Send Orders for Reprints to reprints@benthamscience.ae

2280 *The Open Automation and Control Systems Journal,* 2015, *7,* 2280-2285

Open Access

The Cloud Computing Center Performance Analysis Model Based on Queuing Theory

Zhang Yong-Hua[1,2,*], Zhou Zhen[2], Zeng Fan-Zi[1] and Li Yuan[2]

[1]*College of Computer Science and Electronic Engineering, Hunan University, Changsha, 410012, China*
[2]*Hunan Industry Polytechnic, Changsha, 410208, China*

Abstract: To make comprehensive and objective analysis of the performance of the cloud computing system center, a cloud computing center system analysis model based on queuing theory is proposed. First, to describe the characteristics of user's service changes request arrived at center by using Poisson distribution, the batch arrival system model was established based on queuing theory, then solve the steady-state probability of length on the probability space, finally to analyze the performance indexes as blocking probability, prompt service probability, *etc*. by using simulation experiment. The simulation results indicate that with increased batch arrival requests, the average length of the cloud computing center increases correspondingly. Increasing the length of the buffer can reduce the blocking probability of system, the experimental results can provide valuable reference for cloud service providers.

Keywords: Batch arrival, cloud computing system, performance analysis, queuing theory, the average waiting length.

1. INTRODUCTION

Cloud computing system is a kind of data processing center that takes Internet as the core. Unified schedule storage, software, services and other resources, constitute a virtual computer center, to provide users with on-demand business model. Due to the increasing species of user demand, the users put forward requests of higher service quality of cloud computing system center. Cloud computing center is the core of the cloud computing system, the advantages and disadvantages of whose performance directly determine the pros and cons of cloud computing service quality, thus make the comprehensive and accurate analysis of the performance of cloud computing system center of great significance [1, 2].

For the analysis of cloud computing system center, domestic and foreign scholars, and experts have invested a lot of time and energy to conduct a wide range of research; a lot of cloud computing center performance models came to the fore. Traditional cloud computing center performance analysis models adopt the simulation system such as the realization of CloudSim , they belong to the kind of static analysis model, assuming that cloud computing system center is operating at static state. However, in practical applications, cloud computing systems center has large scale and complex structure, and with time-dependent nature and mutability, it is difficult to establish accurate analysis model by using traditional model. In order to solve the shortage of the traditional model, this paper proposes center analysis model of cloud computing system based on the data aggregation algorithm, on the basis of considering the higher quality of service and reduced system energy consumption; literature proposes the limited performance of heterogeneous cloud center server of energy consumption optimization model, to solve the unreasonable resource allocation problem of cloud computing center. In recent years, with the continuous development of queuing theory, some scholars introduced it into the performance analysis of cloud computing system center, literature puts forward performance analysis method of cloud computing center based on profit maximization. The cloud computing center optimization problem of multiple servers is seen as a M/M/M queuing problem, thus the optimal optimization scheme is obtained based on profit maximization. Literature presents performance analysis model of cloud computing center based on the queuing system M/M/l, which improves the service response time, and reduces the energy consumption of the center. Literature proposes performance analysis model of cloud computing center based on M/G/M/M + r queuing theory [3-5].

For a more accurate and comprehensive analysis of cloud computing center performance, this paper proposes a batch arrival performance analysis model of cloud computing center based on queuing theory. First, the user service request process is seen as Poisson flow with parameter as λ, and the probability distribution of each batch of the user service requests are known, so as to establish a performance analysis model of cloud computing center, and then presents a method for calculating the operating index of static system, finally conducts a test on effectiveness and superiority of the model through the simulation experiment [6].

*Address correspondence to this author at the College of Computer Science and Electronic Engineering, Hunan University, Changsha, 410012;
Tel: 18986139113; E-mail: zhangyong-hua@126.com

1874-4443/15

2. WORKING MECHANISM AND QUEUING MODEL OF CLOUD COMPUTING CENTER

2.1. Working Mechanism of Cloud Computing Center

In cloud computing system, there are usually a large number of user service requests into the data processing center, as shown in Fig. (**1**) [7-10]. Because there are different types of reception center, once the user service requirements arrive, cloud computing center will provide different types of services according to users' adaptive needs, the price of different services, and the working mechanism of the cloud computing center as shown in Fig. (**1**).

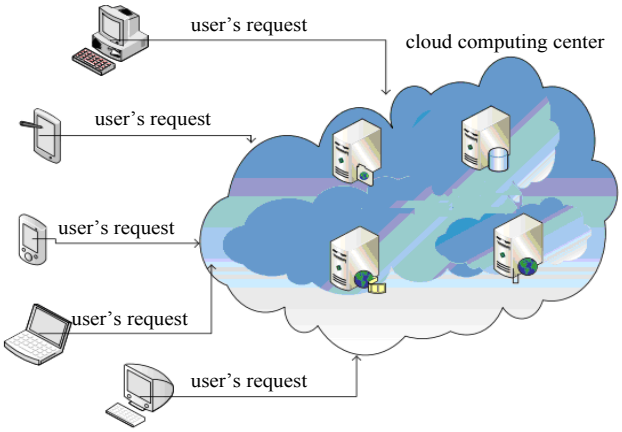


Fig. (1). Working mechanism of the cloud computing center.

2.2. Queuing Model of Cloud Computing Center

Queuing system, an important branch of operational research, is also called stochastic service system, to dispose optimization design problem through the probability of queue. It is successfully applied in the communication system, production management system, *etc*. [11]. Queuing system includes arrival process, queuing rules, service mechanism, for solving performance optimization problem of cloud computing center; specific hypothesis is as follows:

1) The arrival of user service request at the cloud computing system center is random; the arrival time interval between batches of user service requests submits to Poisson flow with parameter λ; all service time required by user service requests is submitted to negative exponential distribution with parameter μ.

2) The number of each batch user service request is a random variable ξ, its probability distribution as: P(ξ=i)=αi , i=1,2,…k.

3) There are m reception desks of cloud computing center (m<k), each reception desk operating independently, in different batch arrival user service request, the service order of which follows the rule of First Come First Served.

4) The capacity of cloud computing system center is limited, that is the system center can contain m user service requests at most, if the capacity of cloud computing system center is full, then the new arrival user service request will leave and search for other server, or it needs to wait in queue.

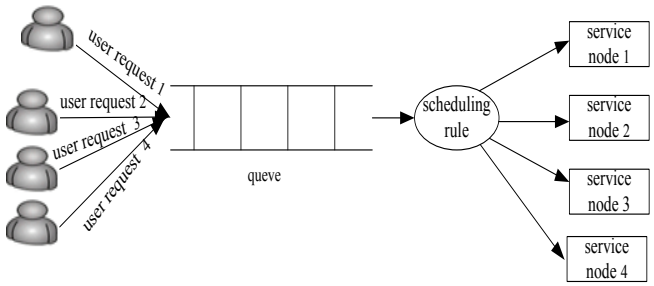


Fig. (2). Queuing model of cloud computing center.

3. PERFORMANCE ANALYSIS MODEL OF CLOUD COMPUTING SYSTEM CENTER

3.1. Queuing Theory Embedded Markov Chain

Markov process refers to the condition of some known random process; the future and past are independent and Markov chain is a time series with discrete time and state as shown in Fig. (**2**). If random process $\{X(t),t\in T\}$ satisfies [12-15]:

1) State space S is countable R.

2) Let $n\geq 1, t_1<t_2<\mathrm{L}<t_n$, if formula (1) sets up, then $\{X(t),t\in T\}$ as Markov chain.

$$P\{X(t_n)<i_n\big|X(t_1)=i_1,\mathrm{L}\ ,X(t_{n-1})=i_{n-1}\}$$
$$=P\{X(t_n)<i_n\big|X(t_{n-1})=i_{n-1}\} \quad (1)$$

Set I as the state space of Markov chain $\{X_n,n\geq 0\}$, then conditional probability is as follows:

$$p_{ij}(m,n)=p\{X_{m+n}=j\big|X_m=i\}$$
$$i,j\in I,m\geq 0,n\geq 0 \quad (2)$$

State space can be divided into several regions according to the specific circumstances of each forecasting object, each region consists of a state, set any state as $E_i\in(E_{1i},E_{2i}]$, $i=1,2\cdots,s$, where s is state number, E_{1i} and E_{2i} as the upper and lower bound of the ith state, respectively, then the state transition probability matrix formula as:

$$P_{ij}^{(n)}=K_{ij}^{(n)}\big/K_i,\quad i,j=1,2\mathrm{L}\ ,s \quad (3)$$

Where, $P_{ij}^{(n)}$ is the probability of E_j *via* n steps transferring, K_i is the frequency of E_i, $K_{ij}^{(n)}$ is the frequency of E_i transferred to E_j *via* n steps.

The calculation formula of state transition probability matrix is obtained as:

$$P^{(n)}=\begin{bmatrix} P_{11}^{(n)} & P_{12}^{(n)} & \text{L} & P_{1s}^{(n)} \\ P_{21}^{(n)} & P_{22}^{(n)} & \text{L} & P_{2s}^{(n)} \\ \text{M} & \text{M} & & \text{M} \\ P_{s1}^{(n)} & P_{s2}^{(n)} & \text{L} & P_{ss}^{(n)} \end{bmatrix} \tag{4}$$

3.2. The Establishment of Performance Analysis Model of Cloud Computing Center

Let X(t) represents the customer service requests number of cloud computing center at time t, according to queuing theory the state space of cloud computing center as E=(0,1,…m), the state transferring process of cloud computing center is shown in Fig. (**3**) [16].

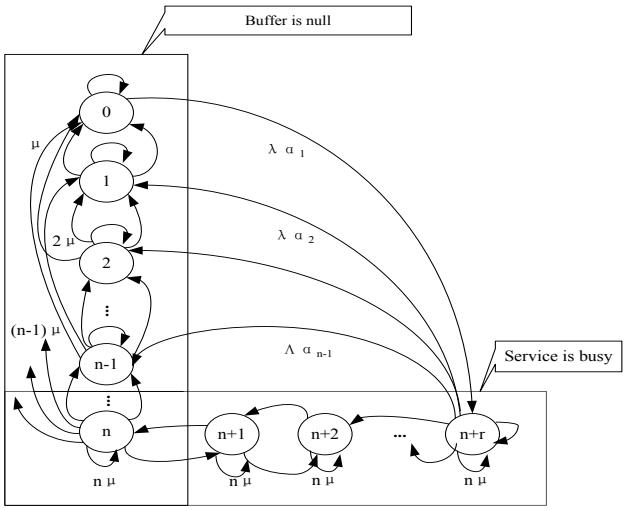


Fig. (3). State transferring process of cloud computing center.

State transition of cloud computing center satisfies the following rules:

1) Anyone state i rightwards can get to its later state as: i+1,i+2,⋯,i+k, and towards to left can get to its adjacent state i-1.

2) Anyone state i towards the left can get to anyone state i-k,⋯,i-2,i-1, while on right state, only state i+1 can get to state i.

3) Take state n as demarcation point, the transfer rate of i(0<i<n) left and right state are iμ and nμ respectively.

Fig. (**3**) shows the limited state Markov process, which has stationary distribution, so the steady state equation set can be obtained based on the Principle of conservation of probability:

$$\begin{aligned} &-\lambda p_0+\mu\cdot p_1=0 \\ &\sum_{j=0}^{i}\left[\lambda\cdot c_{i-j}p_j\right]+p_i\cdot(\lambda+i\mu)+p_{i+1}\cdot(1+i)\mu=0 \quad (0<i<n) \\ &\sum_{j=1}^{i}\left[\lambda\cdot c_j p_{i-j}\right]+p_i\cdot(\lambda+i\mu)+p_{i+1}\cdot(1+i)\mu=0 \quad (n\le i<N) \\ &\lambda\cdot\left[\left(\sum_{j=k}^{k}c_j\right)p_{m-k}+\text{L}+\left(\sum_{j=2}^{k}c_j\right)p_{m-2}+\left(\sum_{j=k}^{k}c_j\right)p_{m-1}\right]-n\mu\cdot p_m=0 \\ &\sum_{i=0}^{m}p_i=1 \end{aligned} \tag{5}$$

Solving on formula (5), the system's steady-state probability distribution can be obtained as p0, p1, …, pm, then the value of length probability pi can be calculated.

4. SIMULATION EXPERIMENT

4.1. Simulation Environment

In order to test the effectiveness and superiority of performance analysis model of the cloud computing system center based on queuing theory, simulation experiment is carried out on a computer with Intel (R) Dual Core (TM) 3.0 GHz CPU, 4 gb of RAM, Windows XP operating system and using MATLAB 2013 tools; the related parameters are shown in Table **1**.

Table 1. Experiment parameter setting.

Parameter Name	Parameter Value
Service node number (m)	100,200,500
Flow intensity (ρ)	0.85
queue buffer length (r)	m/2
covariance (E(x))	[0.5, 10]

4.2. Performance Evaluation Index

To conduct objective and accurate evaluation on the performance of cloud computing center, this paper selects system requests, blocking probability, immediate service probability, and average length as a model performance evaluation indices, which are defined as follows:

$$L_s=\sum_{i=0}^{m}(i\cdot p_i) \tag{6}$$

$$P_b=p_{m+r} \tag{7}$$

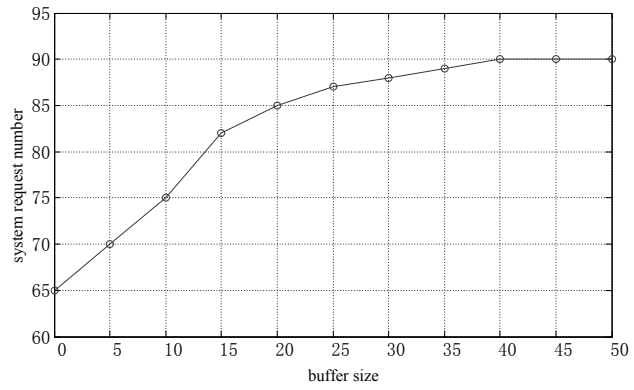
$$P_{iq}=\sum_{i=0}^{m}(p_i) \tag{8}$$

$$L_q=\sum_{i=n}^{m}(i-n)p_i \tag{9}$$

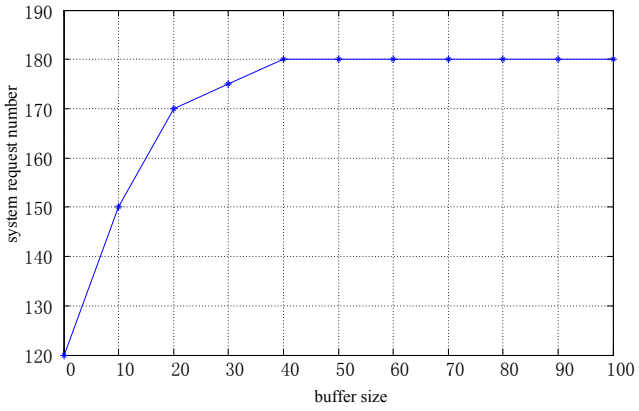
4.3. Result and Analysis

(1) Total System Requests

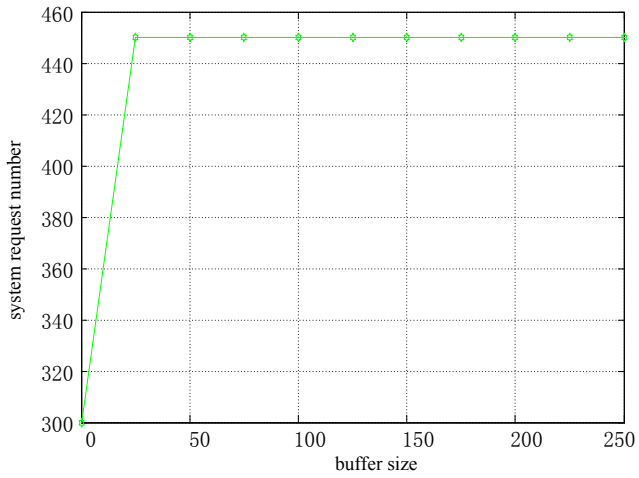
The changing curve of total system requests as shown in Fig. (**4**), shows that with the increase of the input queue buffer length, when service node number m = 100, the growth of the total system requests is quite gentle, when m = 500, then the influence of queue buffer length on the total system requests cannot be observed [17, 18].



(a) m=100



(b) m=200

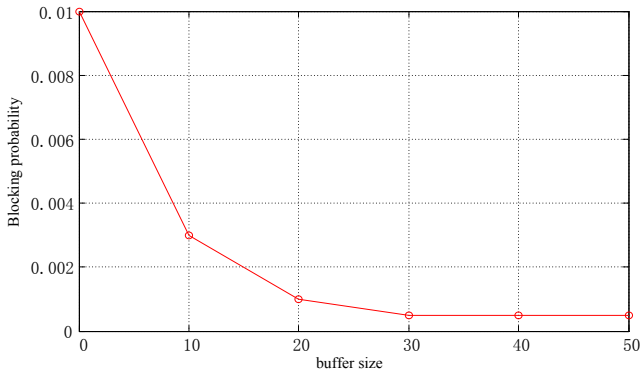


(a) m=500

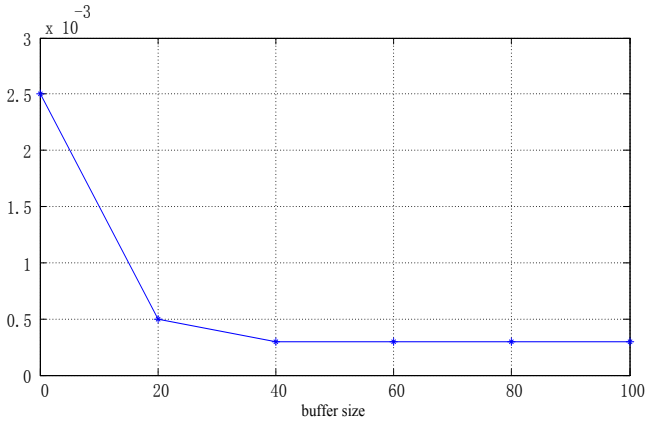
Fig. (4). System Request Number.

(2) Blocking Probability

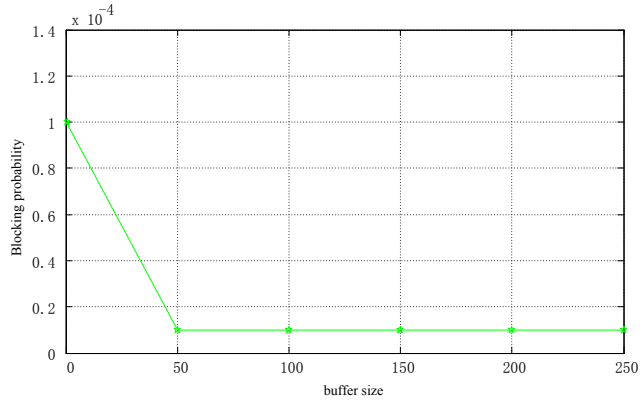
Blocking probability curve is shown in Fig. (**5**). It can be clearly observed from Fig. (**5**) that with the increase of the input queue buffer, and blocking probability is sharply decreased, under the condition of system blocking probability less than 0.2%, the input queue length of buffer should be 10% of this service node number, namely when n = 100, the minimum value of system input queue buffer length should be 10. While increasing the length of the buffer will decline the system blocking probability, it can improve the operation efficiency of the system, but if increased to a certain extent, the system blocking probability is in a stable state.



(a) m=100



(b) m=200

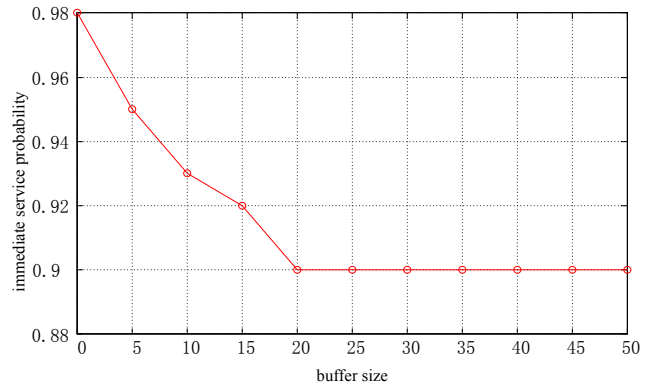


(a) m=500

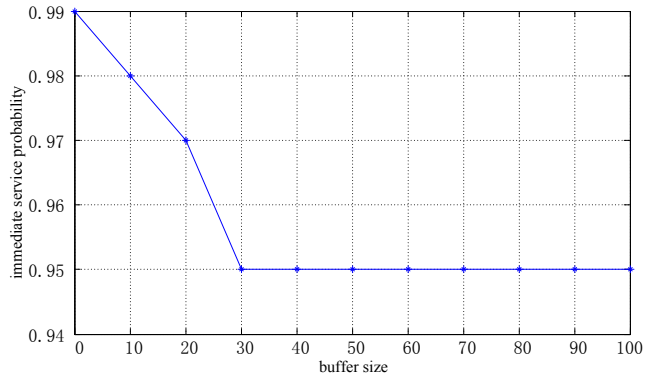
Fig. (5). Blocking Probability.

(3) Immediate Service Probability

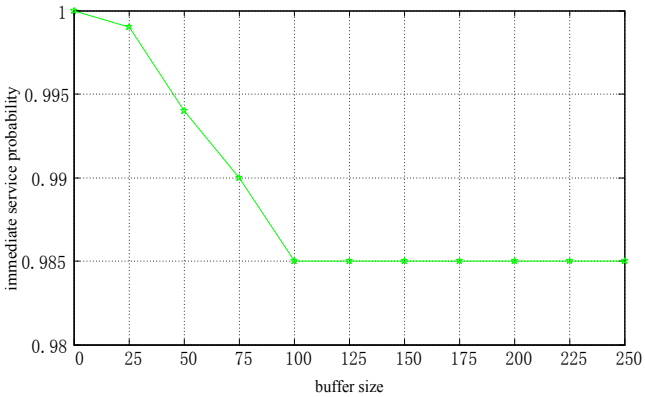
Immediate service probability refers to the probability of user request that can be satisfied without any queuing, which is a key index for evaluating service performance of cloud computing system center. The changing curve of immediate service probability is shown in Fig. (**6**). It can be seen from the Fig. (**6**) that with the increase of the input queue buffer length, the immediate service probability is declined accordingly, which means that when the length of the input queue buffer is smaller, the request arrived at the center of the cloud computing system can obtain immediate service, with no need for queuing. At the same time, it is seen from Fig. (**6c**), the more the service node number, the higher the immediate service probability, the better the system performance is [19].



(**a**) m=100



(**b**) m=200

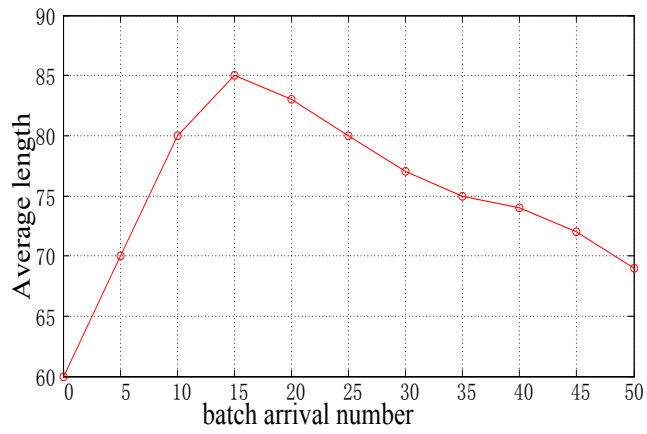


(**c**) m=500

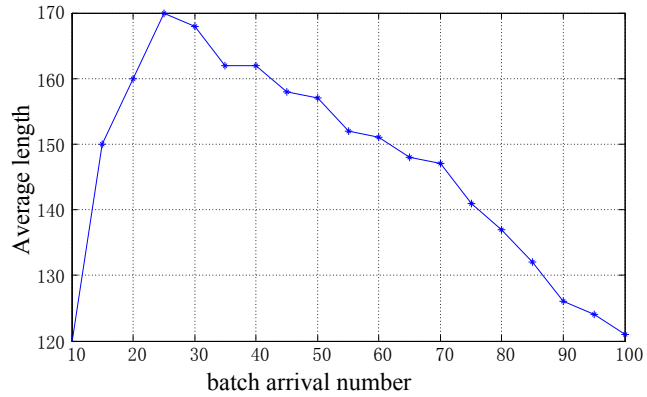
Fig. (6). Immediate service probability.

(4) Average Length Analysis

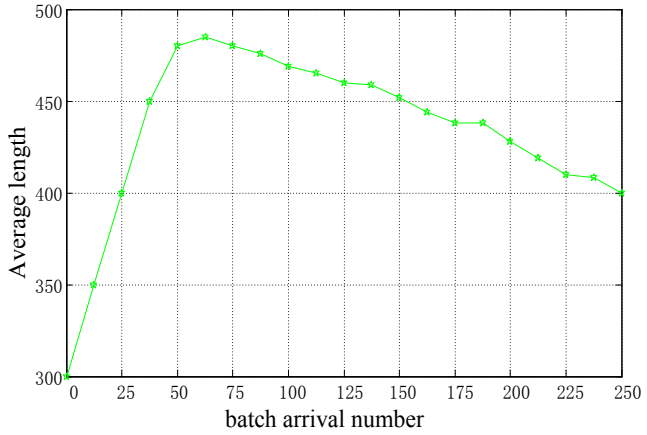
The changing curve of average length is shown in Fig. (**7**). It is known from Fig. (**7**), as the batch service user request number increases, with the continuous increase of average length, up to a certain level, both begin to decline slowly, mainly due to the increase in batch user service requests that occupies large area of queue buffer, and the latter user request service cannot enter the queue buffer and search services. Therefore, the average length is reduced after user requests are satisfied [20].



(**a**) m=100



(**b**) m=200



(**c**) m=500

Fig. (7). Average length.

CONCLUSION

For a more appropriate description of the performance of the cloud computing system center, a performance analysis model of cloud computing center is presented based on queuing theory. The simulation experimental results indicate that the increase of batch arrival request number, the average length of cloud center on this increase, and increasing the buffer length of cloud computing system, can reduce the system blocking probability, and that the more service node number, the higher the immediate service probability, the

results can be the good references for the resource configuration and parameter adjustment of cloud center, which has certain practical application value.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

This work is supported by the Education science department of Hunan provincial research project (12c1032).

REFERENCES

[1] A. Ali-Eldin, J. Tordsson, and E. Elmroth, "An Adaptive hybrid elasicity controller for cloud infrastructures," *IEEE Network Operations and Management Symposium,* 2012, pp. 204-212.

[2] D. Breitgand, and A Epstein, "SLA-aware placement of multi-virtual machine elastic services in compute clouds," In: *IFIP/IEEE International Symposium on Integrated Network Management*, 2011.

[3] J. Fan, "The modified Levenberg-Marquardt method for nonlinear equations with cubic convergence," *Mathematics and Computers*, vol. 81, pp. 447-466, June 2012.

[4] J. Dingde, Z. Xu, P. Zhang, and T. Zhu, "A transform domain-based anomaly detection approach to network-wide traffic," *Journal of Network and Computer Applications ,* vol. 40, no. 5, pp. 292-306, 2014.

[5] J. He, Y. Geng and K. Pahlavan, "Toward accurate human tracking: modelling time-of-arrival for wireless wearable sensors in multipath environment," *IEEE Sensor Journal*, vol. 14, no. 11, pp. 3996-4006, 2014.

[6] J. Wang, L. Zhihan, Z. Xiaolei, F. Jingbao, and C. Ge, "3D graphic engine research based on flash," *Henan Science* , vol. 4, no. 2, pp. 1-15, 2010.

[7] M. Ruina, Z. Lv , Y. Han, and G. Chen, "Research and implementation of geocoding searching and lambert projection transformation based on WebGIS," *Geospatial Information,* vol. 5, no. 13, pp. 101-112, 2009.

[8] S. Li, Y. Geng, J. He, and K. Pahlavan, "Analysis of three-dimensional maximum likelihood algorithm for capsule endoscopy localization*",* In: *International Conference on Biomedical Engineering and Informatics (BMEI)*, 2012, pp. 721-725.

[9] P.E. Gatta, and C. Fierro, "A spatially variant white-patch and gray-world method for color image enhancement driven by local contrast," In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.30, no. 10, pp.1757-1770, 2008.

[10] S. Tianyun, Z. Lv, S. Gao, X. Li, and H. Lv, " 3D seabed: 3D modeling and visualization platform for the seabed, "*IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pp. 1-6, 2014.

[11] T. Alex. B. Lurent, M. Piuzzi, Z. Lu, M. Chavent, M. Baaen, and O. Delalande, "Advances in human-protein interaction-interactive and immersive molecular simulations," In: W. Cai and H. Hong (Eds.) *Protein-Protein Interactions - Computational and Experimental Tools*, 2012, pp. 27-65.

[12] Y. Geng, J. He, H. Deng, and K. Pahlavan, "Modeling the effect of human body on toa ranging for indoor human tracking with wrist mounted sensor, In: *16th International Symposium on Wireless Personal Multimedia Communications (WPMC),* 2013.

[13] Y. Geng, and K. Pahlavan, "On the accuracy of rf and image processing based hybrid localization for wireless capsule endoscopy," In: *IEEE Wireless Communications and Networking Conference (WCNC)*, 2015.

[14] Z. Mengxin, Z. Lv, X. Zhang, G. Chen, and K. Zhang. "Research and application of the 3D virtual community based on WEBVR and RIA," *Computer and Information Science*, vol. 2, no. 1, pp. 84-88, 2009.

[15] Z. Chen, S. M. Arisona, X. Huang, M. Batty, and G. Schmitt, "Detecting the dynamics of urban structure through spatial network analysis," *International Journal of Geographical Information Science* , vol. 28, no. 11, pp. 2178-2199, 2014.

[16] S. Zhou, G. Aggarwal, and R. Chellapa, "Appearance characterization of linear lamebrain object, Generalized photometric stereo and illumination- Invariant face recognition,"*IEEE Transactions on PAMI*, vol. 29, no. 2, pp. 230-245, 2007.

[17] Z. Lv, A. Halawani, S. Feng, S. Rehman, and H. Li, "Touch-less interactive augmented reality game on vision based wearable device," *Personal and Ubiquitous Computing*, vol. 2, no. 1, pp. 1-15 2015.

[18] S. Dang, "Efficient solar power heating system based on lenticular condensation," In: *International Conference on Information Science, Electronics and Electrical Engineering (ISEEE*), 2014, pp. 26-28.

[19] Z. Lv, L. Feng, S. Feng, and H. Li, "Extending touch-less interaction on vision based wearable device," *IEEE Virtual Reality*, 2015, [Available at: http://arxiv.org/ftp/arxiv/papers/1504/1504.0102-5.pdf]

[20] Y. Geng, J. He, H. Deng and K. Pahlavan, "Modeling the effect of human body on toa ranging for indoor human tracking with wrist mounted sensor," In: *International Symposium on Wireless Personal Multimedia Communications (WPMC)*, 2013, pp. 1-6.

Received: July 10, 2015 Revised: August 22, 2015 Accepted: September 11, 2015