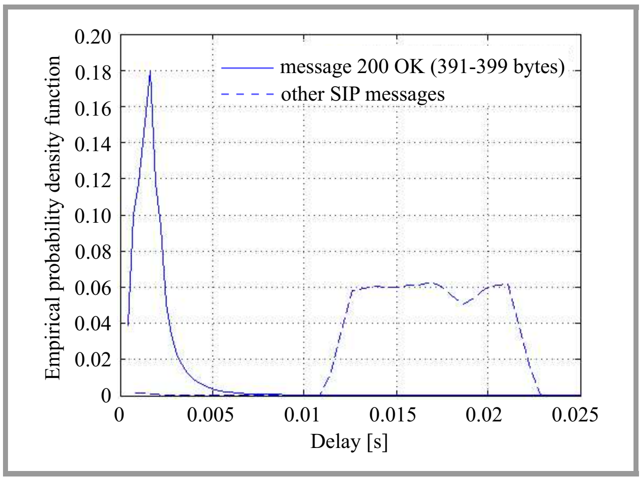

Fig. 7. Empirical probability density functions of delay of 200 OK message (size 391–399 bytes) and other SIP-I messages.

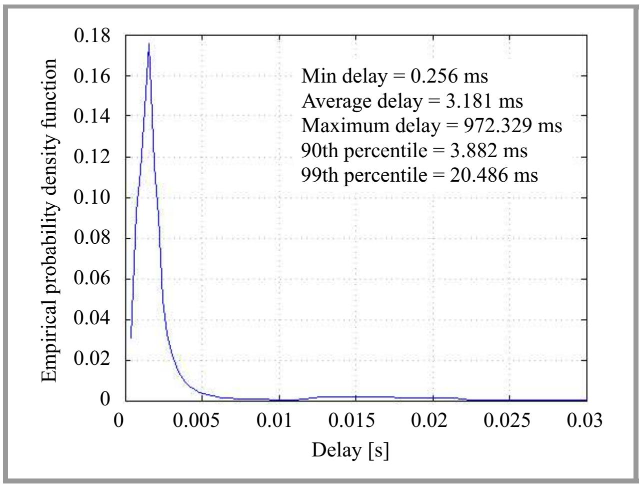

Fig. 8. Empirical probability density function of SIP-I messages transit time through SIP server.

Now we proceed to identify and quantify the factors that contribute the same amount of delay for messages of the same size. The main cause is serialization delay. Serialization delay is proportional to packet size divided by the speed of the output link. All link rates in the monitoring circuit are 1 Gb/s. Although serialization delay of the maximum Ethernet packet size (1500 bytes) is 12 μs we have to take it into account because there are 4 serialization delays.

We refer to the difference between measured delay and all serialization delays as SIP server transit time. The empirical probability density function of transit time is plotted in Fig. 8.

As expected small wide peak in the distribution is still present and high peak is almost not affected. When SIP-I message arriver at SIP server it has to go through network stack and some common and message-specific processing at application level and then again though network stack (see example, e.g. in [15]). This operation takes some time which, when SIP server is under-loaded, indicates minimum transit time experienced by messages (depending either on their size or type or both). We plot the minimum value of the SIP server transit time for each SIP-I message type in Fig. 9.

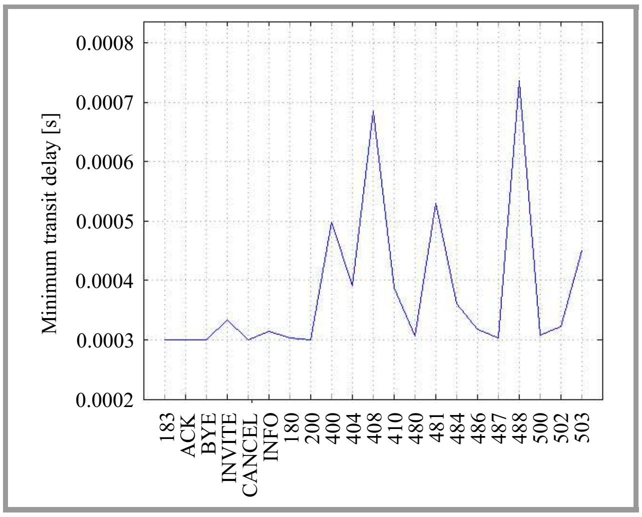

Fig. 9. Minimum transit delay per message type.

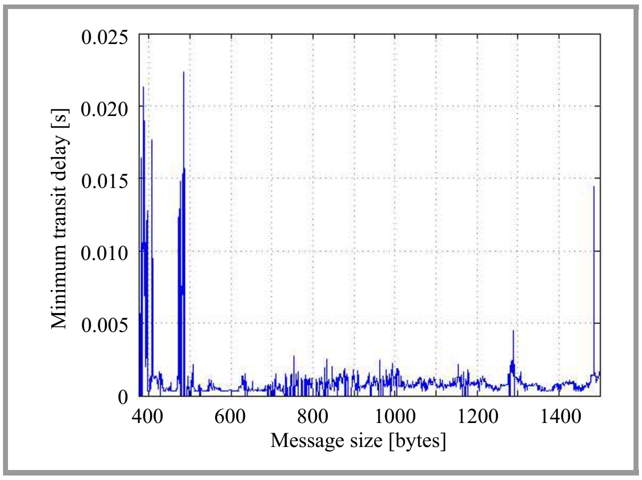

Fig. 10. Minimum transit delay per message size.

One can see that minimum transit time per message type is very low and does not give real picture of minimum delay because each type of message may have very different sizes, carry different information in headers and thus delay should vary significantly. Here it is appropriate to calculate delay characteristics of each type of SIP-I messages. There are a total of 21 type of messages in the data. Basic statistical characteristics per each message type are given in Table 1. It is worth noticing that for 200 OK message average delay differs much from the delay observed in Fig. 7 which is due to the fact that 200 OK messages of other sizes are processes most of the times much faster. In Fig. 10 the minimum transit delay per each unique packet size is plotted.

Table 1
Transit delay (in seconds) per type of SIP message

Message type	Average	Standard deviation
ACK	0.001541	0.001864
BYE	0.001445	0.002541
INVITE	0.002414	0.001977
CANCEL	0.001447	0.001680
INFO	0.008975	0.009001
180	0.002334	0.001870
183	0.004427	0.026801
200	0.005735	0.008445
400	0.001201	0.000388
404	0.001742	0.001696
408	0.003677	0.005602
410	0.001706	0.001829
480	0.001728	0.001754
481	0.001291	0.000914
484	0.001492	0.001420
486	0.001694	0.001681
487	0.001665	0.001612
488	0.001771	0.001671
500	0.001656	0.001656
502	0.001609	0.001575
503	0.001622	0.001688

It can be seen that no simple relationship between two metrics exist. Based on thorough analysis of SIP-I traces we are inclined to consider this minimum transit delay as SIP-I message pure service time. Its empirical probability density is depicted in Fig. 11. Possible explanation of such unstable minimum delays may be the fact that, according to [24], header fields of each new request may have any additional overheads, optional header fields specific to the method and their processing may introduce additional delay. Now if one subtracts this minimum from transit delay for each message in the data, one obtains amount of time each message has to wait inside SIP server. The new empirical probability density function is presented in Fig. 12.

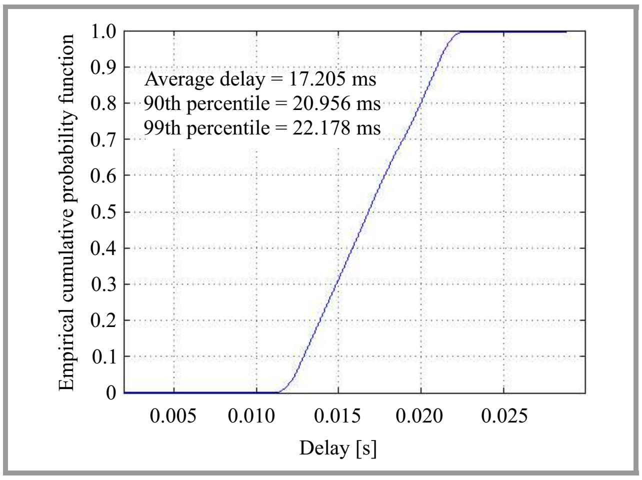

Fig. 11. Empirical cumulative distribution function of the minimum transit delay (service time).

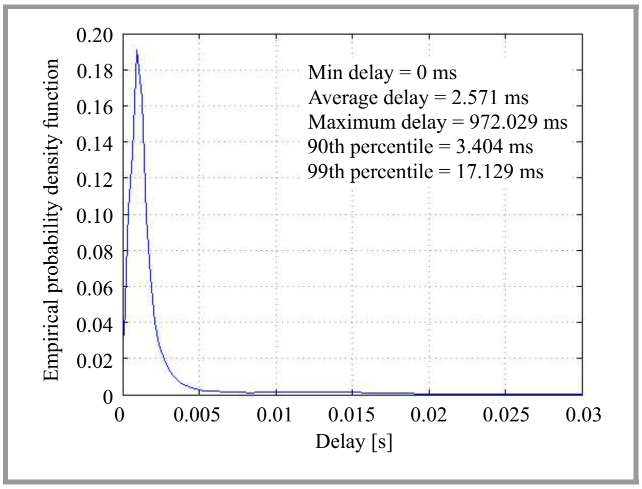

Fig. 12. Probability density of the waiting time in SIP server.

Small peak in the middle related to 200 OK message of 391–399 bytes sizes has disappeared. The distribution is characterized by average low delays, most of the messages exhibit delays close to average (90th percentile is around 3 ms), whereas maximum delay remain high (almost reaches 1 s). In Fig. 13 cumulative distribution function of the waiting time in SIP server is presented. A key observation is that the tail of the delay distribution is long accounting for the presence of high delays. However link utilization when large delays were observed was not very high and therefore long delays must not be due to queueing at the output link. The percentage of messages for which waiting time exceeds 17 ms (99th percentile of all delays) is only 2.1%. We were unable to find possible explanations for such high values. It is not that only one or two types of messages exhibit such long delays. For almost each message type there are packets with high delay values. Once these values have been removed from the data set, we found that statistics of the distribution have changed (average delay decreased to 1.7 ms, 90th percentile decreased to 2.9 ms, 99th percentile dropped to 13 ms). These outliers could be attributed to issues common for routers (e.g. memory locks, poor scheduling etc., see [19]) or SIP-I protocol specific issues but these conjectures were not checked.

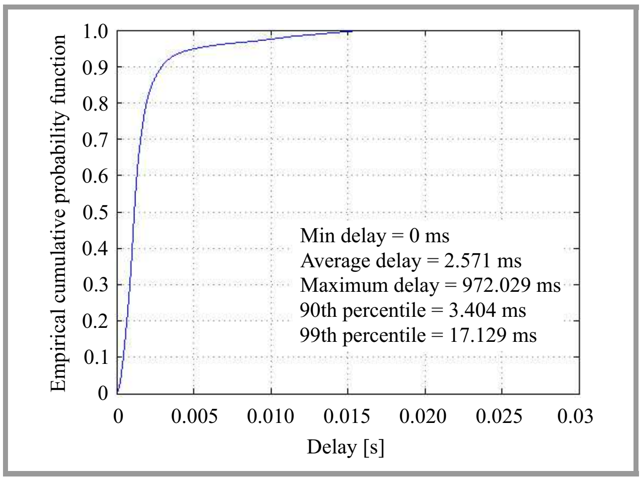

Fig. 13. Empirical cumulative distribution function of the waiting time.

Analysis of the data showed that the distribution of the waiting time is not exponential (squared coefficient of variation equals 26.83). From this it follows that possible queueing model with single (or several but identical) server and exponential service times are not adequate for the considered case. Following the assumption concerning SIP server model stated in Section 1, we tried to fit message waiting times in PH distribution using EM algorithm implemented in [20]. Quantiles of the waiting time estimated from data and of simulated fitted acyclic PH with 15 states are shown in Fig. 14. The simulation was performed using KPC-Toolbox (see [25]).

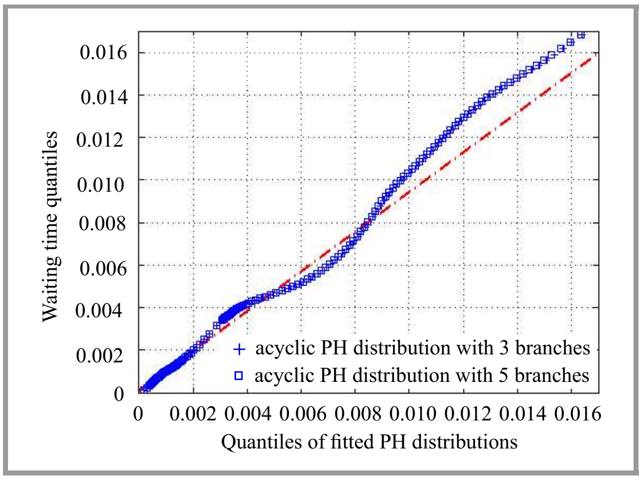

Fig. 14. Q-Q plot of the waiting time and simulated data, using fitted continuous acyclic PH distribution with 15 phases.

The fact that waiting time can be approximated by PH distribution, suggests that in the queueing model of SIP server $\cdot|PH|c$ number of processors is $c = 1$ (see e.g. [21]) with

PH distribution of service times with the same number of phases. In Fig. 15 one can see quantile-quantile plot of the minimum transit delay and simulated data, using fitted continuous acyclic PH distribution with 15 phases.

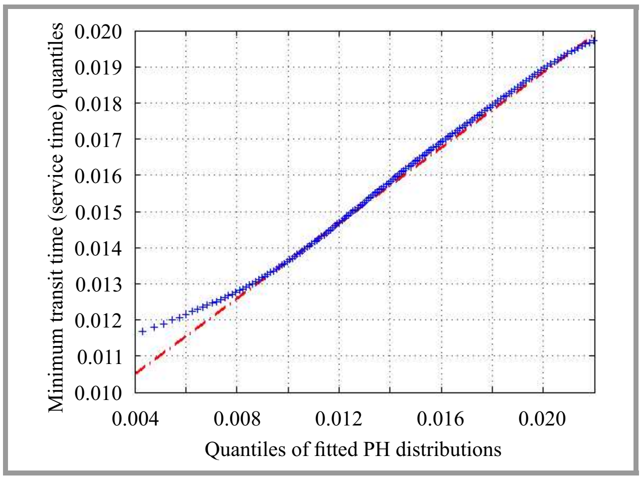

Fig. 15. Q-Q plot of the minimum transit time (service time) and simulated data, using fitted continuous acyclic PH distribution with 15 phases.

With respect to the waiting time one may observe many outliers at high quantiles (note that 99% of all waiting times is below 3.4 ms) and fitting is not absolutely accurate. This may be due to the fact that 2.1% of SIP-I messages suffer extremely high delays and number of such values within the fitted data is not enough for accurate estimation of PH parameters. The same observation is true for minimum transit time (service time) fitting but at low values of quantiles.

4. Conclusion

Paper presents analysis of single-hop delay of SIP-I message going through SIP proxy server operating in carrier's backbone network. According to our results 99% of all SIP-I messages experience less than 21 ms of sojourn delay. We have observer a small number of messages (2.1% of the data volume) which suffer from significantly larger delays and we were unable to find the reason for that. Analysis of the waiting time (queueing delay) and minimum transit time through SIP server revealed that both can be approximated by acyclic continuous phase-type distributions. But due to outliers fitting is not so accurate and thus requires additional validation by virtue of analytical models (e.g. $MMPP|PH|c$ queue).

References

[1] B. Materna, "Threat mitigation for VoIP", in *Proc. 3rd Ann. VoIP Secur. Worksh.*, Berlin, Germany, 2006.

[2] E. Nahum, J. Tracey, and C. Wright, "Evaluating SIP server performance", in *Proc. Int. Conf. Measur. Model. Comp. Syst. ACM SIGMETRICS 2007*, San Diego, CA, USA, 2007, pp. 349–350.

[3] M. Ohta, "Overload protection in a SIP signaling network", in *Proc. Int. Conf. Internet Surveill. Protect. ICISP'06*, Cote d'Azur, France, 2006.

[4] R. Garroppo, S. Giordano, S. Spagna, and S. Niccolini, "Queueing strategies for local overload control in SIP server", in *Proc. IEEE Global Commun. Conf. GLOBECOM 2009*, Honolulu, HI, USA, 2009.

[5] A. Abdelal and W. Matragi, "Signal-based overload control for SIP servers", in *Proc. 7th Ann. IEEE Consumer Commun. Netw. Conf. IEEE CCNC 2010*, Las Vegas, NV, USA, 2010, pp. 1–7.

[6] V. Hilt and I. Widjaja, "Controlling Overload in networks of SIP servers", in *Proc. 16th IEEE Int. Conf. Netw. Protoc. ICNP 2008*, Orlando, FL, USA, 2008, pp. 83–93.

[7] E. Noel and C. R. Johnson, "Novel overload controls for SIP networks", in *Proc. 21st Int. Teletraff. Congr. ITC 21*, Paris, France, 2009.

[8] Y. Hong, C. Huang, and J. Yan, "A comparative study of SIP overload control algorithms", in *Network and Traffic Engineering in Emerging Distributed Computing Applications*, J. Abawajy, M. Pathan, M. Rahman, A. K. Pathan, and M. M. Deris, Eds. IGI Global, 2012, pp. 1–20.

[9] D. Sisalem, "SIP overload control: Where are we today?", in *Trustworthy Internet*, L. Salgarelli, G. Bianchi, and N. Blefari-Melazzi, Eds. Springer, 2011, pp. 273–287.

[10] Hong Y., C. Huang, and J. Yan, "Modelling chaotic behaviour of SIP retransmission mechanism", *Int. J. Paral. Emerg. Distrib. Syst.*, no. 2, pp. 95–122, 2013.

[11] V. Gurbani, L. Jagadeesan, and V. Mendirittam, "Characterizing session initiation protocol (SIP) network performance and reliability", in *Proc. 2nd Int. Service Availabil. Symp. ISAS 2005*, Berlin, Germany, 2005, pp. 196–211.

[12] S. Subramanian and R. Dutta, "Comparative study of M/M/1 and M/D/1 models of a SIP proxy server", in *Proc. Australasian Telecommun. Netw. Appl. Conf. ATNAC 2008*, Adelaide, Australia, 2008, pp. 397–402.

[13] S. Subramanian and R. Dutta, " Measurements and analysis of M/M/1 and M/M/c queueing models of the SIP proxy server", in *Proc. 18th Int. Conf. Comp. Commun. Netw. ICCCN 2009*, San Francisco, CA USA, 2009.

[14] S. Subramanian and R. Dutta, "Performance and scalability of M/M/c based queueing model of the SIP proxy server – a practical approach", in *Proc. Australasian Telecommun. Netw. Appl. Conf. ATNAC 2009*, Canberra, Australia, 2009.

[15] R. Krishnamurthy and G. Rouskas, "Evaluation of SIP proxy server performance: Packet-level measurements and queuing model", in *Proc. Int. Conf. Commun. IEEE ICC'03*, Budapest, Hungary, 2013, pp. 2326–2330.

[16] Y. Gaidamaka, A. Pechinkin, R. Razumchik, K. Samouylov, and R. Sopin, "Analysis of M|G|1|R queue with batch arrivals and two hysteretic overload control policies", *Int. J. Appl. Mathem. Comp. Sci.*, vol. 24, no. 3, pp. 519–534, 2014.

[17] Y. Hong, C. Huang, and J. Yan, "Analysis of SIP retransmission probability using a Markov-Modulated poisson process model", in *Proc. IEEE/IFIP Netw. Oper. Manag. Symp. NOMS 2010*, Osaka, Japan, 2010, pp. 179–186.

[18] Y. Hong, C. Huang, and J. Yan, "Modeling and simulation of SIP tandem server with finite buffer", *ACM Trans. Model. Comp. Simul.*, vol. 21, no. 2, 2011.

[19] C. Diot, C. Fraleigh, S. Moon, K. Papagiannaki, and P. Thiran, "Measurement and analysis of single-hop delay on an IP backbone network", *IEEE J. Selec. Areas Commun.*, vol. 21, no. 6, pp. 908–921, 2003.

[20] F. Bause, P. Buchholz, and J. Kriege, "ProFiDo – the process fitting toolkit Dortmund", in *Proc. 7th Int. Conf. Quantit. Eval. Syst. QEST 2010*, Williamsburg, VA, USA, 2010.

[21] S. Asmussen and J. Moller, "Calculation of the steady state waiting time distribution in GI/PH/c and MAP/PH/c queues", *Queueing Systems: Theory and Appl.*, vol. 37. pp. 9–29, 2001.

[22] P. Abaev, I. Uglov, and R. Razumchik, "Statistical analysis and modelling of SIP traffic for parameter estimation of server hysteretic overload control", *J. Telecommun. Inform. Technol.*, no. 4, pp. 22–31, 2013.

[23] A. Heyde and L. Stewart, "Using the Endace DAG 3.7GF Card With FreeBSD 7.0", Tech. Rep. CAIA 080507A, 2008-05.

[24] "SIP: Session Initiation Protocol", IETF, RFC 3261, 2002-06.

[25] G. Casale, E. Zhang, and E. Smirni, "KPC-toolbox: Simple yet effective trace fitting using markovian arrival processes", in *Proc. 5th Int. Conf. Quantit. Eval. Syst. QEST 2008*, St. Malo, France, 2008, pp. 83–92.

Pavel Abaev received his Ph.D. in Computer Science from the Peoples Friendship University of Russia in 2011. He has been an Associate Professor in the Department of Applied Informatics and Probability Theory of Peoples' Friendship University of Russia. His current research focus is on NGN signalling, QoS analysis of SIP, and mathematical modeling of communication networks.

E-mail: pabaev@sci.pfu.edu.ru
Department of Applied Informatics and Probability Theory
Peoples' Friendship University of Russia
Miklukho-Maklaya st 6
117198 Moscow, Russia

Rostislav Razumchik received his Ph.D. in Physics and Mathematics in 2011. Since then, he has worked as a senior research fellow at the Institute of Informatics Problems of the Russian Academy of Sciences. His current research activities focus on stochastic processes and queuing theory.

E-mail: rrazumchik@ieee.org
Institute of Informatics Problems of RAS
Vavilova st 44-1
119333 Moscow, Russia

Ivan Uglov works at the Department of Information Security, Moscow Technical University of Radio Engineering, Electronics and Automation. His research interests include telecommunication network modeling, simulation and quality of service provisioning.

E-mail: uglov_ivan@mail.ru
Moscow Technical University of Radio Engineering,
Electronics and Automation
Vernadskogo st 78, 21-88
119454 Moscow, Russia