UDC 519.218.31

On Optimization of Energy Consumption in Cloud Computing System

Anastasia V. Daraseliya*, Eduard S. Sopin*†, Vladimir V. Rykov‡

* *Department of Applied Probability and Informatics*
Peoples' Friendship University of Russia
Miklukho-Maklaya str. 6, Moscow, 117198, Russia

† *Institute of Informatics Problems, FRC CSC RAS*
44-2 Vavilov Str., Moscow 119333, Russia

‡ *Departmen of Applied Mathematics and Computer Modeling*
Gubkin Russian State University of Oil and Gas
65 Leninsky Prospekt, Moscow, 119991, Russia

Email: nastyadar6@gmail.com, sopin_es@rudn.university, vladimir_rykov@mail.ru

We constructed mathematical models of cloud computing systems taking into account various mechanisms for increasing energy efficiency in terms of queuing theory, and analytical expressions for the main characteristics of energy consumption and server performance metrics were obtained. We investigated and compered three different energy efficiency improving mechanisms of cloud computing systems: the shutdown delay mechanism, the switch on delay mechanism and the threshold-based switch on mechanism. The general principle of functioning mechanisms for energy efficiency improving is that all mechanisms try to find a middle ground between continuous operation without shutdowns and with switching on as soon as it remains empty. We formulated the energy consumption optimization problem of the cloud computing system for each parameter used in this energy efficiency mechanisms. We conducted a numerical analysis of the formulas for solving the optimization problem of energy consumption in cloud computing system based on the initial data close to the real ones.

Key words and phrases: cloud computing, energy efficiency, queuing system, optimization.

In: K. E. Samouylov, L. A. Sevastianov, D. S. Kulyabov (eds.): Selected Papers of the 12th International Workshop on Applied Problems in Theory of Probabilities and Mathematical Statistics (Summer Session) in the framework of the Conference "Information and Telecommunication Technologies and Mathematical Modeling of High-Tech Systems", Lisbon, Portugal, October 22–27, 2018, published at http://ceur-ws.org

1. Introduction

Recently, the concept of energy efficiency improving in cloud computing systems is becoming popular. There are various methods to implement this. One way to improve energy efficiency is scheduling and load balancing the servers, VMs, and applications [3]. The servers can be put into standby state in order to improve the energy efficiency of a cloud system in case of light load. On the one hand, the switching to standby mode allows to reduce power consumption, and on the other hand, it leads to extra power usage to turn on/off the server. Therefore, it is important to understand under what conditions it will be advantageous to put the server in standby state, and under what conditions it is more profitable to leave it in the operating mode.

Moreover since the service-level agreement (SLA) must not be violated, the provider needs to maintain the required level of energy consumption. However, while maintaining the SLA, one of the parameters is the response time, so here we consider the optimization problem of energy consumption with a constraints on the response time.

2. Modeling of energy efficiency improvement mechanisms

We consider a baseline model [7] as a single-server queuing system described by the Markov process with Processor Sharing policy where the maximum numbers of the customers is C. We do not consider distribution of processing volume of a task in the paper, however, it can be done by means of queuing systems with limited resources [6]. Customers arrive according to the Poisson law with rate λ. Service times, switch on and switch off durations are exponentially distributed with the parameters μ, α and β, respectively. The system state is described by the vector (s, k), where k is the number of customers in the system, s is the server state. Here $s = 0$ means that the system is in the standby mode, s=1 reflects switch-on mode and s=2 and s=3 represent operating and switch off modes, respectively. Arrival of a customer in an empty system cause change of the system state to the switch on mode. After exponentially distributed time with rate α, the system switches to the operating mode, in which serving of customers is started. When the system remains empty in the operating mode, it switches off immediately. Fig. 1 shows the transition intensities diagram for the baseline model. For the base model, the set of states S_1 is represented in the following form: $S_1 = \{(s,k)|s = 1, 2,\ 1 \le k \le C\} \cup \{(s,k)|s = 3,\ 0 \le k \le C\} \cup (0,0)$ [4, 5]. In [9] we derive the system of equilibrium equations, based on the transition intensity diagram [8], [7], which makes it possible to obtain stationary probabilities $p_{s,k}$ that the system is in (s,k) state.

Due to the high energy consumption for shutting down the cloud server, in some cases it's more beneficial to leave it in operating mode pending the arrival of new customers. In [8] we consider the model with server shutdown delay mechanism. In contrast to the base model, where it was assumed that the server shuts down as soon as it remains empty, in this model the system does not switch off immediately, but waits exponentially distributed time with rate γ. If a customer arrives during that waiting period, then the system starts serving. Otherwise, the state is changed to the switch off mode. If a customer arrives during the switch off mode, then the system turns to the switch on mode immediately after the completion of the switch off. Otherwise, the system falls to the stand by mode. Fig. 2 shows the transition intensities diagram for the model with the shutdown delay mechanism. The set of states for this model is represented in the following form: $S_2 = \{(s,k)|s = 1,\ 1 \le k \le C\} \cup \{(s,k)|s = 2, 3,\ 0 \le k \le C\} \cup (0,0)$. In [8] we derive and solve the system of equilibrium equations for the model with shutdown delay mechanism.

Also we consider the model with server switch on delay, as well as in base model, system passes in switch off mode at once after it remains empty. But it does not switch on immediately on arrival of a new customer, and waits exponentially distributed time with rate θ. Fig. 3 shows the transition intensities diagram for this model. For this system, the set of states S_3 is represented in the following form: $S_3 = \{(s,k)|s = 0, 3,\ 0 \le k \le$

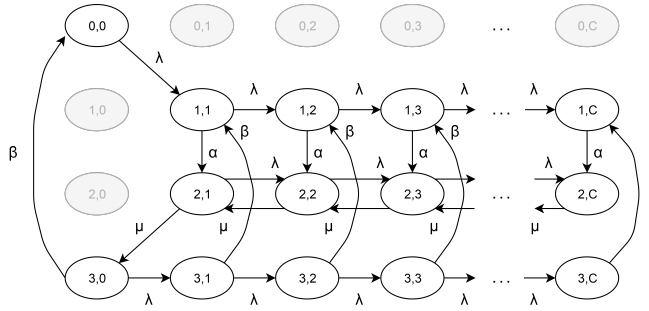

Figure 1. Transition intensities diagram. Baseline mathematical model

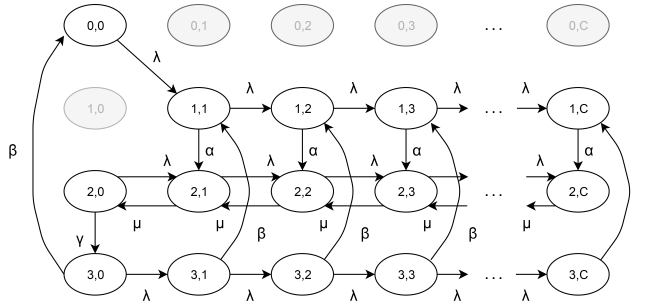

Figure 2. Transition intensities diagram. Mathematical model with the shutdown delay mechanism

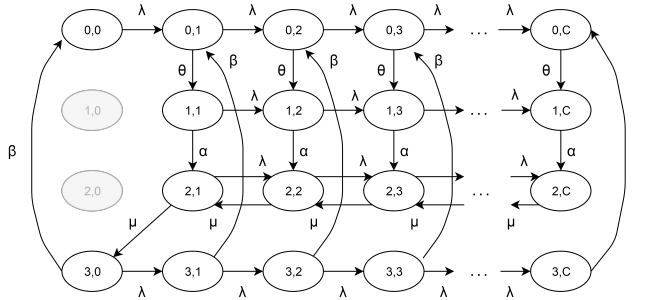

Figure 3. Transition intensities diagram. Mathematical model with the switch on delay mechanism

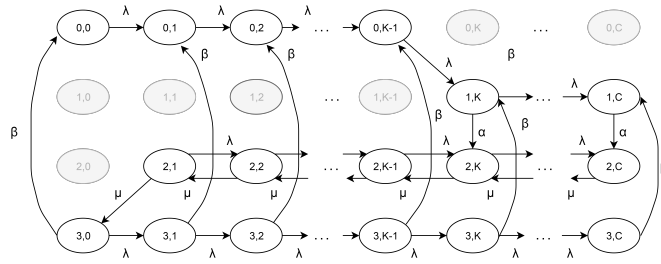

Figure 4. Transition intensities diagram. Mathematical model with the threshold-based switch mechanism

$C\} \cup \{(s,k)|s=1,2,\ 1 \le k \le C\}$. In [9] we derive the system of equilibrium equations for the model with the switch on delay mechanism.

Then we consider the mode with the threshold-based switch mechanism, in which system passes from standby mode in switch on mode only after arrived of a certain number κ of customers. Fig. 4 shows the transition intensities diagram for this model. For this system, the set of states S_4 is represented in the following form: $S_4 = \{(s,k)|s=0,\ 0 \le k \le \kappa-1\} \cup \{(s,k)|s=1,\ \kappa \le k \le C\} \cup \{(s,k)|s=2,\ 1 \le k \le C\} \cup \{(s,k)|s=3,\ 0 \le k \le C\}$. In [9] we derive the system of equilibrium equations for this model with the threshold-based switch mechanism.

We derived the system of equilibrium equations for each model, based on the transition intensity diagrams, which makes it possible to obtain stationary probability distribution of the system. Taking into account the normalization condition and using matrix methods, the system of equilibrium equations can be solved numerically, but we represent the analytical solution in [8].

3. Energy consumption indicators and the performance characteristics of cloud systems

With the system stationary distribution, we calculate the energy consumption indicators. We will assume that in the switch on / off mode, the power consumption is constant and equal to the average values P_1 and P_3, respectively. In the operating mode, the power consumption $P_{2,k}$ depends on the server occupancy. Through $P_{2,max}$ we denoted the maximum value of the server's power consumption in the operating mode, and through $P_{2,min}$ we denoted the power consumption in idle mode. The energy consumption in the standby mode will be calculated by P_0. By analogy with the formula given in [2], we derive the formula for the average server power consumption:

$$P = P_0 \sum_{i=0}^{C} p_{0,i} + P_1 \sum_{i=0}^{C} p_{1,i} + P_3 \sum_{i=0}^{C} p_{3,i} + \sum_{i=0}^{C} P_{2,i} p_{2,i}$$

where

$$P_{2,k} = P_{2,min} + \frac{P_{2,max} - P_{2,min}}{C} k$$

According to Little's law, the average number N of customers in the system is equal to the average effective arrival rate $\lambda(1-\pi)$ multiplied by the average sojourn time T, where blocking probability π is

$$\pi = p_{0,C} + p_{1,C} + p_{2,C} + p_{3,C}$$

The average number N of customers is given by

$$N = \sum_{k=0}^{3} \sum_{i=1}^{C} i p_{k,i}$$

The average response time T follows directly from Little's law and formula (3):

$$T = \frac{\sum_{k=0}^{3} \sum_{i=1}^{C} i p_{k,i}}{\lambda(1-\pi)}$$

4. Optimization problem

In order to understand under what conditions it will be advantageous to put the server in standby state, and under what conditions it is more profitable to leave it in the operating mode, it is necessary to formulate and solve the energy consumption optimization problem for the each parameters of the energy efficiency increasing mechanisms.

The optimization problem can be formulated as

$$\begin{cases} P \to \min, \\ R1 : T \le T_0, \\ R2 : P \ge 0, \end{cases}$$

where the energy consumption P of the cloud system is minimized under constraint T_0 on the average response time threshold.

For a model with the shutdown delay mechanism minimizing the energy consumption P by the parameter γ can be written as follows:

$$\begin{cases} P(\gamma) \to \min, \\ R1 : T \le T_0, \\ R2 : P \ge 0, \end{cases}$$

By analogy, we can write down the minimization problem for models with the switch on delay and the threshold-based switch mechanisms through the parameters θ and κ

$$\begin{cases} P(\theta) \to \min, \\ R1 : T \le T_0, \\ R2 : P \ge 0, \end{cases}$$

$$\begin{cases} P(\kappa) \to \min, \\ R1 : T \le T_0, \\ R2 : P \ge 0. \end{cases}$$

For each of these three mechanisms, the optimization problem was considered separately.

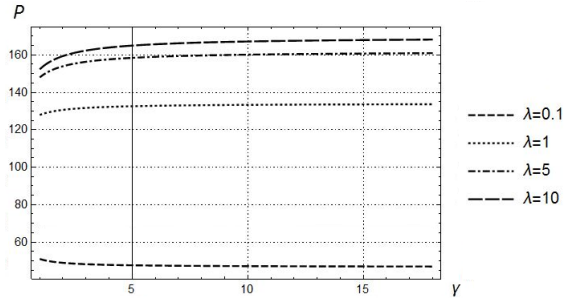

Figure 5. The dependence of the power consumption P on the rate γ. Model with the shutdown delay mechanism

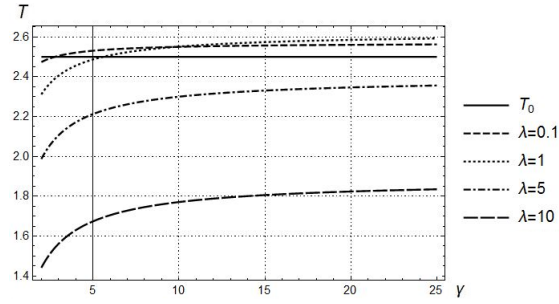

Figure 6. The dependence of the average response time T on the rate γ. Model with the shutdown delay mechanism

5. Numerical analysis

In this section we present results a numerical analysis of the formulas to solved the optimization problem.

On the energy profile of the cloud system installed at the University of Cardiff [1], it can be seen [1] that the inclusion of the server lasts 150 seconds, and the shutdown is 30 seconds. Further, for convenience, it was represented in minutes. The values of P_i were taken from [1], according to which $P_0 = 10$ W, $P_1 = 170$ W, $P_3 = 120$ W, $P_{2,min} = 105$ W and $P_{2,max} = 268$ W.

The results of numerical analysis for the values C=20, μ=20, α=1, β=2 and $T_0 = 2.5$ are presented in Fig. 5–10.

The plots of the server's power consumption for the model with the shutdown delay mechanism (Fig. 5) and for the model with the switch on delay mechanism (Fig. 7) show that the consumed power increases very fast for small values of the arrival flow intensity λ.

The plots of the average response time T(Fig. 6) for the model with the shutdown delay mechanism show that the greatest dependence of the average sojourn time T on the arrival flow intensity λ is observed at values of γ from 1 to 7. Also note, that for the

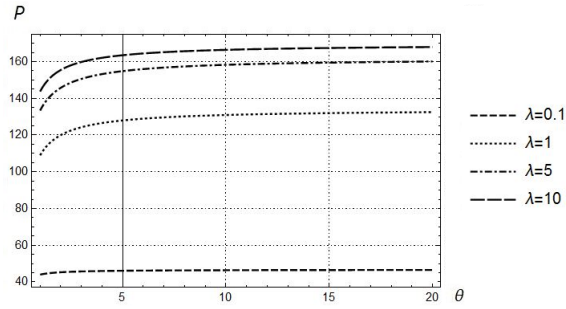

Figure 7. The dependence of the power consumption P on the rate θ. Model with the switch on delay mechanism

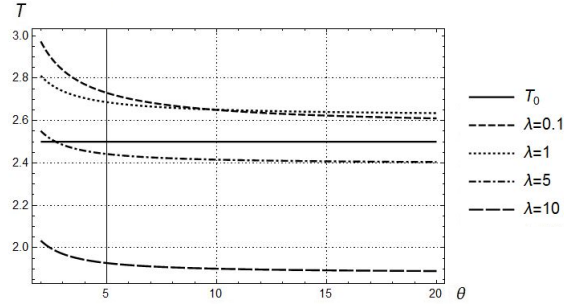

Figure 8. The dependence of the average response time T on the rate θ. Model with the switch on delay mechanism

arrival flow intensity $\lambda = 1$, the condition $R1$ of the optimization problem is performed with γ less than 5. For $\lambda = 5$, the condition $R1$ of the optimization problem is satisfied on the whole segment of the function.

In Fig. 8 note that for small values of θ, the difference in the average sojourn time T is the greatest. For the arrival flow intensity $\lambda = 5$, the condition $R1$ of the optimization problem is fulfilled when the value of θ is greater than 3.

The plots of the average response time T(Fig. 10) for the model with the threshold-based switch mechanism show that for the large values of κ, the difference in the average sojourn time T is the greatest, and vice versa, the small κ values have almost no effect on the average sojourn time. For the arrival flow intensity $\lambda = 5$, the condition $R1$ of the optimization problem is fulfilled when the value of κ is less than 6.

6. Conclusions

We consider a cloud computing system with three different energy efficiency improving mechanisms as a system with Processor Sharing policy. We investigate how a waiting time before a server goes to switch on / off mode and threshold-based switch affects

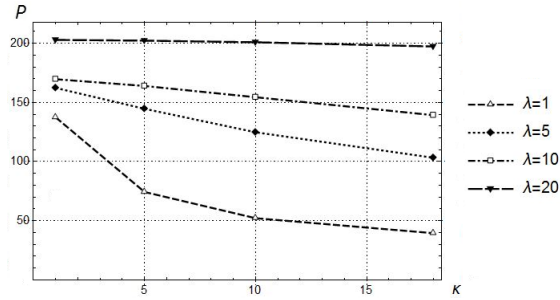

Figure 9. The dependence of the power consumption P on the parameter κ. Model with the threshold-based switch mechanism

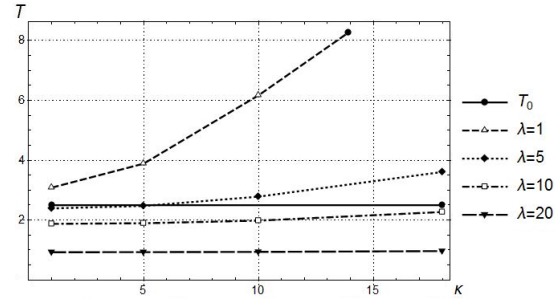

Figure 10. The dependence of the average response time T on the parameter κ. Model with the threshold-based switch mechanism

the energy efficiency of a cloud system. We carried out a numerical analysis of the formulas for solving the energy consumption optimization problem. Numerical analysis showed that that the server switch on mechanism is most efficient in terms of power consumption, but the server shutdown delay mechanism allows the system to work at a lower system load and is more effective in terms of response time. The mechanism with server threshold-based switch on gives an improvement for power, but deterioration in time.

Acknowledgments

The publication has been prepared with the support of the "RUDN University Program 5-100" and funded by RFBR according to the research projects No. 18-07-00576 and No. 19-07-00933.

References

1. J. Conejero, O. Rana, P. Burnap, J. Morgan, B. Caminero, C. Carrion: Analysing Hadoop Power Consumption and Impact on Application QoS. In: Future Generation

Computer Systems, vol.55, Issue C, pp. 213–223. (2016). doi:10.1016/j.future.2015.03.009
2. A. Beloglazov, J. Abawajy, R. Buyya: Energy-aware resource allocation heuristics for efficient management of data centers for Cloud computing. In: Future Generation Computer Systems, vol.28, pp. 755 – 768. (2012). DOI: 10.1016/j.future.2011.04.017
3. G. Valentini et al.: An Overview of Energy Efficiency Techniques in Cluster Computing Systems. In: Cluster Computing, vol. 16, no. 1, pp. 3–15. (2013). doi:10.1007/s10586-011-0171-x
4. Gaidamaka, Y., Pechinkin, A., Razumchik, R., Samouylov, K., Sopin, E. Analysis of an MG1R queue with batch arrivals and two hysteretic overload control policies (2014) International Journal of Applied Mathematics and Computer Science, 24 (3), pp. 519-534. doi:10.2478/amcs-2014-0038
5. Samouylov, K.E., Abaev, P.O., Gaidamaka, Y.V., Pechinkin, A.V., Razumchik, R.V. Analytical modelling and simulation for performance evaluation of sip server with hysteretic overload control (2014) Proceedings - 28th European Conference on Modelling and Simulation, ECMS 2014, pp. 603-609. doi:10.7148/2014-0603
6. Naumov, V., Samouylov, K. Analysis of multi-resource loss system with state-dependent arrival and service rates (2017) Probability in the Engineering and Informational Sciences, 31 (4), pp. 413-419. doi:10.1017/S0269964817000079
7. Daraseliya A.V., Sopin E.S., Energy efficiency analysis of Cloud Computing system with setup and vacation perion of server. In.: Information and telecommunication technologies and mathematical modeling of high-tech systems (ITTMM-2017)», pp. 119–121. (2017).
8. Daraseliya A.V., Sopin E.S., Analysis of an approach to increase energy efficiency of a cloud computing system. In: Selected Papers of the II International Scientific Conference "Convergent Cognitive Information Technologies" (Convergent 2017), vol-2064, pp. 79–87. CEUR Workshop Proceedings, Moscow (2017). http://ceur-ws.org/Vol-2064/paper09.pdf
9. A. V. Daraseliya, E.S.Sopin, A. K. Samuylov, S.Ya. Shorgin, Comparative analysis of the mechanisms for energy efficiency improving in cloud computing systems. In.: The 18th International Conference on Next Generation Wired/Wireless Advanced Networks and Systems (NEW2AN-2018): Internet of Things, Smart Spaces, and Next Generation Networks and Systems, pp 268-276. (2018). doi:10.1007/978-3-030-01168-0_25