



Comparative Analysis of the Mechanisms for Energy Efficiency Improving in Cloud Computing Systems

A. V. Daraseliya¹(✉) , E. S. Sopin^{1,2} , A. K. Samuylov¹ ,
and S. Ya. Shorgin^{1,2}

¹ Peoples' Friendship University of Russia (RUDN University), Moscow, Russia
nastyadar6@gmail.com, {sopin.es, samuylov.ak}@rudn.university,
sshorgin@ipiran.ru

² Institute of Informatics Problems, FRC CSC RAS, Moscow, Russia

Abstract. We consider a cloud computing system with three different mechanisms for increasing the energy efficiency of cloud computing systems. We investigate how energy efficiency of a cloud system is affected by a waiting time before a server goes to switch on/standby mode and threshold-based switch. We developed four mathematical models of Cloud computing system in terms of the queuing system and derived the system of equilibrium equations, which makes it possible to obtain the energy consumption indicators.

Keywords: Cloud computing · Energy efficiency · Queuing system

There are various mechanisms and approaches to improve the energy efficiency of the cloud computing system. One way to improve energy efficiency is to upgrade network equipment by implementing power-saving modes and adaptive transmission rates to achieve proportional power usage. Another example of an approach to improve energy efficiency is scheduling and load balancing the servers, VMs, and applications [4]. Speaking about VMs, several approaches can be taken for improving energy efficiency. For example, VM self-adaption and hardware adjustment and choosing the most efficient physical machines for VM placements. Finally, VM migrations allow dynamic consolidation of physical machines by moving underutilized VMs onto fewer hosts and powering off the unused ones, that permits reduced resource consumption at any given moment [3]. Similar to virtual machines, it is possible to manage the power consumption of cloud computing servers.

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

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. 16-07-00766 and No. 18-07-00576.

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. In this regard, cloud providers employ various mechanisms that decrease server switching number. In this article we consider three mechanisms to improve the energy efficiency of the cloud computing server: the shutdown delay mechanism, the switch on delay mechanism and the threshold-based switch mechanism.

1 Analysis of Mathematical Models of Mechanisms for Increasing the Energy Efficiency of Cloud Systems

We proceed from the fact that arriving customers are distributed evenly on the virtual machine. Since we assumed that the computing resources are distributed evenly across all servers, we will consider a system consisting of one server with Processor Sharing policy.

We consider a base model [6] as a single-server queuing system with C virtual machines. 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.

For the base model, the set of states S_1 is represented in the following form: $S_1 = \{(s, k) | s = 1, 2, 1 \leq k \leq C\} \cup \{(s, k) | s = 3, 0 \leq k \leq C\} \cup (0, 0)$.

We derive the system of equilibrium equations, based on the transition intensity diagram [5, 6], which makes it possible to obtain stationary probabilities $p_{s,k}$ that the system is in (s,k) state:

$$p_{3,0} = \frac{\lambda}{\beta} p_{0,0}; \tag{1}$$

$$p_{1,1} = \frac{\lambda}{(\lambda + \alpha)} p_{0,0} + \frac{\beta}{(\lambda + \alpha)} p_{3,1}; \tag{2}$$

$$p_{1,k} = \frac{\lambda}{(\lambda + \alpha)} p_{1,k-1} + \frac{\beta}{(\lambda + \alpha)} p_{3,k}, \quad 2 \leq k \leq C - 1; \tag{3}$$

$$p_{1,C} = \frac{\lambda}{\alpha} p_{1,C-1} + \frac{\beta}{\alpha} p_{3,C}; \tag{4}$$

$$p_{2,k} = \frac{(\lambda + \mu)}{\mu} p_{2,k-1} - \frac{\lambda}{\mu} p_{2,k-2} - \frac{\alpha}{\mu} p_{1,k-1}, \quad 3 \leq k \leq C; \tag{5}$$

$$p_{2,C} = \frac{\alpha}{\mu} p_{1,C} + \frac{\lambda}{\mu} p_{2,C-1}; \tag{6}$$

$$p_{2,1} = \frac{(\lambda + \beta)}{\mu} p_{3,0}; \tag{7}$$

$$p_{2,2} = \frac{(\lambda + \mu)}{\mu} p_{2,1} - \frac{\alpha}{\mu} p_{1,1}; \tag{8}$$

$$p_{3,k} = \frac{\lambda}{(\lambda + \beta)} p_{3,k-1}, \quad 1 \leq k \leq C - 1; \tag{9}$$

$$p_{3,C} = \frac{\lambda}{\beta} p_{3,C-1}; \tag{10}$$

$$\sum_{i=0}^3 \sum_{j=0}^C p_{i,j} = 1. \tag{11}$$

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 [5] 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. For this model, the set of states is represented in the following form: $S_2 = \{(s, k) | s = 1, 1 \leq k \leq C\} \cup \{(s, k) | s = 2, 3, 0 \leq k \leq C\} \cup (0, 0)$.

The system of equations for the model with shutdown delay mechanism differs from the system of equations for the base model by the following formulas: we added formula (12), and replaced formulas (7) and (8) by formulas (13) and (14).

$$p_{2,0} = \frac{\lambda + \beta}{\gamma} p_{3,0}; \tag{12}$$

$$p_{2,1} = \frac{\lambda + \gamma}{\mu} p_{2,0}; \tag{13}$$

$$p_{2,2} = \frac{(\lambda + \mu)}{\mu} p_{2,1} - \frac{\lambda}{\mu} p_{2,0} - \frac{\alpha}{\mu} p_{1,1}; \tag{14}$$

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 .

For this system, the set of states S_3 is represented in the following form: $S_3 = \{(s, k) | s = 0, 3, 0 \leq k \leq C\} \cup \{(s, k) | s = 1, 2, 1 \leq k \leq C\}$.

The system of equations for the model with switch on delay mechanism differs from the system of equations for the base model by the following formulas: we

added formulas (18) and (19), and replaced formulas (2)–(4) by formulas (15)–(17).

$$p_{1,1} = \frac{\theta}{(\lambda + \alpha)} p_{0,1}; \tag{15}$$

$$p_{1,k} = \frac{\theta}{(\lambda + \alpha)} p_{0,k}, \quad 2 \leq k \leq C - 1; \tag{16}$$

$$p_{1,C} = \frac{\lambda}{\alpha} p_{1,C-1} + \frac{\theta}{\alpha} p_{0,C}; \tag{17}$$

$$p_{0,k} = \frac{\lambda}{\lambda + \theta} p_{0,k-1} + \frac{\beta}{\lambda + \theta} p_{3,k}, \quad 1 \leq k \leq C - 1; \tag{18}$$

$$p_{0,C} = \frac{\lambda}{\theta} p_{0,C-1} + \frac{\beta}{\theta} p_{3,C}; \tag{19}$$

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.

For this system, the set of states S_4 is represented in the following form: $S_4 = \{(s, k) | s = 0, 0 \leq k \leq \kappa - 1\} \cup \{(s, k) | s = 1, \kappa \leq k \leq C\} \cup \{(s, k) | s = 2, 1 \leq k \leq C\} \cup \{(s, k) | s = 3, 0 \leq k \leq C\}$.

The system of equations for this model differs from the system of equations for the base model by the following formulas: we added formula (25), and replaced formulas (2)–(4) by formulas (20) and (21) and formulas (5) and (6) by formulas (22)–(24).

$$p_{1,\kappa} = \frac{\lambda}{(\lambda + \alpha)} p_{1,\kappa-1} + \frac{\beta}{(\lambda + \alpha)} p_{3,\kappa}; \tag{20}$$

$$p_{1,k} = \frac{\lambda}{(\lambda + \alpha)} p_{1,k-1} + \frac{\beta}{(\lambda + \alpha)} p_{3,k}, \quad \kappa + 1 \leq k \leq C - 1; \tag{21}$$

$$p_{2,k+1} = \frac{(\lambda + \mu)}{\mu} p_{2,k} - \frac{\lambda}{\mu} p_{2,k-1}, \quad 2 \leq k \leq \kappa - 1; \tag{22}$$

$$p_{2,k+1} = \frac{(\lambda + \mu)}{\mu} p_{2,k} - \frac{\alpha}{\mu} p_{1,k} - \frac{\lambda}{\mu} p_{2,k-1}, \quad \kappa \leq k \leq C - 1; \tag{23}$$

$$p_{2,C} = \frac{\alpha}{\mu} p_{1,C} + \frac{\lambda}{\mu} p_{2,C-1}; \tag{24}$$

$$\lambda p_{0,k} = \lambda p_{0,k-1} + \beta p_{3,k}, \quad 1 \leq k \leq \kappa - 1; \tag{25}$$

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 [5].

2 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:

$$P = P_0 \sum_{i=0}^C p_{0,i} + P_1 \sum_{i=0}^C p_{1,i} + P_3 \sum_{i=0}^C p_{3,i} + \sum_{i=0}^C P_{2,i} p_{2,i} \tag{26}$$

where

$$P_{2,k} = P_{2,min} + \frac{P_{2,max} - P_{2,min}}{C} k \tag{27}$$

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 . Expressed algebraically the law is

$$N = \lambda(1 - \pi)T \tag{28}$$

where blocking probability π is

$$\pi = p_{0,C} + p_{1,C} + p_{2,C} + p_{3,C} \tag{29}$$

The average number N of customers is given by

$$N = \sum_{k=0}^3 \sum_{i=1}^C i p_{k,i} \tag{30}$$

The average response time T follows directly from formulas (28) and (30):

$$T = \frac{\sum_{k=0}^3 \sum_{i=1}^C i p_{k,i}}{\lambda(1 - \pi)} \tag{31}$$

3 Numeral analysis

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 $C = 20$, $\mu = 20$, $\alpha = 1$, $\beta = 2$ are presented in Figs. 1, 2, 3, 4, 5 and 6.

The plots of the server's power consumption (Fig. 1) for base model show that the consumed power increases very fast for small values of the arrival flow intensity λ , also note that with the increase of waiting time, during which the system doesn't go into standby mode, the power consumption also increases.

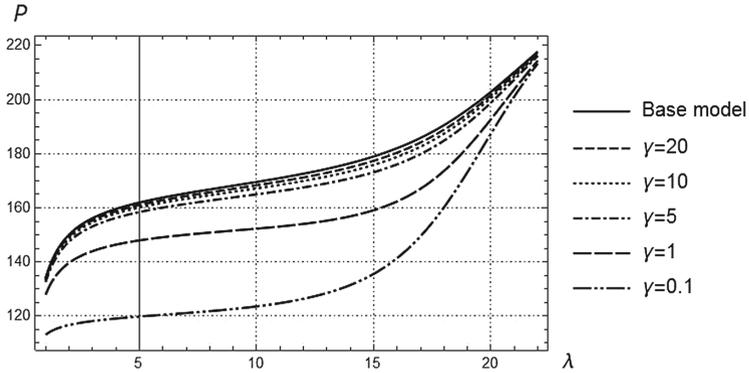


Fig. 1. The dependence of the power consumption P on the arrival flow intensity λ . Comparison of the model with the shutdown delay mechanism and the base model.

In Fig. 2 it is clearly seen 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 5.

The graph of the server power consumption P (Fig. 3) for the model with server shutdown delay increases most sharply at small values of the arrival flow intensity λ , also note that with an increase in the exponential time θ during which

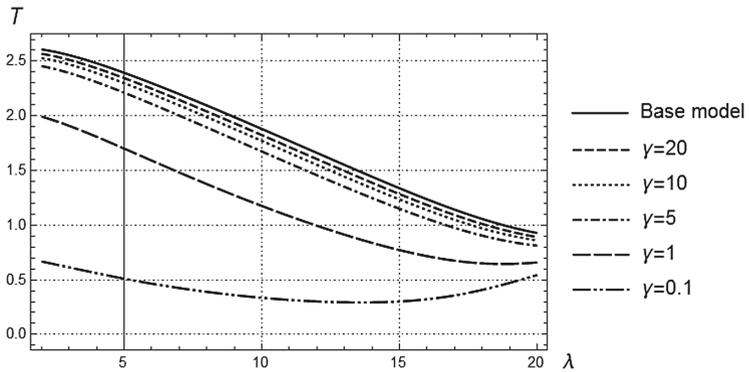


Fig. 2. The dependence of the average response time T on the arrival flow intensity λ . Comparison of the model with the shutdown delay mechanism and the base model.

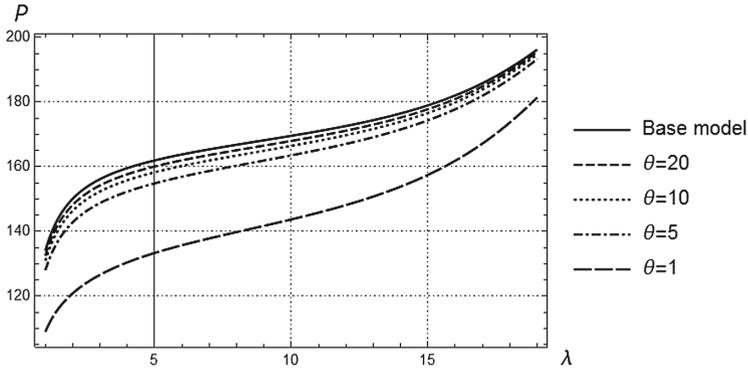


Fig. 3. The dependence of the power consumption P on the arrival flow intensity λ. Comparison of the model with the switch on delay mechanism and the base model.

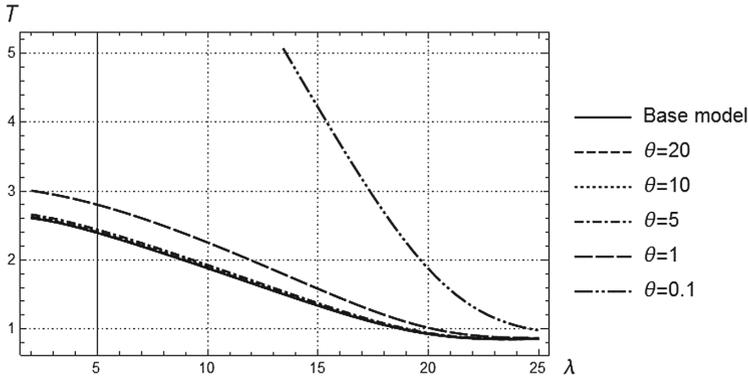


Fig. 4. The dependence of the average response time T on the arrival flow intensity λ. Comparison of the model with the switch on delay mechanism and the base model.

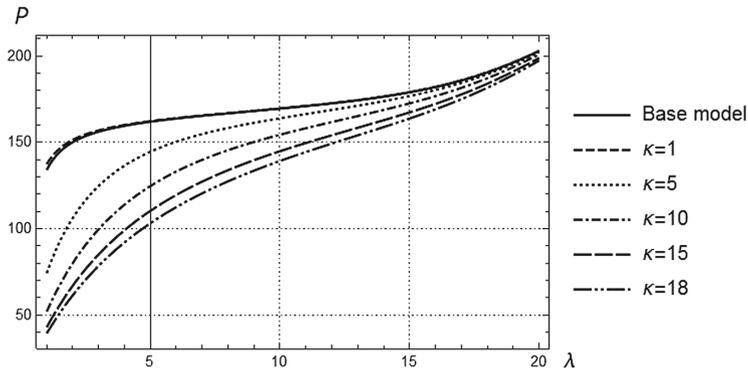


Fig. 5. The dependence of the power consumption P on the arrival flow intensity λ. Comparison of the model with the threshold-based switch mechanism and the base model.

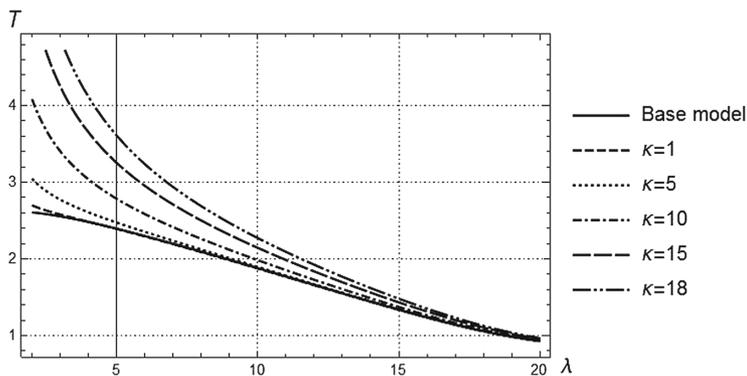


Fig. 6. The dependence of the average response time T on the arrival flow intensity λ . Comparison of the model with the threshold-based switch mechanism and the base model.

the system doesn't go into the operating mode after the first customer is received, the value of the power consumption increases accordingly. Note that for $\theta = 20$, the graph of the model with server shutdown delay is most closely approximated to the base model's graph, where the system was included immediately after the customer was received.

In Fig. 4 note that for small values of θ , the difference in the average sojourn time T is the greatest.

For each of the three models, it was noted that when $\gamma \rightarrow \infty$ (see Figs. 1 and 2), $\theta \rightarrow \infty$ (see Figs. 3 and 4) and $\kappa \rightarrow 1$ (see Figs. 5 and 6), respectively, each of the three systems tends to the initial baseline model.

4 Conclusion

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 conducted a comparative numerical analysis of the mechanisms for improving the energy efficiency of cloud computing based on the initial data close to the real ones. Numerical analysis showed that the mechanism with server shutdown delay gives an improvement in both power and response time, for a mechanism with server enable delay, there is an improvement at almost all intervals of power consumption, except service time at low loads. The mechanism with server thresholding gives an improvement for power, but deterioration in time.

References

1. Conejero, J., Rana, O., Burnap, P., Morgan, J., Caminero, B., Carrion, C.: Analysing Hadoop power consumption and impact on application QoS. *Futur. Gener. Comput. Syst.* **55**(C), 213–223 (2016). <https://doi.org/10.1016/j.future.2015.03.009>
2. Beloglazov, A., Abawajy, J., Buyya, R.: Energy-aware resource allocation heuristics for efficient management of data centers for cloud computing. *Futur. Gener. Comput. Syst.* **28**, 755–768 (2012). <https://doi.org/10.1016/j.future.2011.04.017>
3. Mastelic, T., Brandic, I.: Recent trends in energy-efficient cloud computing. *IEEE Cloud Comput. Mag.* **2**, 40–47 (2015). <https://doi.org/10.1109/MCC.2015.15>
4. Valentini, G.: An overview of energy efficiency techniques in cluster computing systems. *Clust. Comput.* **16**(1), 3–15 (2013). <https://doi.org/10.1007/s10586-011-0171-x>
5. 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>
6. 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)