UDC 519.218

The Analysis of Cloud Computing System as a Queueing System with Several Servers and a Single Buffer

I. S. Zaryadov[*†], **A. A. Kradenyh**[*], **A. V. Gorbunova**[*]

[*] *Department of Applied Probability and Informatics, RUDN University, Miklukho-Maklaya str. 6, Moscow, 117198, Russia*

[†] *Institute of Informatics Problems, FRC CSC RAS IPI FRC CSC RAS, 44-2 Vavilova Str., Moscow, 119333, Russia*

Abstract. The mathematical model of cloud computing system based on the queuing system with the splitting of the incoming queries and synchronization of services is considered. The queueing system consists of a single buffer and N servers ($N > 2$), service times are independent and exponentially distributed. The incoming query enters the system as a whole and only before service is divided into subqueries, each subquery is served by its device. The servers with parts of the same query are considered to be employed as long as the query is not serviced as a whole: the query is handled only when the last of it is out and a new query may be served only when there are enough free servers (the response time is the maximum of service times of all parts of this query). The stationary probability-time characteristics of the system are presented.

Keywords: cloud computing system, splitting of incoming queries, queueing system, response time, stationary probability-time characteristics.

1. Introduction

This paper is devoted to the problem of cloud computing modeling [1]. There are several approaches to the cloud computing systems modeling. For the first approach [3–6] the cloud computing system is modeled via queuing system with K subqueues and each subqueue consists of a buffer with one (several) server. The incoming query is divided into K subqueries and each subquery enters the corresponding subqueue. On this approach the Fork–Join [3, 4] and Split–Merge [7–9] models are based. The Split–Merge model uses the idea of synchronization of servers (only when the all subqueries of the same query finished servicing the new query will be served). The second approach presents the cloud computing system as queuing system with unlimited number of homogeneous servers [10, 11], where the incoming query splits into several parts and each subquery is served by one of the free servers.

Our mathematical model of cloud computing system may be considered as general case of the second approach: the queuing system consists of N servers, a buffer and the incoming queries are splitted only before the start of the service. Also we use the idea of synchronization of services as in [7–9].

Our goal is to derive the analytical expressions for some probability-time characteristics of introduced model.

The paper consists of following sections: the Introduction, the section with system description for the general case (non-homogeneous servers), the section for homogeneous servers where the main results are obtained, in Conclusion the future goals are formulated.

2. The system description for general case

We will construct the mathematical model as a queuing system with N non-homogeneous servers ($N > 2$) and a buffer of size $r \leq \infty$. The incoming by a Poisson law with rate λ query enters the system as a whole and only before the start of its service is divided into N subqueries, the service time of each subquery has exponential distribution with rate μ_i, $i = \overline{1, N}$. The mechanism of synchronization is used — the servers with parts of the same query are considered to be employed as long as the query is not serviced as a whole: the query is handled only when the last part of it is out and a new query may be served only when all servers are free.

The service time for a whole query — the response time η (the main characteristic for cloud computing systems [2]) may be defined as maximum $\eta = \max(\eta_1, \ldots, \eta_N)$ [3–5] or as minimum $\eta = \min(\eta_1, \ldots, \eta_N)$ [12] of times spent by subqueries. In [13] is shown that the analysis of queuing models with the response time defined as minimum is equal to the analysis of well-studied multiserver queuing systems [14]. So we will consider only the case of maximum.

The probability distribution function (PDF) of $\eta = \max(\eta_1, \ldots, \eta_N)$ has the form [6, 13, 15]:

$$P\{\max(\eta_1, \ldots, \eta_N) < x\} = \prod_{i=1}^{N} \left(1 - e^{-\mu_i x}\right). \tag{1}$$

For homogeneous servers ($\mu_i = \mu$, $\forall i = \overline{1, N}$) the (1) takes the form:

$$P\{\max(\eta_1, \eta_2, \ldots, \eta_N) < x\} = \left(1 - e^{-\mu x}\right)^N. \tag{2}$$

We will consider the random process $\nu(t) = \{\xi(t), \vec{\delta}(t)\}$, $\xi(t)$ is the number of queries in the buffer at time t, the vector $\vec{\delta}(t) = (\delta_1(t), \ldots, \delta_N(t))$ describes the servers occupancy ($\delta_i(t) = 1$ if the i-th server is occupied by the i-th part of a query and $\delta_i(t) = 0$ otherwise). The set of states is defined as $\mathcal{X} = \{0\} \cup \{I, (\delta_1, \ldots, \delta_N)\}$, where $I = \overline{0; r}$, $\delta_1, \ldots, \delta_N$ take values 0 or 1. The probability $P\{\xi(t) = I, \vec{\delta}(t) = \vec{\delta}\}$ is denoted as $p_{I,\vec{\delta}}(t)$ and the probability $P\{\nu(t) = 0\} = P\{\xi(t) = 0, \vec{\delta}(t) = \vec{0}\}$ of system being empty as $p_0(t)$. The steady-state probabilities (on the assumption of steady-state regime existence) are $p_{I,\vec{\delta}}$ and p_0.

3. The special case. Stationary probability-time characteristics

Let's assume that all servers are homogeneous, then we may redefine the random process $\nu(t)$ as $\nu(t) = \{\xi(t), \delta(t)\}$, $\delta(t)$ — the number of occupied servers for non-empty system. The steady-state probabilities are p_0 (the system is empty) and $p_{i,j}$, $i = \overline{0,r}$, $j = \overline{1,N}$, for case when there i queries in the buffer and j servers are occupied by subqueries, and satisfy the following systems of equations ($r = \infty$):

$$\begin{cases} \lambda p_0 = \mu p_{0,1}, \\ (\lambda + j\mu)\, p_{0,j} = (j+1)\mu p_{0,j+1}, \quad j = \overline{1, N-1}, \\ (\lambda + N\mu)\, p_{0,N} = \lambda p_0 + \mu p_{1,1}, \\ (\lambda + j\mu)\, p_{i,j} = \lambda p_{i-1,j} + (j+1)\mu p_{i,j+1}, \quad i \geq 1, j = \overline{1, N-1}, \\ (\lambda + N\mu)\, p_{i,N} = \lambda p_{i-1,N} + \mu p_{i+1,1}, \quad i \geq 1, \end{cases} \tag{3}$$

The normalization condition is:

$$p_0 + \sum_{i=0}^{\infty} \sum_{j=1}^{N} p_{i,j} = 1. \tag{4}$$

If we define the probability distribution π_j, $j = \overline{0,N}$, of number of occupied servers and probability distribution $\tilde{\pi}_i$, $i \geq 0$, of number of queries in the buffer, then from (3) and (4) we obtain:

$$\begin{cases} \lambda \tilde{\pi}_0 = \mu(p_{0,1} + p_{1,1}), \\ \lambda \tilde{\pi}_i = \mu p_{i+1,1}, \quad i \geq 1. \end{cases} \tag{5}$$

and

$$\begin{cases} \pi_0 = p_0, \\ \pi_j = \dfrac{\lambda}{j\mu}, \quad j = \overline{1,N}. \end{cases} \tag{6}$$

From (6) and normalization condition (4) for π_j, $j = \overline{0,N}$,the probability p_0 is obtained:

$$p_0 = 1 - \frac{\lambda}{\mu} \sum_{j=1}^{N} \frac{1}{j}.$$

The estimation of p_0 for the system with inhomogeneous servers:

$$1 - \frac{\lambda}{\min(\mu_1, \ldots, \mu_N)} \sum_{j=1}^{N} \frac{1}{j} \leq p_0 \leq 1 - \frac{\lambda}{\max(\mu_1, \ldots, \mu_N)} \sum_{j=1}^{N} \frac{1}{j}.$$

If we define as $\omega(s)$ the Laplace-Stieltjes transformation (LST) of waiting time PDF for an arbitrary query, $\omega_{i,j}(s)$ — the LST of waiting time PDF for the incoming query when there are i, $i \geq 0$, other queries in the buffer and j, $j = \overline{1,N}$, servers are occupied, then:

$$\omega(s) = p_0 + \sum_{i=0}^{\infty}\sum_{j=1}^{N}\omega_{i,j}(s)p_{i,j} = p_0 + \sum_{i=0}^{\infty}\omega_N^i(s)\sum_{j=1}^{N}\omega_j(s)p_{i,j}, \qquad (7)$$

where $\omega_j(s)$ is the LST of PDF (2) for $\eta = \max(\eta_1, \ldots, \eta_j)$, $j = \overline{1,N}$, [15]:

$$\omega_j(s) = \sum_{i=1}^{j}(-1)^{i-1}\sum_{k=1}^{C_j^i} k\frac{i\mu}{s+i\mu}.$$

4. Conclusions

The brief introduction to the mathematical model of cloud computing system based on the queuing system with the splitting of the incoming queries and synchronization of services was presented.

Our future goals are to evaluate probabilities $\tilde{\pi}_i(5)$, $i \geq 0$, and $p_{i,j}$ (3), $i \geq 0, j = \overline{1,N}$, LST (7) for N homogeneous servers. We also will try to generalize the model for case of $N = \alpha K$ servers, where K is a fixed number of subqueries for a query. And, of course, the model with inhomogeneous servers should be derived.

Acknowledgments

The publication was prepared with the support of the «RUDN University Program 5-100» and partially supported by RFBR grants No 15-07-03007, No 15-07-03406 and No 14-07-00090.

References

1. *Buyya R., Broberg J., Goscinski A. M.* Introduction to cloud computing // Cloud Computing: Principles and Paradigms. — John Wiley & Sons Inc. — 2011. — P. 3–42.
2. *Khazaei H., Misic J., Misic V. B.* A Fine-Grained Performance Model of Cloud Computing Centers // IEEE Transactions on Parallel and Distributed Systems. — 2012. — Vol. 24, no. 11. — P. 2138–2147.
3. *Flatto L., Hahn S.* Two parallel queues created by arrivals with two demands // SIAM Journal on Applied Mathematics. — 1984. — Vol. 44, no. 5. — P. 1041–1053.

4. *Nelson R., Tantawi A. N.* Approximate analysis of fork/join synchronization in parallel queues // IEEE Transactions on Computers. — 1988. — Vol. 37, no. 6 — P. 739–743.
5. *Thomasian A.* Analysis of fork/join and related queueing systems // ACM Computing Surveys (CSUR). — 2014. — Vol. 47, no. 17. — P. 17.1–17.71.
6. *Gorbunova A. V., Zaryadov I. S., Matushenko S. I., Sopin E. S.* The Estimation of Probability Characteristics of Cloud Computing Systems with Splitting of Requests // Proceedings of the Nineteenth International Scientific Conference: Distributed computer and communication networks: control, computation, communications (DCCN-2016, Russia) / Communications in Computer and Information Science. — 2016. — Vol. 678. — P. 418–429.
7. *Duda A., Czachórski T.* Performance evaluation of fork and join synchronization primitives // Acta Informatica. — 1987. — Vol. 24, no. 5. — P. 525–533.
8. *Kim M. Y., Tantawi A. N.* Asynchronous disk interleaving: Approximating access delays // IEEE Transactions on Computers. — 1991. — Vol. 40, no. 7. — P. 801–810.
9. *Fiorini P. M.* Exact Analysis of Some Split Merge Queues. // SIGMETRICS Performance Evaluation Review. — 2015. — Vol. 43, no. 2 — P. 51–53.
10. *Moiseeva S., Sinyakova I.* Investigation of Queueing system GI(2)|M2|∞ // Modern Probabilistic Methods for Analysis and Optimization of Information and Telecommunication Networks / Proc. of the Int. Conf. — 2011. — P. 219–225.
11. *Moiseeva S., Sinyakova I.* Investigation of Output Flows in the System with Parallel Service of Multiple Requests // Problems of Cybernetics and Informatics (PCI-2012) : IV International Conference (IEEE). Baku, Azerbaijan. — 2012. — P. 180–181.
12. *Joshi G., Soljanin E., Wornell G.* Efficient redundancy techniques for latency reduction in cloud systems // arXiv preprint arXiv:1508.03599. — 2015.
13. *Gorbunova A. V., Kradenyh A. A., Zaryadov I. S.* The mathematical model of a cloud computing system // Proceedings of the Nineteenth International Scientific Conference: Distributed computer and communication networks: control, computation, communications (DCCN-2016). — Vol. 3: Youth School–Seminar. — 2016. — P. 169–175.
14. *Bocharov P.P., D'Apice C., Pechinkin A.V., Salerno S.* Queueing Theory. — VSP, Utrecht, Boston, 2004.
15. *Harrison P., Zertal S.* Queueing Models with Maxima of Service Times // Computer Performance Evaluation. Modelling Techniques and Tools. Lecture Notes in Computer Science. — 2003. — Vol. 2794. — P. 152–168.