Service Performance Analysis of Cloud Computing Server by Queuing System

Ruijuan Wang[1], Guangjun Zai[1(✉)], Yan Liu[2,3], and Haibo Pang[1]

[1] School of Software, Zhengzhou University, Zhengzhou, China
zaiguangjun@zzu.edu.cn
[2] School of Mathematics and Statistics, Zhengzhou University, Zhengzhou, China
[3] Henan Key Laboratory of Financial Engineering, Zhengzhou, China

Abstract. Performance analysis of cloud computing server provides the basis for ensuring Quality of Service (QoS), and the service strategy of server will directly affect the analysis of performance indicators. The performance indicators of QoS are usually defined in the form of Service Layer Agreement (SLA), such as the average response time, the average queue length, immediate service probability and so on. In this work, Service performance analysis models based on $Geo/G/1$ queuing system and queuing system with the vacation of the server are proposed. In these models, we analyze the main performance indicators of cloud computing server for the different parameters: the time between arrive of the task, the time of service, and the time of the provision of vacation. Furthermore, we discuss the optimizing concurrent number of the cloud computing.

Keywords: $Geo/G/1$ queuing system · Cloud computing · Performance analysis · Response time

1 Introduction

Cloud computing is the further development of distributed computing, parallel processing and grid computing, and quickly becomes a hot spot in industry and academic research [1, 2]. QoS performance metrics typically includes response time, blocking probability and probability of immediate service. If the cloud computing center is assumed to be a random service system, each computing resource (e.g., network server, database server) can be deemed as a service platform. To build a stochastic service system, the QoS performance metrics can be discussed by the research method of queuing theory.

The performance indicators of QoS are usually defined in the form of Service Layer Agreement (SLA). SLA is a consultative agreement between customers and service providers, including service availability, performance, and data protection and security [3]. Accurate prediction of customer service needs can enable service providers to avoid over provisioning resources in order to satisfy SLA. Customer request is a variable load. It is difficult to satisfy SLA and achieve optimal utilization of resources by dynamically

S. Deng et al. (Eds.): MobiCASE 2021, LNICST 434, pp. 42–53, 2022.
https://doi.org/10.1007/978-3-030-99203-3_3

configuring computing resources [4]. However, we can analyze the service performance by modeling the cloud computing center, and obtain the probability distribution of the request response time and other performance indicators, and discuss the factors that affect the performance of the cloud computing center.

At present, the use of queuing theory calculation service performance analysis and modeling centers on the cloud technology research work is very little to this theory research. In 2012, KHAZAEI H et al. [5] used the M/G/m queuing model to analyze the relationship between the average response time, average queue length and the number of service station. B. Yang et al. [6] in 2009 and H. W. He et al. [7] in 2014 proposed the approximate analysis model of computing center of M/M/n/n+r queuing system based on the cloud, which obtained the user request response distribution function of time and other important performance metrics QoS through solving the model. K. Q. Xiong provided a M/M/1 Queuing System, which was a response time model for servers providing Web services. The probability distribution of the response time and its mean value are given [8]. In 2013, H. Wang set up a set of mapping rules that mapped the various "internal causes" and "external causes" affecting the performance of Web services to G/G/1 FCFS, M/G/1 PS and M/G/ infinity queuing nodes. Further more, he gave a group of Web services based on the queuing network and proposed the combination of performance analysis index system and its calculation formula [9].

In a queuing system, service stations will adopt various strategies of customer reception in some time, and this service is temporarily interrupted time (usually is random variable) is referred to the vacation. In 2015, we presented a queuing system part of the service station based on asynchronous multiple vacation and the service process of cloud computing center corresponding queuing system, by the analysis of the corresponding queuing model, to investigate the performance index and optimize the allocation of resources in cloud computing [10].

2 Cloud Computing Server Queuing System Model

The user sends task to the cloud computing center and a suitable computing node provides service. After service completes, the task leaves. Compute node contains different computing resources, such as network server, database server. The assumption of a computing node corresponds to queuing service of one system, the user sends the corresponding queuing system task arrival process, user services complete the service process corresponding to the queuing system, the process of resource allocation in cloud computing server can correspond to a queue model.

So the process of resource allocation in cloud computing server may correspond to a part of the queuing system $Geo/G/1$ [11, 12].

(i) The time between arrive of the task is time series $\{\tau_k, k = 1, 2, \cdots\}$, and τ_k is independent and identical distribution (i.i.d.). The distribution of τ_k is the Geometric Distribution with the parameter p,

$$P\{\tau = j\} = p(1-p)^{j-1}, j = 1, 2, \cdots, 0 < p < 1$$

(ii) In the system there is one server, and the time of service is variable χ. The distribution of χ is a Discrete Distributions

$$P\{\chi = j\} = g_j, j = 1, 2, \cdots$$

Cumulative distribution function of service time is $G(j)$, and the Probability generating function is.

$$G^*(z) = \sum_{j=1}^{\infty} g_j z^j, |z| < 1$$

And the average service time per request is $\mu(1 \leq \mu < \infty)$. Customer service according to FCFS service rules.

(iii) The arrival time and the service time are independent.
The variable $N(n)$ is the number of customer in the service system, that is, queue length of the queuing system. Supposing $\rho = p\mu$ is the traffic intensity of the service system, and $\overline{p} = 1 - p$. We can analyze the quality of service by a discrete time $Geo/G/1$ Queuing service system.

3 Model Analysis and Service Performance Metrics by *Geo*/*G*/1

The $N(n)$ represents the task numbers in the cloud computing system at the time of n, and distribution of steady queue length

$$p_j = \lim_{n\to\infty} P\{N(n) = j\}, j = 0, 1, 2, \cdots$$

(i) when $\rho \geq 1, p_j = 0, j = 0, 1, 2, \cdots$.
(ii) when $\rho < 1$, the recursion expressions of $\{p_j, j = 0, 1, 2, \cdots\}$ for the transient queue length distribution is [12]

$$p_0 = 1 - \rho,$$

$$p_1 = \frac{(1-\rho)(1-G^*(\overline{p}))}{G^*(\overline{p})},$$

$$p_j = \frac{1}{G^*(\overline{p})}\left\{p(1-\rho)\sum_{n=j-1}^{\infty}\sum_{k=n+1}^{\infty} g_k \binom{n}{j-1} p^{j-1}\overline{p}^{n-j+1} + \sum_{i=1}^{j-1} p_{j-i}\left(1 - G^*(\overline{p}) - \sum_{m=1}^{i}\sum_{k=m}^{\infty} g_k \binom{k}{m} p^m \overline{p}^{k-m}\right)\right\}, j = 2, 3, \cdots$$

where $j \leq 0, \sum_{i=1}^{j} = 0$.

This is related to service performance index of cloud computing center.

(1) **The probability of immediate service**
If a customer arrives and the existence of station service is free, customer service can be accepted immediately without waiting. The probability of this happening calls immediate service probability. By the teady queue length distribution, it is immediate service probability that can be got:

$$p_0 = 1 - \rho.$$

(2) **The time of the server idle**
The server idle is the time when the system has just turned empty until the time when a customer arrives. Supposing the variable $\hat{\tau}$ is the server idle, we can get that the $\hat{\tau}$ is the Geometric distribution

$$P\{\hat{\tau} = j\} = p(1-p)^{j-1}, \quad j = 1, 2, \cdots, 0 < p < 1$$

And the server busy period is the time when the system has a customer arrives until the time when the last customer leaves, expressed by variable b.Average time of continuous work of cloud computing server is the expectation of b, and its value is

$$E[b] = \begin{cases} \frac{\rho}{p(1-\rho)}, & \rho < 1, \\ \infty, & \rho \geq 1. \end{cases}$$

(3) **Average number of clients waiting for service on the server (total system request)**
According to the definition of the expectations, average number of clients waiting for service on the server is the average queue length $E[N]$ of the queuing system (the number of requests for queuing):

$$E[N] = \rho + \frac{p^2}{2(1-\rho)} E\left[\chi^2 - \chi\right],$$

where $E\left[\chi^2\right] = \sum_{j=1}^{\infty} j^2 g_j$.

(4) **The distribution of response time**
The sum of queue waiting time and accepting the service time is called the request response time. They are mutually independent. In the queuing model, the waiting time can be calculated by the following ways.

When $\rho < 1$, the waiting time $W(t)$ can be decomposed into the sum of all service time of requests for queuing in $Geo/G/1$ queuing system. The distribution of $W(t)$ is

$$H_W(t) = \sum_{j=0}^{\infty} p_j \times G^{(j+1)}(t),$$

where $G^{(j)}(t)$ is the j-Fold Convolution of $G(t)$.

(5) **The average response time**
The average response time of requests is the sum of the average waiting time of customers and the average business hours of customers in queuing system, so the average response time $T = E[W]$, That is,

$$\begin{aligned} T &= E[W] = E[N] \times E[\chi] \\ &= \mu\left(\rho + \frac{p^2}{2(1-\rho)} E\left[\chi^2 - \chi\right]\right). \end{aligned}$$

4 Service Performance Metrics by *Geo*/*G*/1 with the Vacation

In practical application, a service cloud computing server may interrupt occurs. On the other hand, in order to optimize the service system and improve the system of economic benefit, the relative leisure time server is to engage in other work. The vacation policy can provide great flexibility for the design and optimization of process control system. So when the cloud computing center operates computing resources in the dynamic configuration, the vacation of the server is needed to be considered to achieve the optimal allocation of resources utilization. Because the queue system with a variety of vacation policy is complex, the study and calculation of indicators is difficult. So the research on QoS performance of current did not consider the vacation of service station.

In a queuing system, service stations will adopt various strategies of customer reception in some time, this service is temporarily interrupted time (usually is random variable) is referred to the vacation. This paper presents a queuing system, in which part of the service station is based on asynchronous multiple vacation, and the service process of cloud computing center corresponds to queuing system, and by the analysis of the corresponding queuing model, investigates the performance index and optimizes the allocation of resources in cloud computing.

The user sends the cloud computing task for service in a suitable computing node. After service is completed, the task leaves. Computing node contains different computing resources, such as network server, database server. The assumption of a computing node corresponds to queuing service of one system, the corresponding queuing system task arrival process which the user sends and the service process corresponding to the queuing system which user services complete, can be the process of resource allocation in Cloud Computing corresponds to a queue model. The real cloud computing services is generally part of server vacation leave, leave for asynchronous start, and in the absence of the user can be repeatedly leave. So the process of resource allocation in cloud computing center may correspond to a part of the service station asynchronous multiple vacation queuing system $Geo/G/1$ [13, 14].

The system model discussed in this paper, the following specific vacation policy:

(1) The time of task arrival interval is time series $\{\tau_k, k = 1, 2, \cdots\}$, and τ_k is i.i.d. The distribution of τ_k is the Geometric Distribution with the parameter p,

$$P\{\tau = j\} = p(1-p)^{j-1}, j = 1, 2, \cdots, 0 < p < 1;$$

(2) In the system there is one service station, and the time of service is variable χ. The distribution of χ is a Discrete Distributions

$$P\{\chi = j\} = g_j, j = 1, 2, \cdots$$

Customer service according to FCFS service rules;

(3) A service platform not only can complete a customer service system and meet in the customer waiting, but also can continue for the next customer service.

(4) When there is no customers waiting for service in the system, the service station will be in the provision of vacation. If the customers arrive in the time of the provision of vacation, the service station must begin to service. If there are not customers in the time of the provision of vacation, the service station starts the vacation.
The time of the provision of vacation is random variable Y with the Discrete Distributions

$$P\{Y = j\} = y_j, j = 0, 1, 2, \cdots$$

Cumulative distribution function of provision time is $Y(j)$, and the Probability generating function is $Y^*(z) = \sum_{j=1}^{\infty} y_j z^j$, $|z| < 1$, and the average provision time is $E(Y)$.

(5) A service station vacation is continue, if the system is still no customers waiting for service, or the number of customer is less than N_0.

(6) If the number of the customer in the system is more than N_0, then the service station vacation is end. In particular, when the first customer of the system arrives, the service station starts the service immediately.

(7) The arrival time, the service time and the provision time are independent.
So we can analyze the quality of service by a discrete time $Geo/G/1$ Queuing service system with vacation.

The steady distribution of the task number in the cloud computing system is $p_j = \lim_{n\to\infty} P\{N(n) = j\}, j = 0, 1, 2, \cdots$,

(i) when $\rho \geq 1, p_j = 0, j = 0, 1, 2, \cdots$;

(ii) when $\rho < 1$, the recursion expressions of $\{p_j, j = 0, 1, 2, \cdots\}$ for the transient queue length distribution is [14, 15]

$$p_0 = \frac{1-\rho}{1 + (N_0 - 1)Y(\overline{p})},$$

$$p_j = p_0 \left\{ p\theta_j + Y(\overline{p}) + pY(\overline{p}) \sum_{k=1}^{j-1} \theta_k \right\}, j = 1, 2, \cdots, N_0 - 1,$$

$$p_j = pp_0 \left\{ \theta_j + Y(\overline{p}) \sum_{k=1}^{N-1} \theta_{j-N+k} \right\}, j = N_0, N_0 + 1, \cdots$$

where $j \leq 0$, $\sum_{i=1}^{j} = 0$,

$$\theta_1 = \frac{1 - G^*(\overline{p})}{pG^*(\overline{p})},$$

$$\theta_j = \frac{1}{G^*(\overline{p})} \left\{ \sum_{n=j-1}^{\infty} \sum_{k=n+1}^{\infty} g_k \binom{n}{j-1} p^{j-1} \overline{p}^{n-j+1} + \sum_{i=1}^{j-1} \theta_{j-i} \left[1 - G^*(\overline{p}) - \sum_{m=1}^{i} \sum_{k=m}^{\infty} \binom{k}{m} g_k p^m \overline{p}^{k-m} \right] \right\} \quad j = 2, 3, \cdots$$

So this can be related to service performance index of cloud computing.

(1) **The time of the service busy**
The variable b is the server busy period, and the Probability generating function of the variable b is $B^*(z) = \sum_{j=1}^{\infty} P\{b = j\} z^j$, $|z| < 1$ and $B^*(z) = G^*((\overline{p} + pB^*(z))z)$.The expectation of b is

$$E[b] = \begin{cases} \frac{\rho}{p(1-\rho)}, & \rho < 1, \\ \infty, & \rho \geq 1. \end{cases}$$

(2) **Average number of clients waiting for service on the server**
According to the definition of the expectations, Average time of continuous work of cloud computing server is the average queue length $E[N]$ of the queuing system (the number of requests for queuing):

$$E[N] = \rho + \frac{p^2}{2(1-\rho)} E\left[\chi^2 - \chi\right] + \frac{Y(\overline{p})}{1 + (N_0 - 1)Y(\overline{p})} \times \frac{N_0(N_0 - 1)}{2},$$

where $E\left[\chi^2\right] = \sum_{j=1}^{\infty} j^2 g_j$.

(3) **The distribution of requests for queuing number**
The Probability generating function of the number of the queuing task variable N in the cloud computing system is $\pi(z) = \sum_{j=0}^{\infty} p_j z^j$, $|z| < 1$.
When $\rho < 1$, we have

$$\pi(z) = \frac{(1-\rho)(1-z)G^*(\overline{p} + pz)}{G^*(\overline{p} + pz) - z} \times \frac{1 - z + Y(\overline{p})(z - z^N)}{(1-z)\left[1 + (N_0 - 1)Y(\overline{p})\right]}.$$

(4) **The average response time**
The average response time of requests is the sum of the average waiting time and average business hours in queuing system for customers. So the average response time $T = E[W]$, That is,

$$\begin{aligned} T &= E[W] \geq E[N] \times E[\chi] \\ &= \mu\left(\rho + \frac{p^2}{2(1-\rho)} E\left[\chi^2 - \chi\right] + \frac{Y(\bar{p})}{1 + (N_0 - 1)Y(\bar{p})} \times \frac{N_0(N_0 - 1)}{2}\right). \end{aligned}$$

5 Model Analysis and Service Performance Index

Assuming that the system simulation time in minutes, the arrival interval of task obeys the Geometric Distribution with $p = 0.2$, the service completion time obeys the Discrete Distributions with $\mu = 2$. We can analyze service performance index of the cloud computing system.

5.1 Service Performance Index with no Vacation

(1) The immediate service probability is

$$p_0 = 1 - \rho = 1 - p\mu = 0.6.$$

(2) The expectation of the server busy period is

$$E[b] = \frac{\rho}{p(1-\rho)} = 0.3333.$$

(3) The steady distribution of the queuing task p_j is in the Table 1:

Table 1. The stead distribution of the queuing task

	p_0	p_1	p_2	p_3	p_4	p_4
Probability	0.8523	0.1094	0.0268	0.0083	0.0015	0.0006

(4) Average number of clients waiting for service on the server

$$E[N] = 0.1971$$

(5) The average response time is

$$T = E[W] = E[N] \times E[\chi] = 0.3942.$$

5.2 Service Performance Index with Vacation

The time of the provision of vacation is variable Y that is the Geometric Distribution with parameter q. The number of the customer in the system is more than N_0, then the service station vacation is end.

(1) When $p = 0.2, \mu = 2, N_0 = 5$, we discuss the stead distribution of the queuing task at $q = 0.01, or\ q = 0.1,\ or\ q = 0.3$, such as Fig. 1.

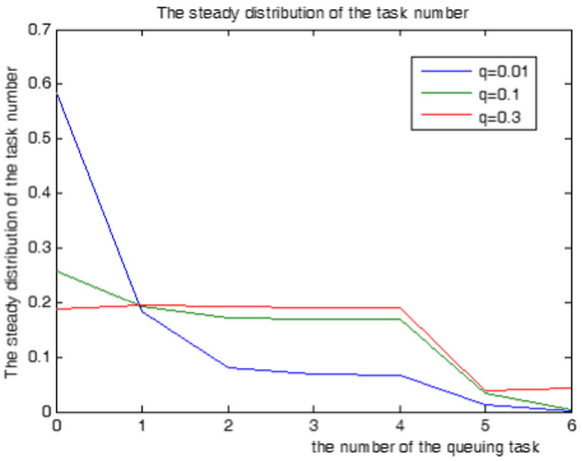

Fig. 1. The stead distribution of the queuing task at $q = 0.01, or\, q = 0.1,\ or\, q = 0.3$.

By the stead distribution, we can calculate the average number of queuing task in the Table 2.

Table 2. The average number of clients waiting for service on the server by q.

	$q = 0.01$	$q = 0.1$	$q = 0.3$
The average number of clients waiting for service on the server	0.89	1.92	2.14

(2) When $p = 0.2, \mu = 2, q = 0.05$, we discuss the stead distribution of the queuing task at $N_0 = 1, or\, N_0 = 3,\ or\, N_0 = 5$, such as Fig. 2.

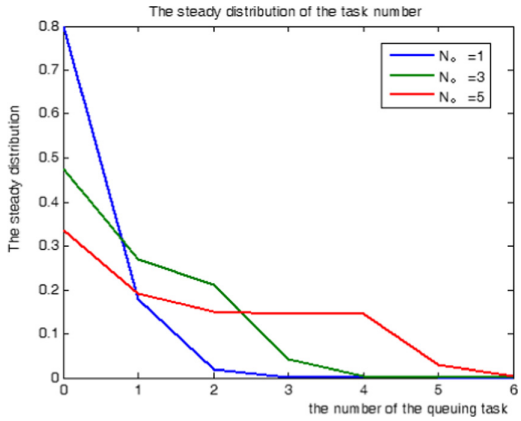

Fig. 2. The stead distribution of the queuing task at $N_0 = 1, or\, N_0 = 3,\ or\, N_0 = 5$.

By the stead distribution, we can calculate the average number of clients waiting for service on the server in the Table 3.

Table 3. The average number of clients waiting for service on the server by N_0.

	$N_0 = 1$	$N_0 = 3$	$N_0 = 5$
Average number of clients waiting for service on the server	0.23	0.84	1.67

(3) When $N_0 = 5, \mu = 2, q = 0.05$, we discuss the stead distribution of the queuing task at $p = 0.01, or\, p = 0.05,\ or\, p = 0.2, or\, p = 0.4$, such as Fig. 3.

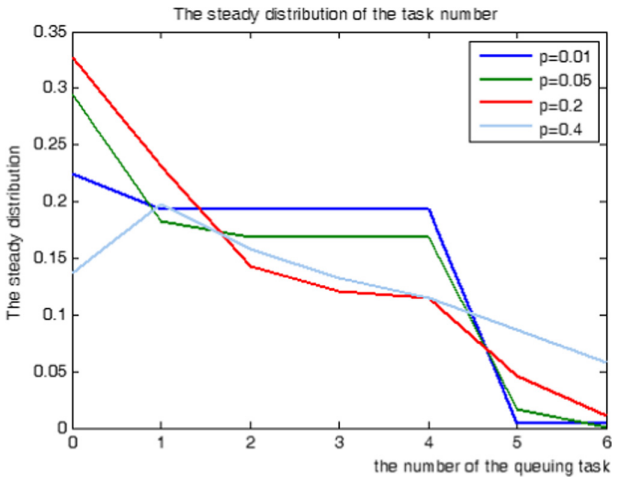

Fig. 3. The stead distribution of the queuing task at $p = 0.01, or\, p = 0.05,\ or\, p = 0.2, or\, p = 0.4$.

By the stead distribution, we can calculate Average number of clients waiting for service on the server in the Table 4.

Table 4. The average number of clients waiting for service on the server by p.

	$p = 0.01$	$p = 0.05$	$p = 0.2$	$p = 0.4$
The average number of clients waiting for service on the server	1.95	1.78	1.67	3.2

5.3 Optimizing Concurrent Number of the Cloud Computing

The concurrent number of the cloud computing is not enough to be more than the mean. For example, when $p = 0.2, \mu = 2, N_0 = 5, q = 0.05$, the stead distribution of the queuing task is in the Table 5.

Table 5. The stead distribution of the queuing task

	p_0	p_1	p_2	p_3	p_4	p_5	p_6	p_7
Probability	0.3362	0.1907	0.1500	0.1455	0.1450	0.0290	0.0032	0.0004

In this system, the average number of queuing task is 1.67. So

$$P\{N > E(N)\} = P\{N > 1.67\} \approx 0.373.$$

That is, the Probability of the concurrent number is not enough which is 37.3% if the concurrent number is the mean of queuing task. For the Probability of the concurrent is enough which will be greater than 99%, we must get the concurrent number which is more than 6 ($p_0 + p_1 + p_2 + \cdots + p_6 > 99\%$).

Acknowledgement. The thesis is supported by Postgraduate Education Reform and Quality Improvement Project of Henan Province (YJS2021AL008). I shall extend my thanks to Yan Liu for all his kindness and help. I'd like to thank my school for providing the experimental environment.

References

1. Buyya, R., Yeo, C.S., Venugopal, S., et al.: Cloud computing and e-merging IT platforms: vision, hype, and reality for delivering computing as the 5th utility. Futur. Gener. Comput. Syst. **25**(6), 599–616 (2009)
2. Zhang, Q., Cheng, L., Boutaba, R.: Cloud computing: state-of-the-art and research challenges. J. Internet Serv. Appl. **1**(1), 7–18 (2010). https://doi.org/10.1007/s13174-010-0007-6
3. Patel, P., Ranabahu, A.H., Sheth, A.P.: Service level agreement in cloud computing [C/OL] (2013). http://knoesis.wright.edu/library/download/OOPSLA_cloud_wsla_v3.pdf
4. Xiong, K., Perros, H.: Service performance and analysis in cloud computing. In: Proceedings of the 2009 World Conference on Services, pp. 693–700. IEEE, Piscataway (2009)
5. Khazaei, H., Misic, J., Misic, V.B.: Performance analysis of cloud computing centers using m/g/m/m+r queueing systems. IEEE Trans. Parallel Distrib. Syst. **23**(5), 936–943 (2012)
6. Yang, B., Tan, F., Dai, Yuan-Shun., Guo, S.: Performance evaluation of cloud service considering fault recovery. In: Jaatun, M.G., Zhao, G., Rong, C. (eds.) CloudCom 2009. LNCS, vol. 5931, pp. 571–576. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-10665-1_54
7. He, H., Fu, Y., Yang, Y., Xiao, T.: Service performance analysis of cloud computing center based on M/M/n/n + r queueing model. J. Comput. Appl. **34**(7), 1843–1847 (2014)
8. Xiong, K.: Web services performance modeling and analysis. In: Proceeding of the 6th International Symposium on High Capacity Optical Networks and Enabling Technologies, Alexandria, Egypt, pp. 1–6 (2009)

9. Wang, H., Huang, M.-H., Long, H.: The performance analysis of web services composition based on queueing network with G/G/1-FCFS, M/G/1-PS and M/G/∞ Nodes. Chin. J. Comput. **36**(1), 22–38 (2014). https://doi.org/10.3724/SP.J.1016.2013.00022
10. Zai, G., Liuyan: Service performance analysis of cloud computing center based on vacation Queueing system. J. Comput. Inf. Syst. **11**(19) 7029–7036 (2015)
11. Tian, N.S.: Stochastic service systems with vacations. Peking University Press, Beijing, pp. 299–315 (2001)
12. Tang, Y.: Queueing Theory: Basic and Analytical Techniques. Science Press, Beijing (2006)
13. Tang, Y.: Queue-length distribution and capacity optimum design for Geo/G/1 queueing system with delayed N-policy and set-up time. Math. Appl. **24**(3), 567–574 (2011)
14. Wei, Y.-Y., Tang, Y.-H., Gu, J.-X.: Queue length distribution and numerical calculation for Geo/G/1 Queueing system with delayed N-policy. Syst. Eng.-Theory Pract. **31**(11), 2151–2160 (2013)
15. Wei, Y.-Y., Tang, Y.-H., Gu, J.-X.: Queue size distribution and capacity optimum design for Geo/G/1 queueing system with delayed D-policy. **33**(4), 996–1005 (2013)