Performance Evaluation of Cloud Computing Accounting for Expenses on Information Security

Anatoly Khomonenko, Sergey Gindin
Petersburg State Transport University
Saint Petersburg, Russia
khomon@pgups.ru, sgindin@gmail.com

***Abstract*—An approach to the performance evaluation of cloud computing accounting for the expenses on information security is proposed. For this purpose the model of multi-channel non-Markov queueing system with a "warm-up" is used. When applied to information security at cloud computing warm-up can be understood as: user/terminal authentication, network encryption, caching database queries, on-demand resource allocation, etc. The study demonstrates the influence of the "warm-up" process on the cloud computing performance and shows the need to collect data and examine "warm-up" patterns to assure that the system capabilities are appropriate for significantly different levels and patterns of demand. Obtained results of numerical experiments show the influence patterns which the "warm-up" process has on the system performance.**

I. Introduction

Increased access to high-speed Internet, as well as innovations in virtualization and distributed computing, have boosted public interest in Cloud computing, which, now being widely used, is transforming the information technology. As applications and data migrate to the cloud, it is transforming not just where computing is done, but, fundamentally, how it is done. Cloud computing now offers several advantages in terms of scalability, maintainability, high availability, performance. Performance is probably the most important competitive aspect of cloud-based solutions generally accessed from the web, where the quality of service (QoS) depends heavily on the performance level. As Buytaert mentiones in [1], for Amazon and Google even 100 ms of extra load time result in substantial and costly drops in revenue.

Measuring the performance is done by evaluating certain factors and criteria throughout the product lifecycle. It starts from early design phase with the preliminary efficiency evaluation, which is then supplemented by stress tests on prototypes and production copies. Using an integrated approach to measurements allows estimating in advance the planned performance level and preventing service unavailability under load, helping to plan the resources costs and allocation policies for the given parameters of the system configuration and performance metrics. With effective planning of computing resources during the development phase, the development time frames could be reduced and, consequently, the overall costs may be lowered, too. Examples of early works on cloud performance subject are present by Gong in [2] and by Xiong and Perros in [3].

Mathematical modeling plays an important role in analyzing the performance parameters of modern cloud computing systems. As they grow bigger and become more complex, the mathematical models respectively take steps from simple ones, when the solution is obtained analytically, towards models, where it is only possible to compute the solution during simulation process. Performance estimates are commonly based on the queueing theory, a discipline within the mathematical theory of probability, which studies waiting lines, or queues. In queueing theory a model of queueing system (QS) is constructed and its lifecycle is scrutinized as a stochastic process to predict the probability characteristics of efficiency such as queue lengths and waiting times.

This paper examines multichannel non-Markov queues with "warm-up", accounting for expenses on information security. The request that comes to a system before the security checks are passed only begins to be serviced after a certain time of system warm-up. The examples of warm-up processes are: negotiation of the session key while reestablishing the encrypted communication, authentication of the user and assigning a security token, triggering the guard security timers.

II. Existing models and related work

For queueing theory Kendall's notation is the standard system to describe and classify a queueing node. It uses three factors written A/S/n, where A references to the time between arrivals to the queue, S — to the service time distribution and n denotes the number of servers at the node. It has since been extended to A/S/n/K/N/D where K and D mean the capacity of the queue and queueing discipline and N denotes the size of the population of jobs to be served. Best studied models are those with up to n channels and where both the service time distribution and the distribution for the time between arrivals follow an exponential distribution: M/M/n. Because of the assumptions made, such models known as Markov models, have limited appliance and do not fit for most practical systems. A good overall status report on various models in cloud computing using queueing theory for a specific resource allocation task is present by Murugesan, Elango and Kannan in [4].

The most examined are the relatively simple M/M/n models. The simplest example is M/M/1 queue, for which textbooks pertaining to performance evaluation usually present the results to compute the steady state distribution of number of requests. Another well studied class is the one-channel models with

specific flow characteristics, discussed e.g. by Ryzhikov in [5] or Szekli, Disney and Hur in [6], which analyze the behavior of MR/GI/1 queues with positively correlated arrivals, or queueing systems with modified service mechanism as reviewed by Kaczynski in [7] or Takahashi in [8]. Vilaplana, Solsona and Teixido in [9] use an M/M/n performance model for scalable cloud computing simulation.

The biggest interest has recently been focused on the investigations in multichannel non-Markovian queues where flows are approximated by phase-type distributions. For example Bubnov, Khomonenko and Tyrva in [10] examine the C_2/M/n model with Coxian distribution to forecast software reliability characteristics, such as number of corrected errors, required debugging time, etc. Brandwajn and Begin in [11] propose a semi-numerical approach to compute the steady-state probability distribution for the number of requests at arbitrary and at arrival time instants in Ph/M/c-like systems.

Cox showed in [12] that an arbitrary distribution of length of a random variable can be represented by a compound of exponential stages or phase-type distribution. The advantage of such a representation is that it ensures convenience of approximation of the random process to a Markov process and gives the power of creating and solving the system of equations describing the behavior of the corresponding model.

Described here multichannel non-Markov QS with warm-up require more complex mathematical description, compared to the Markov models, e.g. the request flow can be recurrent or represented by an arbitrary stochastic function. Examples of previous works addressing QS with warm-up are by Kolahi in [13] or by Kreinin in [14] for the characteristics of single channel QS, or by Bin Sun and Dudin in [15] studying the MAP/PH/n multichannel QS with warm-up and broadcasting service discipline. Mao and Humphrey in [16] examine the influence of the warm-up during virtual machine startup in the cloud system.

III. Expenses on Information Security

We consider the problem of estimating the efficiency of distributed processing based on information security expenses. The major information security threats to cloud computing as identified by Cloud Security Alliance in [17] are:

1) Difficulty in moving avegare servers to the computing cloud.
2) Dynamism of virtual machines.
3) Vulnerabilities within the virtual environment.
4) The presence of inactive virtual machines.
5) The complexity of perimeter protection and the need for differentiation of the network.

The most effective means of protection of the cloud computing are:

1) Data security. Encryption.
2) Data protection during the transmission.
3) Authentication.
4) Users isolation.

Known QS models do not fully allow to take into account the expenses of information security in the cloud computing. To solve the above problem the authors proposed in [18], [19] to use the set of models of multi-channel non-Markov QS with warm-up, approximated by two-phase hyperexponential distributions. Data servers with multi-core architecture are modelled by a multichannel QS, the number of channels in the QS is the number of processors. These models of multichannel non-Markov QS with warm-up allow to take into account the above-mentioned expenses on information security for cloud computing. The physical nature and mathematical formalization of warm-up processes are further studied by the authors in [18].

IV. Modeling a Class of Cloud Systems with Warm-Up

In the cloud the network infrastructure is shared among the consumers, so the information traveling over the cloud may often be vulnerable to interception. A sample cloud system with logical separation by virtualization layers with inter-layer security and physical separation of computing environments reviewed by Okuhara, Shiozaki and Suzuki in [20], is illustrated in Fig. 1.

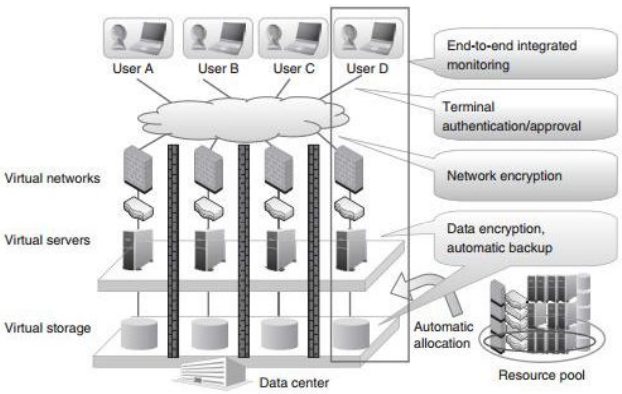

Fig. 1. Separation of computing environments

It illustrates an architecture of cloud computing and highlights the associated information security operations adding computational costs.

Quite rightly, many of today's data privacy requirements and standards include, as a baseline level of protection, a mandate to protect data in motion. While the provider can choose to encrypt selected data at the application level or within databases or other storage environments, the bulk protection of data flowing over a network provides a blunt but very effective instrument for adding an extra layer of security.

To study the cloud systems with described warm-up it is useful to introduce an enhanced notation A/W/S/n, which compared to original Kendall's notation contains additional W denoting the warm-up time distribution. Principally, in real systems all three properties might be non-Markov. In this paper the authors examine a cloud system as a QS model where A – the incoming flow – is approximated by two-phase hyperexponential distribution (H_2) and W – is the Poisson (i.e. exponential) warm-up process, the system type is H_2 / M / M / n.

Let's describe the parameters of a model which is set up. For modeling simplicity let's denote status of a multichannel

non-Markov QS in the form of a set of microstates. Microstates are all sorts of states which the system might be in while in operation.

The QS with the classification $H_2/M/M/n$ has the microstate diagram, as shown in Fig. 2. The arrival process has hyperexponential distribution H_2 with parameters: λ_1 and λ_2 – arrival rates of corresponding phases, a_1 – probability of choosing the first phase, $a_2 = 1 - a_1$ – probability of choosing the second phase; parameters μ_p and μ describe the intensity of the warm-up and serving processes in the system.

The column on the left shows the number of requests in the system. An n-channel system after reaching the full load (on the diagrams – layers with numbers greater than n) eventually stabilizes.

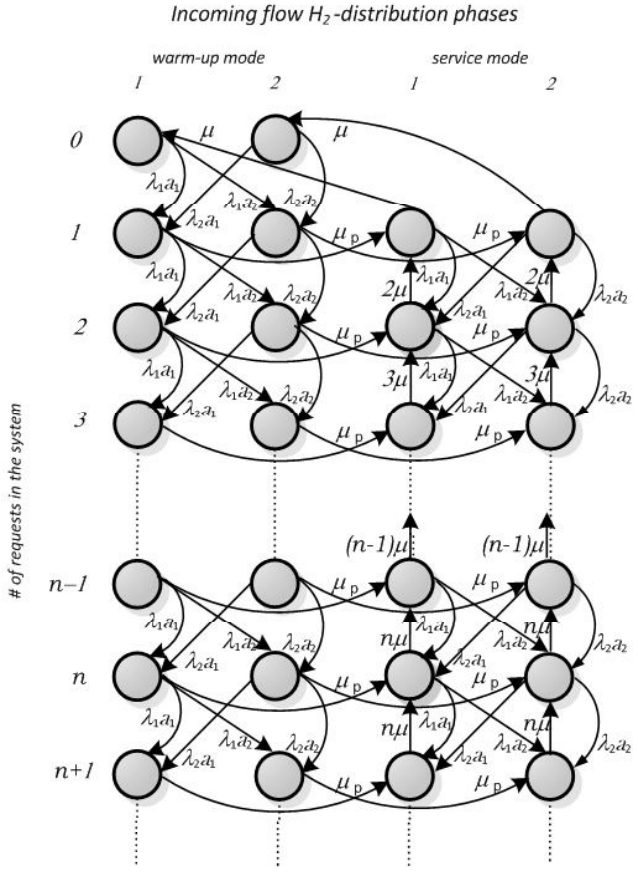

Fig. 2. Microstates diagram for H2/M/M/n system

V. Computation Analysis

Computation of described QS $H_2/M/M/n$ in this paper is studied for the case of n = 3. Computation algorithm is based on a numerical method for the steady-state probabilities of a GI/G/n QS of a general class, as introduced by Takahashi and Takami in [21], since this method is proven to be stable even at the large-scale computations.Denote as Sj the set of system microstates when exactly j requests are served, and by σj – number of elements in Sj. Then from the microstates diagram, analytically the following matrices describing the system are defined:

- Aj[σj×σj+1] – in Sj (request arrival),
- Bj[σj×σj-1] – in Sj-1 (request service completion),
- Cj[σj×σj] – in Sj (request service in progress),
- Dj[σj×σj] – leaving microstates of tier j (a diagonal matrix).

For each tier j denote by vectors $\gamma_j = \{\gamma_{j,1}, \gamma_{j,1}, ..., \gamma_{j,\sigma_j}\}$ the probability that a QS is in microstate (j, i), j = 0,1,… Then it is possible to write the system of vector-matrix balance equations describing transitions between microstates:

$$\gamma_0 D_0 = \gamma_0 C_0 + \gamma_1 B_1,$$
$$\gamma_j D_j = \gamma_{j-1} A_{j-1} + \gamma_j C_j + \gamma_{j+1} B_{j+1}, \quad j = 1,2,...$$

For described QS the matrices A, B, C, D are the following:

$$A_0 = \begin{bmatrix} a_1\lambda_1 & a_2\lambda_1 & 0 & 0 \\ a_1\lambda_2 & a_2\lambda_2 & 0 & 0 \end{bmatrix}, \quad A_1 = \begin{bmatrix} a_1\lambda_1 & a_2\lambda_1 & 0 & 0 \\ a_1\lambda_2 & a_2\lambda_2 & 0 & 0 \\ 0 & 0 & a_1\lambda_1 & a_2\lambda_1 \\ 0 & 0 & a_2\lambda_1 & a_2\lambda_2 \end{bmatrix},$$

$$A_j = A_1, \quad j > 1;$$

$$B_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad B_1 = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ \mu & 0 \\ 0 & \mu \end{bmatrix}, \quad B_2 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 2\mu & 0 \\ 0 & 0 & 0 & 2\mu \end{bmatrix},$$

$$B_3 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 3\mu & 0 \\ 0 & 0 & 0 & 3\mu \end{bmatrix}, \quad B_j = B_3, \quad j > 3;$$

$$C_0 = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, \quad C_1 = \begin{bmatrix} 0 & 0 & \mu_p & 0 \\ 0 & 0 & 0 & \mu_p \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad C_j = C_1, \quad j > 1;$$

$$D_0 = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}, \quad D_1 = \begin{bmatrix} \lambda_1 & 0 & 0 & 0 \\ 0 & \lambda_2 & 0 & 0 \\ 0 & 0 & \lambda_1 + \mu & 0 \\ 0 & 0 & 0 & \lambda_2 + \mu \end{bmatrix},$$

$$D_2 = \begin{bmatrix} \lambda_1 & 0 & 0 & 0 \\ 0 & \lambda_2 & 0 & 0 \\ 0 & 0 & \lambda_1 + 2\mu & 0 \\ 0 & 0 & 0 & \lambda_2 + 2\mu \end{bmatrix}, \quad D_3 = \begin{bmatrix} \lambda_1 & 0 & 0 & 0 \\ 0 & \lambda_2 & 0 & 0 \\ 0 & 0 & \lambda_1 + 3\mu & 0 \\ 0 & 0 & 0 & \lambda_2 + 3\mu \end{bmatrix},$$

$$D_j = D_3, \quad j > 3.$$

An iterative numerical method introduced by Takahashi and Takami in [21] is then used to solve the system and find the steady-state microstates probability distribution. This classical iterative method is chosen for it's well known convergence

properties [22]. To the authors' best knowledge, little is known about global convergence theorem and speed of convergence for competitive iterative algorithms except the direct successive substitution method.

Finally the mean parameters of requests servicing process are obtained using Little's law and the waiting time distributions for waiting time in the queue, time in the system, the distribution of the number of requests in the queue and in the system are obtained using the results of [23] to compute the Laplace-Stieltjes transform.

VI. Experiments and practical results

A Java program [24] has been written by the authors to implement the described computation algorithm. It has been designed to perform actions on matrices in general, and therefore it allows obtaining results for both presented models as well as for models of other systems within the comparable classes of phase-type distributions.

Since it is not possible to perform benchmarking experiments using real-world cloud environments, a series of extensive experimentation computations has been run with various input parameters in order to explore the behavior of the modeled systems. The initial data has been produced from test statistics gathered for a sample cloud system as shown in Fig 1. To collect it a simulation performance testing in an Apache JMeter tool [25] was run with 300 threads in parallel and independently performing queries of various kinds, including:

- Session startup.
- Authentication and Login.
- Conducting business transactions.
- Management of user activity limiters.
- Reports Generation.

The following computational resources were used to perform the testing:

Application servers have the following characteristics:

- CPU Intel Xeon X7560 (2.26 GHz), 8 cores.
- Hyperthreading disabled.
- 512Gb RAM.
- Memory allocated in the range of 6-8 Gb per node.
- Average CPU load in the range 0.3-0.5 per core.

Network segments have the following characteristics:

- Nodes reside in the same logical network segment.
- 100/1000 Mbps Ethernet ports are used.

After obtaining the system performance characteristics and adjusting the model parameters, it was useful to experiment with the model parameters, changing the request arrival process parameters (e.g.: mean arrival rate $\widetilde{f}_1$, the coefficient of variation υ_A for distribution of the interval between adjacent requests) and the warm-up parameter μ_p which allow to study the influence of the warm-up on system performance characteristics under different scenarios and determine the optimal conditions for receiving the desired QoS level.

In a simulation for a system corresponding to the model shown on Fig. 2 with static parameters: $\mu = 1.85$, $\widetilde{f}_1 = 0.5977$, $\upsilon_A = 1.5$, $a_1 = 0.6$ and mean warm-up time changing between 0.1 and 1.9 seconds the following results were received for the expected waiting time in the queue at different values, as described in Fig. 3.

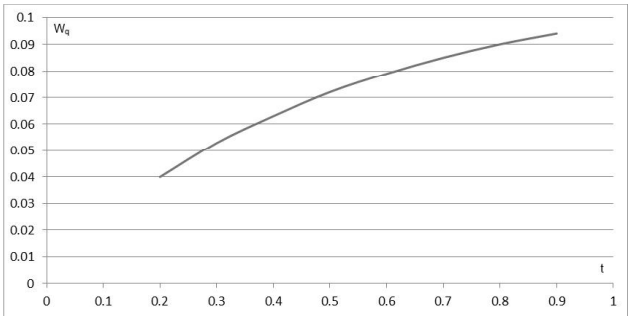

Fig. 3. Expected time in the queue in the H_2/ M/M/3 system at different values of mean warm-up time

The corresponding experiment results shown in Table I listing the expected number of customers in the system – *L*, expected queue length – *Lq*, waiting time in the queue – *Wq*, and expected time in system, including service time – *W* at different values of the mean warm-up time between 0,1 and 0,9 seconds.

TABLE I. Expected average characteristics in the system H_2/M/M/3 at different values of mean warm-up time

#	t = 0,1	t = 0,2	t = 0,3	t = 0,5	t = 0,7	t = 0,8	t = 0,9
L	1,292	1,481	1,628	1,835	1,972	2,023	2,067
Lq	0,083	0,128	0,167	0,227	0,269	0,286	0,3
W	0,273	0,313	0,345	0,389	0,417	0,429	0,438
Wq	0,026	0,04	0,053	0,072	0,085	0,09	0,094

Results show that the "warm-up" time change may significantly affect the waiting time and service time, especially when the "warm-up" and service times are of the same order of magnitude. On the right side of the Fig., when the "warm-up" time is similar or even bigger to the service time, reducing the "warm-up" would result in a substance speed up of the service. But in the left side of Fig. 1, the performance curve gets steeper as the "warm-up" time starts to be similar or less than the service time.

The accuracy of the performance estimation results here depend on combination of both "warm-up" and service time. To investigate into the area under which the calculations stay stable more modelling experiments have been conducted for a range of coefficient of variation υ_A both below and above 1,0. Due to the nature of the hyperexponential distribution, to experiment on subrange below 1.0 the equivalent model with generalized two-phase Erlang distribution is used and the parameters alignment is performed to fit the two models together. The method of models fitting is based on solving the system of transformation equations from initial distribution to the same generalized two-phased Coxian distribution, the

approach is further described by Bubnov, Khomonenko and Tyrva in [10].

The results of comparison the two models are shown at Fig. 4. It shows the stationary distribution of the number of requests in the system.

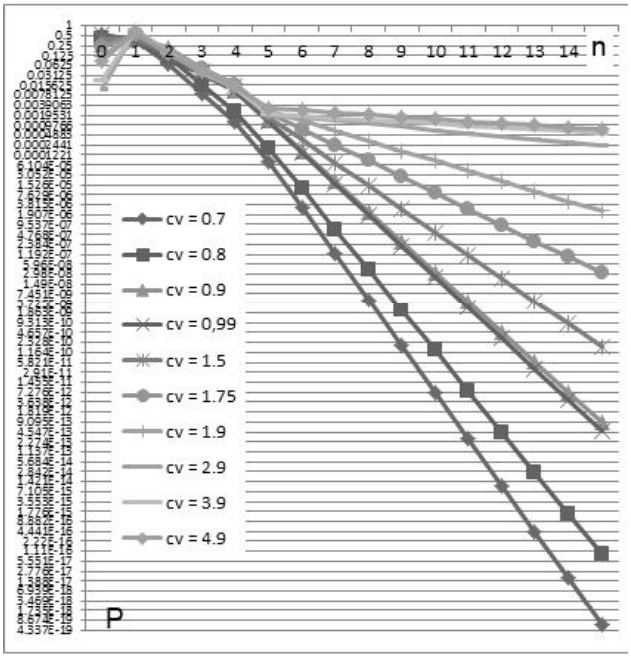

Fig. 4. Stationary distribution of the number of requests in the system with E_2/M/M/3 and H_2/ M/M/3 models

E_2/M/M/3 model is used for υ_A values from 0,7 to 0,99; H_2/M/M/3 model is used for υ_A values from 1,0 to 4,5; 1,0 is the boundary value for the both. The distributions at given values of υ_A are forming the internally-unified and smoothly varying function set which allows to accurately predict the changes of the system behavior at the whole range of υ_A variance.

Further increase of the accuracy of modeling is to improve the accuracy of determining the input parameters — the empirical mean time of the system warm-up and the values, describing the arrival rate.

Note: other approaches to the performance analysis for cloud systems are also possible. Examples of other approaches to the performance analysis of the cloud systems could be found in works [26] by Bruneo and [27] by Kliazovich, Bouvry and Khan.

VII. Discussion

The study demonstrates the influence of the “warm-up” process on the cloud computing performance and shows the need to collect data and examine “warm-up” patterns to assure that the system capabilities are appropriate for significantly different levels and patterns of demand that might be relevant during a given time period.

This study also supports the usefulness of considering the “warm-up” process in guiding the analysis of the QS. The results are relevant for cloud distributive systems where resources are tight relative to potential demand because, in such situations even relatively small changes can have a substantial impact on delays and waiting times and thus may affect the QoS.

Despite analytic models such as queueing models can never capture all characteristics of an actual operational system, still it has been demonstrated over many years, that modelling can be invaluable in providing decision support that greatly improves performance, particularly in complex distributed environments.

The study illustrates the proposed approach to data analysis for cloud computing, where queueing models can be used to identify possible changes that can decrease the delays in system service; it also provides the grounds to obtain meaningful estimates of the capacity planning. An analytic model, in combination with a carefully developed, appropriate computation analysis, can provide an objective evaluation of what additional resources are required to meet a given standard of performance.

The applicable area for the proposed approach is verified by cross-checking between models with two different disctributions equivalently transformed against each other. This is also useful for mutual testing of the characteristics of models with H_2-and E_2-distributions. Basis for such testing would be the principle of reciprocity – test input data, specifically selected to obtain theoretically predicted results, is computed on different models, testing whether they produce the same (similar) results or not.

Whereas the “warm-up" process is modelled precisely for an arbitrary distribution with known moments, assumptions are made in this paper regarding the residual service times of services and arrival process in order to obtain the approximations. In real cloud systems the memoryless properties of the service time distribution are still quite acceptable, while the operating characteristics are far more sensitive to changes in the parameters of the arrival process, as outlined by van Hoorn and Seelen in [28].

This is done in a purpose to stress out an effect which information security expenses have on performance characteristics and examine their engineering relevance particularly. Practical cloud systems differ in the architecture of their information security measures and technologies (or generally speaking, in the process of the computational context preparation) often without devoting enough attention to the impact that those “warm-up” expenses would have on system performance characteristics.

VIII. Conclusions

In the present paper, the class of systems H_2/M/M/n with “warm-up” defined by hyper exponential approximation was studied to examine the effect of information security expenses on the cloud computing performance. The key parameters of system performance were investigated and plotted. Obtained results of numerical experiments show the influence patterns which the “warm-up” process has on the system performance. The studied model being a multi-channel, featuring unbounded queue and considering the non-Markov “warm-up” process, it determines the practical importance of presented results in assessing the efficiency of cloud systems with multiple

processing nodes, being theoretically capable of approximating any arbitrary distribution of the real cloud systems.

The further study is to be focused on expanding of the characteristic distribution function of the described models on to the networks of QS with multiple nodes, on expanding the research on hyper exponential distribution case with coefficients represented by complex numbers and on expanding the modelling class on multichannel systems with both the "warm-up" time distribution and the inter-arrival rate distribution approximated by the phase-type distributions at the same time.

ACKNOWLEDGMENT

This work was financially supported by the Petersburg State Transport University, at which the authors are members of the Information Technology Systems department. Many thanks to our fellow colleagues and in particular to Vladimir Bubnov for very useful discussions held on the topic of this paper.

REFERENCES

[1] Blog of Dries Buytaert, founder of Drupal, Web: http://buytaert.net/faster-is-better

[2] C. Gong et al., "The characteristics of cloud computing". *39th International Conference on Parallel Processing Workshops (ICPPW)*, IEEE Press, 2010, pp. 275-279.

[3] K. Xiong, H. Perros, "Service performance and analysis in cloud computing" *World Conference on Services-I*, IEEE Press, 2009, pp. 693-700.

[4] R.Murugesan, C.Elango, S.Kannan, "A Status Report on Resource Allocation in Cloud Computing Using Queuing Theory", *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, Volume 3 Issue 11, November 2014, ISSN:2278-1323, pp. 3603-3608

[5] Yu. I. Ryzhikov, "Distribution of the Number of Requests in a Queueing System with warm-up" *Probl. Peredachi Inf.*, 9:1 (1973), pp. 88–97.

[6] R. Szekli, R. L. Disney, S. Hur. "MR/GI/1 queues with positively correlated arrival stream." *J. Appl. Prob.*, Vol. 31, 1994. pp. 497–514.

[7] W. H. Kaczynski, "Transient Queueing Analysis", *INFORMS Journal on Computing*, Vol. 24, No. 1, Winter 2012, pp. 10–28

[8] Y. Takahashi, "A Single-Server Queueing System with Modified Service Mechanism: An Application of the Diffusion Process to System Performance Measure Formulas" *WASEDA BUSINESS & ECONOMIC STUDIES*, 2005 NO. 41, pp. 19-28.

[9] J. Vilaplana, F. Solsona, I. Teixido, "A performance model for scalable cloud computing", *Proceedings of the 13th Australasian Symposium on Parallel and Distributed Computing (AusPDC 2015)*, Sydney, Australia, 27 - 30 January 2015, pp. 50-60

[10] V.P. Bubnov, A.D. Khomonenko, A.V. Tyrva, "Software reliability model with coxian distribution of length of intervals between errors detection and fixing moments", *Proceedings — 35th Annual IEEE International Computer Software and Applications Conference Workshops, COMPSACW*, 2011, pp. 310-314.

[11] A. Brandwajn, T. Begin, "A recurrent solution of Ph/M/c/N-like and Ph/M/c-like queues", *Journal of Applied Probability 49*, 1 (2012), pp. 84-99.

[12] D.R. Cox, "A use of complex probabilities in the theory of stochastic processes", *Proc. Cambr. Phil. Soc.* –1955. –V. 51, № 2, pp. 313-319.

[13] S.S. Kolahi, "Simulation Model, Warm-up Period, and Simulation Length of Cellular Systems", *Second International Conference on Intelligent Systems, Modelling and Simulation (ISMS)*, 2011, pp. 375 – 379.

[14] Ya. Kreinin, "Single-channel queueing system with warm up," *Automation and Remote Control*, 41, 6, 1980, pp. 771-776.

[15] Bin Sun, A. N. Dudin, "The MAP/PH/N multi-server queuing system with broadcasting service discipline and server heating", *Automatic Control and Computer Sciences,* July 2013, Volume 47, Issue 4, pp. 173-182.

[16] M. Mao, M. Humphrey, "A performance study on the vm startup time in the cloud", *IEEE 5th International Conference on Cloud Computing (CLOUD)*, IEEE Press, 2012, pp. 423-430.

[17] Cloud Security Alliance, "Security Guidance for Critical Areas of Focus in Cloud Computing v3.0", 2011, Web: https://cloudsecurityalliance.org/group/security-guidance/

[18] S. I. Gindin, A. D. Khomonenko, S. E. Adadurov. "Numerical calculation of multichannel queueing system with recurrent input flow and warm-up". *Proceedings of the Petersburg State Transport University.* - 2013, #4 (37), pp. 92-101.

[19] A.D. Khomonenko, S.I. Gindin, "Stochastic models for cloud computing performance evaluation", *Proceedings of the 10th Central and Eastern European Software Engineering Conference in Russia*, 2014 (CEE SECR'14). Article No. 20. Web: http://dl.acm.org/citation.cfm?id=2687233

[20] M. Okuhara, T. Shiozaki, T Suzuki. "Security Architectures for Cloud Computing", *Fujitsu Sci. Tech. J.* vol.46, Oct. 2010, no. 4, pp. 397–402.

[21] Y. Takahashi, Y. Takami. "A numerical method for the steady-state probabilities of a GI/G/c queuing system in a general class" *J. of the Operat. Res. Soc. of Japan.* – 1976. – V. 19, N 2. – pp. 147-157.

[22] W. Cao, W.T. Stewart, "Iterative Aggregation / Disaggregation Techniques for Nearly Uncoupled Markov Chains", *Journal of the ACM*, Vol.32 (1985), pp. 702-719.

[23] A.D. Khomonenko, "Waiting time distribution in queueing systems of type $GI_q/H_k/n/R\leq\infty$", *Automation and Remote Control.*, 1990, #8, pp. 91–98.

[24] S.I. Gindin, A.D. Khomonenko, S.V. Matveev, "Program for probability-time characteristics calculation in multichannel queueing systems with "warm-up" and its testing approach", *Modern problems of science and education*, 2014, № 4, Web: http://www.science-education.ru/118-13872

[25] Apache Jmeter official website, an Apache Software Foundation application designed to load test functional behavior and measure performance, Web: http://jmeter.apache.org/

[26] D. Bruneo, "A stochastic model to investigate data center performance and qos in iaas cloud computing systems" *IEEE Transactions on Parallel and Distributed Systems*, 2014, pp. 560-569.

[27] D. Kliazovich, P. Bouvry, S.U. Khan, "Simulation and Performance Analysis of Data Intensive and Workload Intensive Cloud Computing Data Centers", *Optical Interconnects for Future Data Center Networks*. Springer, 2013, pp. 47-63.

[28] M.H. van Hoorn, L.P. Seelen, "Approximations for the GI/G/c Queue", *Journal of Appl. Probability* Vol. 23, No. 2 (Jun., 1986), pp. 484-494.