ISSN 0146-4116, Automatic Control and Computer Sciences, 2016, Vol. 50, No. 1, pp. 37–45. © Allerton Press, Inc., 2016.
Published in Russian in Avtomatika i Vychislitel'naya Tekhnika, 2016.

Analysis of Quality-of-Service Metrics in IMS Networks[1]

N. Kulikov
R&D PROTEI, Bol'shoi Sampsonievskii 60A, St. Petersburg, 194044 Russia
e-mail: kulikov@protei.ru

Received June 12, 2015; in final form, September 25, 2015

Abstract—A network constructed according to the IMS architecture consists of different modules that sequentially process signaling messages transferred during providing communication services. Delays that occur when processing signaling messages determine the quality of service for subscribers. International recommendations define the quality-of-service metrics, which include not only the mean, but also the 95% quantile. The paper proposes an approach to analyzing the quality-of-service characteristics. It is assumed that IMS-network nodes work as multi-server queues with general service time distribution.

Keywords: IMS, M/G/k queue, SIP, QoS, cumulant analysis

DOI: 10.3103/S0146411616010077

1. INTRODUCTION

Recently, Next Generation Networks (NGN) has been actively constructed based on the latest international recommendations. There is a strong trend towards the convergence of different networks that has determined the appearance of unique architecture, i.e., the IP Multimedia Subsystem (IMS), which allows a variety of subscribers to access unified telecommunications infrastructure resources, regardless of the type of network access. The most recently published IMS construction concept (Release 7) allows provide a single set of services to all subscribers, including mobile (in the networks 1-4G), wireless, Wi-Fi, Wi MAX and fixed-line subscribers, including access via Ethernet and xDSL.

The modernization of the PSTN or ISDN up to IP-based next-generation network architecture has a number of advantages, including the integration of different subscribers connected via a number of technologies to a single architecture, increased revenues by expanding the range of services based on a single info-communication platform, lowered operating costs for the operator, etc. One of the important tasks for IMS architecture is the support of standardized metrics of quality of service (QoS) level. The main cause of the problem is a delay when processing signaling messages in IMS-network modules.

The rest of the paper is organized as follows. Chapter II introduces a simplified model of the IMS-network core. Chapter III describes QoS metrics based on international recommendations for ISDN and IP networks. Section IV contains an estimate of the algorithm based on a cumulant analysis of QoS metrics.

2. ARCHITECTURE OF IMS NETWORK

IP Multimedia Subsystem concept is open network architecture, introduced for next generation networks. It supports a wide range of services for circuit-switched and packet networks. The IMS has two main signaling protocols: SIP—Session Initiation Protocol [1] for communication sessions management and DIAMETER [2] for interaction with a subscriber database server. SIP is a very flexible protocol, which allows for users to manage their services and to mix media streams within a session. Furthermore, it is universal for all network access types.

The base of an IMS network is the core, which consists of a set of specialized modules that make various functions needed to service subscribers. In simplified form, the structure of the IMS core could be represented as follows [3]. The main functional elements of the core of the IMS network, which are involved in the processes of registration, switching, RTP packet routing, charging calls is Call Session Control Functions (CSCF). It is divided into three types of modules, i.e., Proxy (P CSCF), Session (S-CSCF), and Interrogating (I-CSCF) functions. The Home Subscriber Server (HSS) database is a central repository of information about user profiles and IMS network services.

[1] The article was translated by the authors.

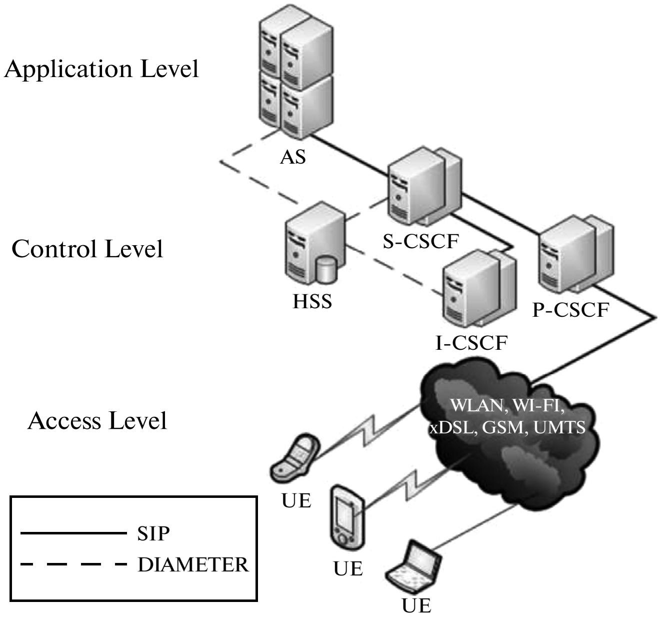

Fig. 1. Simplified scheme of the IMS core.

3. QUALITY OF SERVICE IN IMS NETWORK

3.1. Quality-of-Service Metrics in IMS Network

QoS support is one of fundamental requirement for IMS-network [3]. Fully quality of service analysis in IP-network includes an end-to-end investigation of the parameters taking into consideration both the functional and physical structure of the network.

In Recommendation ITU-T Y.1531 [5], which is devoted to measuring the performance of IP-based networks, the basic services delay is considered to be the main QoS metric. From the point of view of this document, the Call Setup Delay (CSD) is the time interval between connection initiation and confirmation that it has been successfully established. Delays caused by the UE and the time of processing intermediate signal messages are not taken into account.

A more recent Internet Engineering Task Force (IETF) recommendation [6] introduced several metrics that reflect delays associated with the establishment of the connection, i.e., session request delay (SRD), and destruction, i.e., session disconnect delay (SDD). Unfortunately, this recommendation does not introduce numerical characteristics for the proposed indicators.

During the migration from ISDN to networks based on IMS architecture, characteristics of provided services must not be degraded, both in terms of the mathematical expectation and 95% quantile. Therefore, it is appropriate to apply numerical requirements specified in Recommendation E.721 [7] developed for ISDN networks to metrics defined in the recommendations [5, 6]. Metrics should be evaluated taking into account the peculiarities of the exchange of signaling messages in VoIP networks. Thus, IMS networks could be introduced QoS metrics as follows (Table 1).

Table 1. Quality-of-service metrics

Metric	Mean (ms)	95% (ms)
Session Request Delay	3000	6000
Answer signal delay	750	1500
Call release delay	400	600

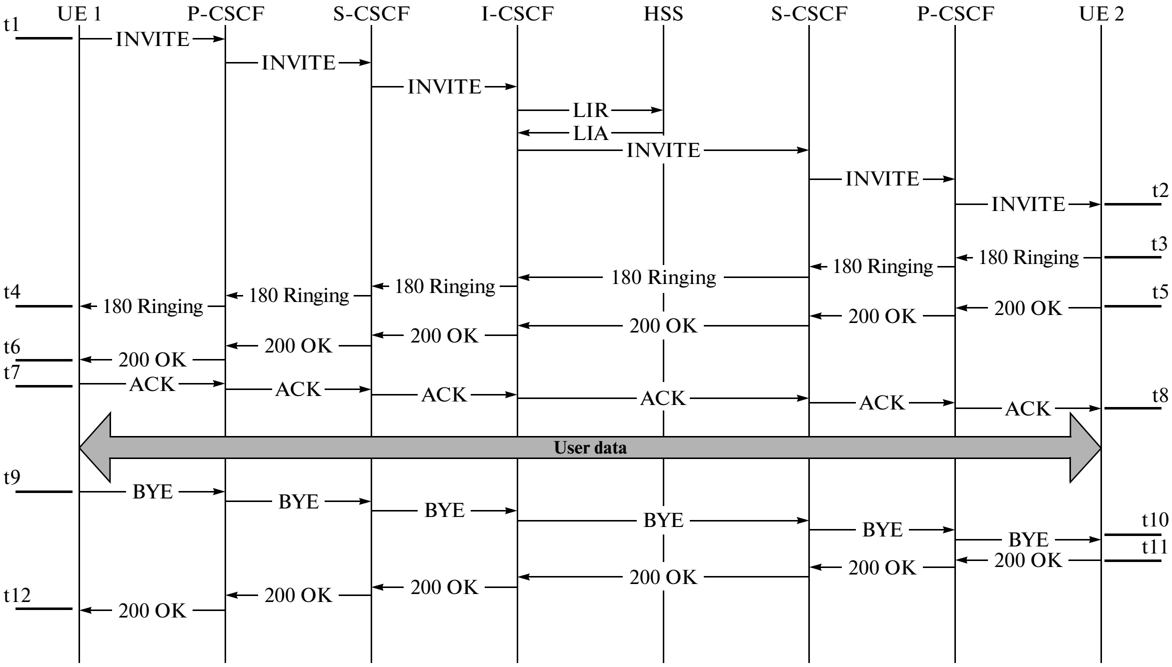

Fig. 2. Services delay in IMS network.

3.2. Services That Provide a Model of Delay Estimation

Let the random variable of SRD in the IMS network be referred to as T_{SRD}. In the case of a successful call, it is determined by the duration of the transmission of the INVITE message from calling UE to receive a 180 Ringing message, except the duration of processing messages in the UEs themselves (Fig. 2).

$$T_{\text{SRD}} = (t4 - t1) - (t3 - t2). \qquad (1)$$

This random consists of number of processing delays of different messages in each network node ($T_{\text{P-CSCF}}$, $T_{\text{I-CSCF}}$, $T_{\text{S-CSCF}}$, and T_{HSS}). At the same time, we assume that the transmission of signaling messages delay within the IMS core is a negligible quantity because the IMS core is usually situated within one site and modules are connected via high-speed channels.

Thus, the IMS core could be considered to be a multiphase queuing system, which consists of a set of independent queues that consequently signal messages with a certain speed. The value of standardized QoS metrics will be determined by the total delay in each phase of the system.

For example, the value of T_{SRD} for successful connection is defined as follows:

$$T_{\text{SRD}} = 2T^{(\text{P-CSCF})}_{\text{INVITE}} + 2T^{(\text{S-CSCF})}_{\text{INVITE}} + 2T^{(\text{I-CSCF})}_{\text{INVITE}} + T^{(\text{I-CSCF})}_{\text{LIA}} + T^{(\text{I-CSCF})}_{\text{LIR}} + 2T^{(\text{P-CSCF})}_{\text{Ringin}} + 2T^{(\text{S-CSCF})}_{\text{Ringin}} + 2T^{(\text{I-CSCF})}_{\text{Ringin}}. \qquad (2)$$

To estimate these random variables, it is necessary to assess the parameters of incoming signaling messages stream and determine characteristics of the core modules of IMS network.

To evaluate traffic characteristics in the next generation network, a segment of IP-telephony of the Moscow City Telephone Network (MGTS) was studied. This segment is constructed in accordance with the IMS architecture, and serves more than 2.5 million subscribers. [8] In the research, it was studied parameters of signaling traffic, which transmitted to the IMS-core. The most interesting were INVITE message, which initiate establishment of new connections, and the total sum of all messages streams, corresponding to user sessions management.

In accordance with a variety of classic researches, from Erlang [9], we see in Fig. 3a that INVITE messages flow is exponential, because associated with call attempts. For the entire set of signaling messages exponential character is not obvious, however, empirical evidence shows that the coefficient of variation for the entire set of signaling SIP messages, close to one that leads to the assumption of exponential distribution of the flow of all incoming messages.

Several researches [10, 11] indicates that the service time distribution in IMS network nodes is not exponential. As shown in [12], for different types of signaling SIP messages, IMS-core modules (primarily

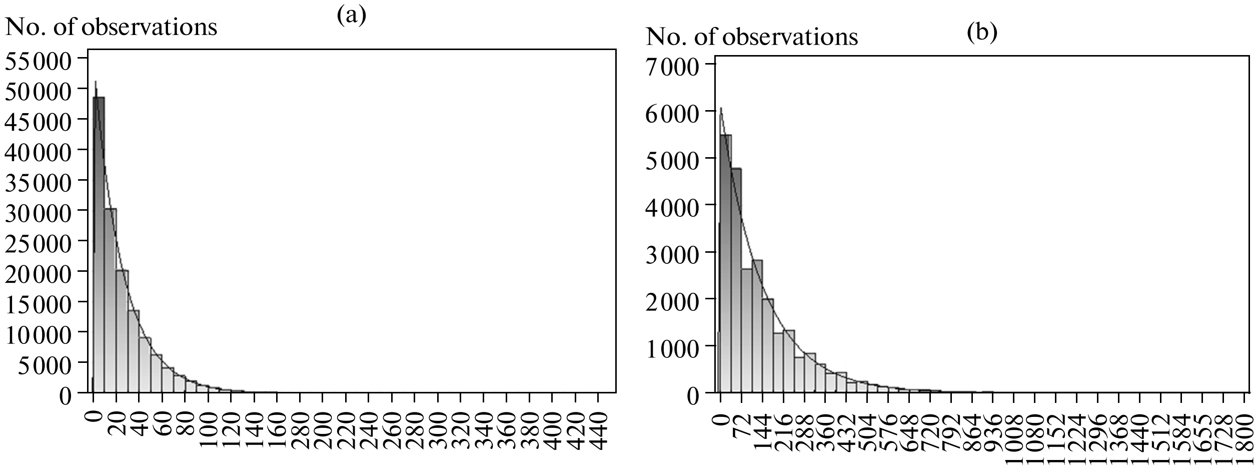

Fig. 3. PDF between incoming messages duration to IMS-core: (a) INVITE messages only, (b) all call management messages.

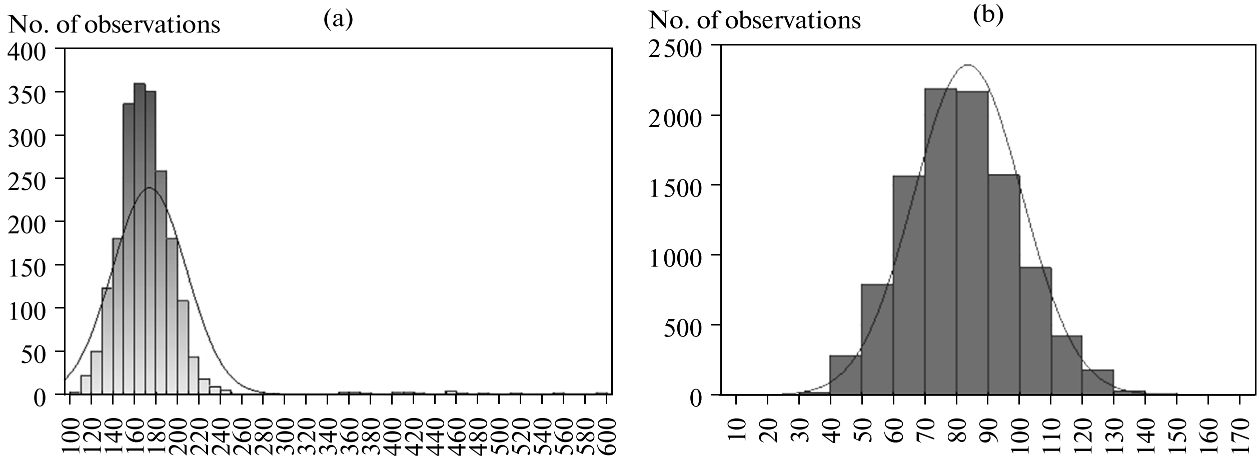

Fig. 4. PDF response time delay in IMS-core: (a) INVITE messages, (b) BYE messages.

CSCF modules) make a finite number of operations. Thus, it could be assumed that the service time in core nodes is distributed on a finite time interval. As the service time distribution function, it is convenient to choose generalized beta distribution, as its coefficient of variation can vary on a wide range.

However, the service time characteristics for HSS module differ from other elements of the IMS. Since this module is a database, processing time has slowly decreasing, so-called “heavy-tailed”, distribution function. For analytical study it may be selected Weibull–Gnedenko distribution as a law of service delay in HSS. The influence of the subscriber database on request delay in the IMS-network can be clearly seen in Fig. 4, which shows the histogram of the response delay of SIP messages for real IMS network in MGTS. We can see the “heavy tail” for processing INVITE messages, which other signaling SIP messages do not have, since they does not require processing in the HSS.

Another feature of the IMS-core module is that the actual hardware used to construct them, based on a multiprocessor technologies such as ATCA or AXIe [13]. Assuming that, after each network node is passed, the characteristics of message flow does not change and we can define each IMS-network node in terms of Kendall classification as an $M/G/k$ queue. The validity of the assumption is confirmed by real observations in the MGTS network; in particular, it is noted that the SIP messages traffic is exponential at both the input and output of different service modules.

From the classical probability theory point of view, if we want to determine the standardized quality of services metrics, it is necessary to determine the distribution function of the response time delay for signaling messages at IMS-network nodes, then determine the overall distribution of the overall random variable (for example, by the convolution mechanism), then find the mean and 95% quantile. Unfortunately, for $M/G/k$ systems, this algorithm cannot be applied due to the lack of accurate probability parameters formulas for this type of system. Therefore, we will resort to the cumulant analysis machine [14, 15], which allows one to find any of the standardized metrics.

Cumulants could be expressed in terms of moments as follows [15]:

$$\begin{gathered}\chi_1 = E[X^1], \quad \chi_2 = E[X^2] - E[X^1]^2, \quad \chi_3 = E[X^3] - 3E[X^1]E[X^2] + 2E[X^1]^3, \\ \chi_4 = E[X^4] - 3E[X^2]^2 - 4E[X^1]E[X^3] + 12E[X^1]^2E[X^2] - 6E[X^1]^4.\end{gathered} \tag{3}$$

The formulas of the first four cumulants have the following meanings. The first two are the mead and dispersion. The third cumulant can be called skewness, and the fourth is kurtosis [15]. Cumulant analysis is often a more convenient mathematical technique than the analysis of the moments.

The cumulants set $\chi_1, \chi_2, \chi_3, ..., \chi_k, ...$ could serve as an identical representation of the random variable probability distribution. An important property of the cumulants is that the cumulants of the random variable obtained as the sum of two independent random variables is equal to the sum of cumulants of initial random variables. The distribution function of the service delay can be expressed in terms of total k-th order cumulant [15], which are expressed as the sum

$$\chi_k^{\eta} = \sum_{j=1}^{N} \chi_k^{\zeta_j}, \tag{4}$$

where $\eta = \xi_1 + \xi_2 + ... + \xi_N$ is the summation of cumulants of the waiting time $W(t)$ and the service time $B(t)$ distributions for all messages and nodes involved in service providing. Expression (4) could only be applied to investigating the IMS network if all modules are mutually independent. For $M/M/k$ tandem queues, it is shown [16] that response time delays at each node are not mutually independent. Despite this, in order of approximation, when the total response time parameters estimate each unique queue itself could be considered independently [17]. We extend this assumption to the system with an arbitrary service law and check this approximation at the end of the article.

4. ESTIMATES OF CUMULANTS OF THE RESPONSE TIME DELAY FOR *M/G/k* QUEUES

To estimate the characteristics of the response time delay, it is necessary to find cumulants of the waiting time distribution $W(t)$, and service time distribution $B(t)$. It is not a difficult task to determine cumulants for the $B(t)$ function, for known distribution law, as most laws have well-known formulas for calculating moments. However, estimation of the waiting time distribution $W(t)$ characteristics for $M/G/k$ queues is not trivial.

Since there are many approximate solutions for the mean waiting time [18–21], there is no exact definition of the probability distribution for this class of queues. The question of determining high-order moments and, therefore, the 95% quantile of the waiting time distribution is badly reviewed in the literature.

According to [22], the waiting time moments are associated with moments of the queue length by the following relation:

$$E[W^r] = \frac{E[Q(Q-1)\dots(Q-r+1)]}{\lambda^r}, \quad r = 1, 2, \dots, \tag{5}$$

where Q is the conditional number of messages, waiting in the queue when all service modules are busy. A paper [23] proposes an approximation of moments generating function $P_q(z)$ for an unconditional number of messages in the queue.

The approximation is based on the following two assumptions:

1. If, at the moment when message leaves the queue, n messages are left behind in the system with $1 \le n \le k$, then the time until the next message leaves is distributed as $(B_1^e, \dots, B_n^e)$, where $B_1^e, \dots, B_n^e$ are independent random variables with the same probability distribution function $B_e(t)$ as follows:

$$B_e(t) = \mu \int_0^t (1 - B(x))dx, \quad t \ge 0.$$

2. If, at the moment when message leaves the queue, n messages are left behind in the system with $n > k$, then queue $M/G/k$ could be considered as $M/G/1$, where the service unit works k times faster.

Assuming that the delay in waiting buffer probability π_w is the same for $M/G/k$ and $M/M/k$ queues, work [23] gives generating function of queue length for $M/G/k$ as follows:

$$P_q(z) = \lambda p_{k-1}^{\exp}\frac{\alpha(z)}{1-\lambda\beta(z)}, \tag{6}$$

where

$$\alpha(z) = \int_0^{\infty}(1-B_e(t))^{k-1}(1-B(kt))e^{-\lambda(1-z)t}dt, \quad \beta(z) = \int_0^{\infty}(1-B(kt))e^{-\lambda(1-z)t}dt.$$

The probability that there are $k - 1$ messages in the system could be easily obtained based on the assumption that the probability of a processing delay (which corresponds to the presence in queue exactly k messages) is as follows [23]:

$$\pi_w = \frac{\rho}{(1-\rho)}p_{k-1}. \tag{7}$$

At the same time [24],

$$\pi_w = \frac{(k\rho)^k}{k!(1-\rho)}\left[\sum_{j=0}^{k-1}\frac{(k\rho)^j}{j!}+\frac{(k\rho)^k}{k!(1-\rho)}\right]^{-1}. \tag{8}$$

Therefore we can rewrite expression (5) as follows:

$$P_q(z) = \pi_w\mu k(1-\rho)\frac{\alpha(z)}{1-\lambda\beta(z)}. \tag{9}$$

Thus, we need to find the queue length moments. It is necessary to calculate the derivatives of expression (9) at point $z = 1$. Then, based on the data given in [23, 25], we can assume that the moments of the queue length distribution of $M/G/k$ queue refer to corresponding queue length distribution moments of the $M/M/k$ queue, as well as the moments of the waiting-time distribution, i.e., $\frac{E[Q]}{E[Q_{\exp}]} = \frac{E[W]}{E[W_{\exp}]}$; $\frac{E[Q(Q-1)]}{E[Q(Q-1)_{\exp}]} = \frac{E[W^2]}{E[W^2_{\exp}]}$, etc. Then, using expression (5), the following relationships for the waiting time moments of the $M/G/k$ queue were obtained:

$$\frac{E[W]}{E[W_{\exp}]} = \left(\frac{\lambda(1-\rho)}{\rho}\gamma_1+\frac{\rho}{2}(1+c_S^2)\right), \tag{10}$$

$$\frac{E[W^2]}{E[W^2_{\exp}]} = \left(\frac{\lambda^2(1-\rho)^2}{\rho^2}\gamma_2+\frac{\lambda(1-\rho)}{\rho}(1+c_S^2)\gamma_1+\frac{\rho^2}{4}(1+c_S^2)^2+\frac{\lambda^3(1-\rho)E[B^3]}{6k^3\rho^2}\right), \tag{11}$$

$$\begin{aligned}\frac{E[W^3]}{E[W^3_{\exp}]} = &\left(\frac{\lambda^3(1-\rho)^3}{2\rho^3}\gamma_3+\frac{\lambda^2(1-\rho)^2}{2\rho}(1+c_S^2)\gamma_2+\left(\frac{\lambda(1-\rho)\rho}{4}(1+c_S^2)^2+\frac{\lambda^4(1-\rho)^2E[B^3]}{6k^3\rho^3}\right)\gamma_1\right.\\ &\left.+\frac{\lambda^3(1-\rho)E[B^3]}{6k^3\rho}(1+c_S^2)+\frac{\lambda^4(1-\rho)^2E[B^4]}{24k^4\rho^3}(1+c_S^2)+\frac{\rho^3}{8}(1+c_S^2)^3\right),\end{aligned} \tag{12}$$

$$\begin{aligned}\frac{E[W^4]}{E[W^4_{\exp}]} = &\left(\frac{\lambda^4(1-\rho)^4}{6\rho^4}\gamma_4+\frac{\lambda^3(1-\rho)^3}{4\rho^2}\gamma_3+\left(\frac{\lambda^2(1-\rho)^2}{4}(1+c_S^2)+\frac{\lambda^5(1-\rho)^3E[B^3]}{6k^3\rho^4}\right)\gamma_2\right.\\ &+\left(\frac{\lambda(1-\rho)\rho^2}{8}(1+c_S^2)^3+\frac{\lambda^4(1-\rho)^2E[B^3]}{6k^3\rho^2}(1+c_S^2)+\frac{\lambda^5(1-\rho)^3E[B^4]}{24k^4\rho^4}\right)\gamma_1+\frac{\rho^4}{16}(1+c_S^2)^4\\ &\left.+\frac{\lambda^3(1-\rho)E[B^3]}{8k^3}(1+c_S^2)^2+\frac{\lambda^6(1-\rho)^2E^2[B^3]}{36\rho^4k^6}+\frac{\lambda^4(1-\rho)^2E[B^4]}{24\rho^2k^4}(1+c_S^2)+\frac{\lambda^5(1-\rho)^3E[B^5]}{120\rho^4k^5}\right),\end{aligned} \tag{13}$$

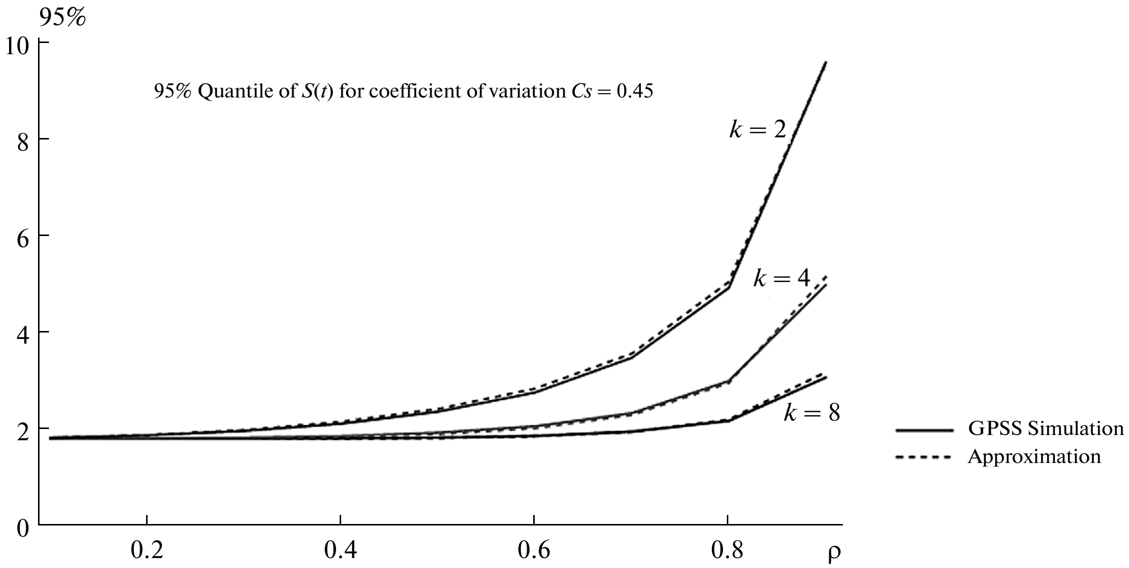

Fig. 5. Comparison of the 95% quantile for different loads of the *M/G/k* queue. Coefficient of variation *Cs* = 0.45.

where $\gamma_j = \int_0^{\infty} t^{j-1}(1 - B_e(t))^k dt$.

The moments of waiting time distribution in the system with exponential service time distribution are defined in a known manner as follows [24]:

$$E[W_{\exp}^j] = \frac{j!P_k}{k^j\mu^j(1-\rho)^j}. \tag{14}$$

Applying relations (10)–(13) to Eq. (3), it is not difficult to calculate cumulants of the waiting time distribution in *M/G/k* queue. As noted above, cumulant analysis is a very useful mathematical tool. In particular, it allows one to calculate the 95% quantile of the probability distribution taking advantage of the Edgeworth series [14] or Cornish–Fisher expansion [26]. At the same time, the Cornish–Fisher approach yields better results in a wide range of distribution functions and it is necessary to know only first four cumulants as follows [27]:

$$x_\alpha = M + \frac{(\sigma^2+1)z_\alpha}{2} + \frac{(z_\alpha^2-1)\chi_3}{6} + \frac{(z_\alpha^3-3z_\alpha)\chi_4}{24} - \frac{(2z_\alpha^3-5z_\alpha)\chi_3^2}{36}, \tag{15}$$

where M is the mean, σ is the standard deviation, z_α is the α-quantile of normal distribution, and χ_3 and χ_4 are the third and fourth cumulants of the distribution.

4.1. Verification of the Proposed Model

To verify the proposed analytical method, a simulation program has been developed using the GPSS World Student Version, which allows one to investigate both standalone queues and the entire IMS network core, which process voice calls flows. As source information for the simulation model were used empirical data collected on the MGTS network. The real IMS-network research has showed that the response delay distribution function in real IMS-modules have a coefficient of variation *Cs* = 0.45. Queues that serve SIP messages with two, four, and eight serving modules that match processor cores or cards in the telecommunications cassette were researched.

Figure 5 shows plots of the 95% quantile of the response time distribution. Uninterrupted lines show the dependences obtained by simulation, and the dashed ones were derived analytically. The graph shows the dependence for coefficient of variations in the service time distribution *Cs* = 0.45 and for different numbers of servers in the queue, where each of the servers has a service intensity of μ = 1. Cases for a coefficient of variation of 0.3 and 0.8 were also investigated. A comparison of the results shows that the difference between analytical values and values obtained by simulation does not exceed 3%, which is sufficient for the practical application of the proposed approximation.

In order to estimate the standardized QoS metrics, it is necessary to calculate the cumulants of the response time delay of each SIP message type in each of the core's nodes. Then, it is necessary to produce the sequential summation of cumulants that constituted the normalized value; then, it is possible to calculate the mean and 95% quantile. As noted in Section 3, this approach is possible if the delays in message response time in each module are mutually independent variables.

A simulation model was developed in order to test this hypothesis that reflects the signaling messages exchanged in the IMS core, when connections are established and destroyed according to the scenario in Fig. 2. It was considered an example where the intensity of incoming calls $\lambda_{\text{INVITE}} = 0.8$, the intensity of service unit in each of the nodes $\mu = 1$ and in each node, the number of service units $k = 8$. Then the maximum load there is on the I-CSCF $\rho_{\text{I-CSCF}} = 0.7$. For the rest of the modules, it is not difficult to see that the load is $\rho_{\text{P-CSCF}} = 0.6$, $\rho_{\text{S-CSCF}} = 0.6$, and $\rho_{\text{HSS}} = 0.1$.

With a sequential computation of the cumulants, we can calculate the mean and quantile of standardized metrics of quality of service. A comparison of the results obtained analytically and using the simulation GPSS model is given in the Table 2 below.

Table 2. Comparison of quality-of-service metrics

Quality of service metric	PSD		ASD		CRD	
	M	$x_{95\%}$	M	$x_{95\%}$	M	$x_{95\%}$
Simulation GPSS	12.58	15.67	10.53	13.30	10.50	13.27
Approximation	12.48	15.23	10.41	12.92	10.41	12.92
Error (%)	0.77	2.81	1.18	2.87	0.90	2.65

5. CONCLUSIONS

The quality of service in IMS networks is characterized by random variables of various services delays. While there are many studies on estimating the average network delays, the issue of multiqueue systems quantile estimation remains insufficiently investigated. The study in this article, demonstrates the possibility of considering modules of the IMS network while providing services, such as mutually independent *M/G/k* queues. The analytical apparatus, which allows one to calculate the parameters of the response time distribution of the separate queue and trough "end-to-end" delay has been proposed. A comparison of the calculations results with the data obtained by simulation proves the possibility of using this method in practice.

REFERENCES

1. Rosenberg, J. et al., SIP: Session Initiation Protocol RFC 3261, *IETF,* June 2002.
2. Calhoun, P. et al., Diameter Base Protocol RFC 3588, *IETF,* September 2003.
3. 3GPP, Network architecture, TS 23.002 (V12.4.0), March 2014.
4. 3GPP, Policy and charging control architecture, TS 23.203 (V12.4.0), March 2014.
5. ITU-T Recommendation Y.1531, 2007. SIP-based Call Processing Performance.
6. Malas, D. et al., Basic Telephony SIP End-to-End Performance Metrics RFC 6076, *IETF,* January 2011.
7. ITU-T Recommendation E.721, 1999. Network grade of service parameters and target values for circuit-switched services in the evolving ISDN.
8. MGTS introduces IMS platform and accelerates launch of new services. http://mgts.ru/company/press/news/344479/#
9. Erlang, A., The theory of probabilities and telephone conversations, *Nyt Tidsskrift for Matematik,* 1909, vol. 20, pp. 33–39.
10. Yang, S. and Chen, W., SIP multicast-based mobile quality-of-service support over heterogeneous IP multimedia subsystems, *Mobile Comput., IEEE Trans.,* 2008, vol. 7, no. 11.
11. Abaev, P., Razumchik, R., and Uglov, I., Statistical analysis of message delay in SIP proxy server, *J. Telecommun. Inf. Technol.,* 2014, no. 4.
12. Poikselka, M. and Mayer, G., *The IMS: IP Multimedia Concepts and Services,* John Wiley and Sons, Ltd., 2009, 3rd ed.
13. Ahson, A. and Ilyas, M., *IP Multimedia Subsystem (IMS) Handbook,* Boca Raton, FL: Taylor & Francis Group, 2009.

14. Cramer, H., *Mathematical Methods of Statistics,* Princeton: Princeton University Press, 1946.
15. Malakhov, A., *Kumulyantnyi analiz sluchainykh negaussovykh protsessov i ikh preobrazovanie* (Cumulant Analysis of Random Non-Gaussian Process and Their Transformation), Sovetskoe Radio: Moscow, 1975.
16. Burke, P.J., The dependence of sojourn times in tandem M/M/s queues, *Oper. Res.,* 1969, no. 17.
17. Kiessler, P.C., Melamed, B., Yadin, M., and Foley, R.D., Analysis of a three node queueing network, *Queueing Syst.,* 1988, no. 3.
18. Lee, A.M. and Longton, P.A., Queuing processes associated with airline passenger check-in, *J. Oper. Res. Soc.,* 1959, no. 10, pp. 56.
19. Hokstad, P., Approximations for the M/G/m Queue, *Oper. Res. (INFORMS),* 1978, vol. 26, no. 3, pp. 510–523.
20. Kollerstrom, J., Heavy traffic theory for queues with several servers, *J. Appl. Probab. (Appl. Probab. Trust),* 1974, vol. 11, no. 3, pp. 544–552.
21. Boxma, O. and Cohen, J., Approximations of the mean waiting time in an M/G/s queuing system, *Oper. Res. (INFORMS),* 1979, vol. 27, no. 6, pp. 1115–1127.
22. Bertsimas, D. and Nakazato, D., The distributional Little's law and its applications, *Oper. Res.,* 1995, pp. 298–310.
23. Tijms, H.C., *A First Course in Stochastic Models,* West Sussex: A John Wiley and Sons, Ltd., 2003.
24. Kleintrok, L., *Teoriya massovogo obsluzhivaniya. Mashinostroenie* (Queueing Theory. Mechanical Engineering), Moscow, 1979.
25. Hanbali, A. Al, Alvarez, E.M., and Heijden, M.C. van der, Approximations for the waiting-time distribution in an M/PH/c priority queue, *OR Spectrum,* 2015, vol. 37, no. 2, pp. 529–552.
26. Cornish, E.A. and Fisher, R.A., Moments and cumulants in the specification of distributions, *Revue de l'Institut International de Statistique,* 1937, no. 5, pp. 307–320.
27. Kulikov, N., Analysis of QoS estimation methods in IMS-networks, *Telecommun. Transp.,* 2014, no. 10, pp. 96–99.