The $M/M/2$ Queue System with Flexible Service Policy

You Lyu, Shengli Lv, *Member, IAENG* and Xiaochen Sun

***Abstract*—This paper studies the $M/M/2$ queue system with flexible service policy. When the system has at least two customers in it, every server service for one customer separately. On the other hand, when the system has only one customer in it, the two servers service for the one customer collectively at the same time. For such a queue system, we derive the steady-state probabilities of system states, the steady-state queue length, waiting time and sojourn time of an arbitrary customer. Numerical experiments have been done to show the system performances which are different from the classical $M/M/2$ queue system.**

***Index Terms*—queue system, Markov process, flexible service policy, queue length, waiting time, sojourn time.**

I. Introduction

THE stochastic service system theory, which is also called queuing theory, is a kind of efficient implement for analyzing and calculating the manufacturing system, it is an important aspect of operational research. Many researchers have studied optimization problems of queue systems, and their optimal designs and strategies focus on the behavior of customers. For instance, Burnetas and E-conomou [1] analyzed the customers equilibrium strategies in several Markovian queues. Economou and Kanta [2] studied the equilibrium balking strategies in the M/M/1 queueing system with an unreliable server. Guo and Hassin [3] and Sun et al. [4] have introduced server vacation policies to economic analysis models. Li et al. [5], [6] deal with the balking behavior of customers in the economic analysis of the $M/M/1$ queue with breakdowns. Recently, Li and Li [7] considered an M/M/1 retrial queue with working vacation, orbit search and balking. They obtained the necessary and sufficient condition for system to be stable, the stationary probability distribution and some performance measures.

For economic or optimization objective, we should not only pay attention to the behavior of customers, but also consider the service policy of the servers. Lan and Tang [8] studied the departure process and the optimal control strategy for a discrete-time $Geo/G/1$ queueing model, where the system operates under the control of multiple server vacations and Min(N, V)-policy.

In the past decades, the analysis of multi-server queueing systems has received considerable attention in view of their applicability in manufacturing industries, communication networks and supply-chain systems [9], [10], [11], [12]. The different service policies bring different performances in multi-server queueing systems. Many research works of multi-server queueing systems defined the service policy as one-to-one (OTO) policy. OTO policy is that one server can service for only one customer at the same time, and one customer can be serviced by only one server at the same time yet. This OTO policy causes some servers be in idle when the number of customers is less than the number of the servers. However, the facts suggest otherwise, many cases indicate that in order to enhance the economic efficiency, the many-to-one (MTO) service policy is introduced to the multi-server queueing systems. MTO service policy is that one server can service for only one customer at the same time, but one customer can be serviced by several servers at the same time. Actually, MTO service policy is common in practical systems. For examples, a large transport logistics centre has many forklifts, a truck can be serviced by two or more forklifts at the same time. In addition, multi-core processor technology has been used in computer technology widely[13], two or more computing cores may work together on one task at the same time. These examples, all indicate MTO service policy is used in many service systems, but till this date it is still difficult to indicate any studies that deal with MTO service policy in a multi-server queue system study.

Manuscript received April 22, 2019; revised January 15, 2020.

You Lyu is with the School of Mathematics, Tianjin University, Tianjin 300350, PR China.

Shengli Lv (corresponding author) is with the School of Science, Yanshan University, Qinhuangdao, Hebei 066004, PR China. e-mail:qhdddlsl@163.com.

Xiaochen Sun is with the School of Mathematics, Tianjin University, Tianjin 300350, PR China.

II. Model description

We consider the $M/M/2$ queueing system with an infinite waiting room where customers arrive according to a Poisson process with intensity λ. The system has two identical servers, one server can service for only one customer at the same time, but one customer can be serviced by two servers together at the same time. This service policy is called as two-to-one(TTO) service policy. When the system has at least two customers, each server services for one customer respectively. Otherwise, when the system has only one customer in the system, the two servers service for the one customer together at the same time. If another customer arrive before the completion of the service for the single customer, one of the two servers turns to service for the coming customer immediately. When a customer is serviced by one server, the service time is exponential distribution with the parameter of μ. On the other hand, when a customer is serviced by two servers together, the service time density function is

$$f(x) = \begin{cases} 2\mu q e^{-2\mu q t}, & t \geq 0, \\ 0, & \text{other}, \end{cases}$$

where λ, μ and q are constants, they are all greater than zero. Generally, it is faster for two servers to service one

customer compared to a single server servicing one customer. Though it is faster with two servers, sometimes two servers may produce interaction effects when they work together, so we introduce the interactional parameter q.

Let $X(t)$ be the number of customers in system at time t, then $\{X(t), t \geq 0\}$ is a random process with state space

$$\Omega = \{i, i \geq 0\},$$

According to the birth-and-death process theory [14], the steady-state condition is

$$\rho = \frac{\lambda}{2\mu} < 1.$$

Under the steady-state condition, the steady-state probabilities can be calculated. We let

$$p_i = \begin{cases} \lim\limits_{t\to\infty} P\{X(t) = i\}, & i = 0, 1, 2, \cdots, \\ 0, & \text{others}, \end{cases}$$

then

$$\sum_{i=0}^{\infty} p_i = 1.$$

The system state transfer rate matrix is as follows[14]:

$$Q = \begin{bmatrix} -\lambda & \lambda & & & \\ 2\mu q & -\lambda - 2\mu q & \lambda & & \\ & 2\mu & -\lambda - 2\mu & \lambda & \\ & & \ddots & \ddots & \ddots \end{bmatrix},$$

then the balance equations of the queue system are as follows:

$$\begin{cases} \lambda p_0 - 2\mu q p_1 = 0, \\ \lambda p_0 - (\lambda + 2\mu q)p_1 + 2\mu p_2 = 0, \\ \lambda p_i - (\lambda + 2\mu)p_{i+1} + 2\mu p_{i+2} = 0, \quad (i \geq 1). \end{cases} \tag{1}$$

Solving Eq. (1) yields

$$p_0 = \frac{2\mu q - \lambda q}{2\mu q - \lambda q + \lambda}, p_i = (\frac{\lambda}{2\mu})^i \frac{2\mu - \lambda}{(2\mu - \lambda)q + \lambda}, (i \geq 1). \tag{2}$$

So p_0 is the probability of the system is empty. The probability of a customer need not to wait denoted by P_{NW} is

$$P_{NW} = p_0 + p_1 = \frac{4\mu^2 q - 2\mu\lambda q + 2\mu\lambda - \lambda^2}{4\mu^2 q - 2\mu\lambda q + 2\mu\lambda}.$$

III. Steady-state queue length

Supposing $\rho < 1$, and letting N_W be the queue length of waiting customers, we have

$$P\{N_W = 0\} = \sum_{i=0}^{2} p_i, P\{N_W = k\} = p_{2+k}, k = 1, 2, \cdots.$$

Letting $\overline{N}_W$ be the steady-state queue length of waiting customers, we have

$$\begin{aligned} \overline{N}_W = E[N_W] &= \sum_{i=3}^{\infty}(i-2)p_i \\ &= \sum_{i=3}^{\infty}(i-2)(\frac{\lambda}{2\mu})^i \frac{2\mu - \lambda}{(2\mu - \lambda)q + \lambda} \\ &= \frac{2\mu - \lambda}{(2\mu - \lambda)q + \lambda}(\frac{\lambda}{2\mu})^3 \sum_{i=3}^{\infty}(i-2)(\frac{\lambda}{2\mu})^{i-3} \\ &= \frac{2\mu - \lambda}{(2\mu - \lambda)q + \lambda}(\frac{\lambda}{2\mu})^3 (\sum_{i=1}^{\infty} x^i)'|_{x=\frac{\lambda}{2\mu}}. \\ &= \frac{\lambda^3}{[(2\mu - \lambda)q + \lambda]2\mu(2\mu - \lambda)}. \end{aligned} \tag{3}$$

In steady state, letting N_S be the number of customers being served. Then, we have

$$P\{N_S = k\} = p_k, \quad k = 0, 1, \quad P\{N_S = 2\} = \sum_{i=2}^{\infty} p_i.$$

In steady state, letting $\overline{N}_S$ be the steady-state number of customers being served. Then, we have

$$\begin{aligned} \overline{N}_S = E[N_S] &= p_1 + 2\sum_{i=2}^{\infty} p_i \\ &= \frac{\lambda}{2\mu}\frac{2\mu - \lambda}{(2\mu - \lambda)q + \lambda} + 2\sum_{i=2}^{\infty}(\frac{\lambda}{2\mu})^i \frac{2\mu - \lambda}{(2\mu - \lambda)q + \lambda} \\ &= \frac{2\mu\lambda + \lambda^2}{[(2\mu - \lambda)q + \lambda]2\mu}, \end{aligned} \tag{4}$$

further, the steady-state queue length $\overline{N}$ is

$$\overline{N} = \overline{N}_W + \overline{N}_S = \frac{2\mu\lambda}{[(2\mu - \lambda)q + \lambda](2\mu - \lambda)}. \tag{5}$$

Remark 1. Letting $q = \frac{1}{2}$, Eq. (5) reduces to

$$\overline{N} = \frac{4\mu\lambda}{4\mu^2 - \lambda^2}, \tag{6}$$

and Eq. (6) is the steady-state queue length of the classical $M/M/2$ queue system [14].

IV. Waiting time and sojourn time

A. Waiting time

In this section, we derive the waiting time of an arbitrary customer. In steady state, we denote the total waiting time of an arbitrary customer by W_q with distribution function $W_q(t)$.

Theorem 1. If $\rho < 1$, the waiting time distribution function of an arbitrary customer is

$$\begin{aligned} W_q(t) &= P\{W_q \leq t\} \\ &= 1 - \frac{\lambda^2}{(2\mu - \lambda)2\mu q + 2\mu\lambda} e^{-(2\mu - \lambda)t}, t \geq 0, \end{aligned} \tag{7}$$

and the steady-state waiting time is

$$\overline{W}_q = \frac{\lambda^2}{[(2\mu - \lambda)q + \lambda]2\mu(2\mu - \lambda)}.$$

Proof 1) For $t = 0$, the customer need not to wait, so the number of the customer in the system is less than two, then we have

$$W_q(0) = P\{W_q = 0\} = p_0 + p_1 = \frac{4\mu^2 q - 2\mu\lambda q + 2\mu\lambda - \lambda^2}{(2\mu - \lambda)2\mu q + 2\mu\lambda}.$$

2) For $t > 0$, the customer must to wait, so the number of the customer in the system is not less than two. Using Eq. (2) we have

$$W_q(t) = P\{W_q = 0\} + P\{0 < W_q \le t\} = W_q(0) + \sum_{i=2}^{\infty} P\{0 < W_q \le t | N = i\} p_i$$

where N is the number of customers of an new coming customer sees before him. If there are $i(> 1)$ customers before an new coming customer, the waiting time of this new coming customer is the sum of $(i-1)$ customers' successive departure time intervals, and the density function is Erlang-$(i-1)$ distributed with the mean of 2μ. Then, we have

$$P\{0 < W_q \le t | N = i\} = \int_0^t \frac{2\mu(2\mu x)^{i-2}}{(i-2)!} e^{-2\mu x} dx.$$

Therefore

$$\begin{aligned} W_q(t) &= W_q(0) + \sum_{i=2}^{\infty} (\frac{\lambda}{2\mu})^i \frac{2\mu - \lambda}{(2\mu - \lambda)q + \lambda} \cdot \int_0^t \frac{2\mu(2\mu x)^{i-2}}{(i-2)!} e^{-2\mu x} dx \\ &= \frac{4\mu^2 q - 2\mu\lambda q + 2\mu\lambda - \lambda^2}{(2\mu - \lambda)2\mu q + 2\mu\lambda} \\ &+ \frac{2\mu - \lambda}{(2\mu - \lambda)q + \lambda} (\frac{\lambda}{2\mu})^2 \sum_{i=2}^{\infty} \int_0^t \frac{2\mu(\lambda x)^{i-2}}{(i-2)!} e^{-2\mu x} dx \\ &= \frac{4\mu^2 q - 2\mu\lambda q + 2\mu\lambda - \lambda^2}{(2\mu - \lambda)2\mu q + 2\mu\lambda} \\ &+ \frac{2\mu - \lambda}{(2\mu - \lambda)q + \lambda} (\frac{\lambda}{2\mu})^2 \int_0^t 2\mu e^{(\lambda - 2\mu)x} dx \\ &= \frac{4\mu^2 q - 2\mu\lambda q + 2\mu\lambda - \lambda^2}{(2\mu - \lambda)2\mu q + 2\mu\lambda} \\ &+ \frac{\lambda^2}{(2\mu - \lambda)2\mu q + 2\mu\lambda} (1 - e^{-(2\mu - \lambda)t}) \\ &= 1 - \frac{\lambda^2}{(2\mu - \lambda)2\mu q + 2\mu\lambda} e^{-(2\mu - \lambda)t}. \end{aligned} \tag{8}$$

Using Eq. (8), we obtain the steady-state waiting time as follows:

$$\begin{aligned} \overline{W}_q &= \int_0^{\infty} t dW_q(t) \\ &= 0 \cdot \frac{4\mu^2 q - 2\mu\lambda q + 2\mu\lambda - \lambda^2}{(2\mu - \lambda)2\mu q + 2\mu\lambda} \\ &+ \int_{0^+}^{\infty} \frac{t\lambda^2(2\mu - \lambda)}{(2\mu - \lambda)2\mu q + 2\mu\lambda} e^{-(2\mu - \lambda)t} dt \\ &= \frac{\lambda^2}{[(2\mu - \lambda)q + \lambda]2\mu(2\mu - \lambda)}, \end{aligned} \tag{9}$$

comparing Eq. (9) with Eq. (3), Little's formula [14] holds.

Remark 2. Letting $q = \frac{1}{2}$, Eq. (9) reduces to

$$\overline{W}_q = \frac{\lambda^2}{\mu(4\mu^2 - \lambda^2)}, \tag{10}$$

and Eq. (10) is the steady-state waiting time of an arbitrary customer in the classical $M/M/2$ queue system [14].

B. Sojourn time

We denote the sojourn time of a customer by W with distribution function $W(t)$, and denote the service time of a customer by X with distribution function $X(t)$, so $W(t)$ is the sum of $W_q(t)$ and $X(t)$, and $W_q(t)$ is independent of $X(t)$.

According to the model assumptions, if the system is empty after a customer leaves, the service for the leaving customer is completed by two servers collectively with probability of p_0. On the other hand, if the system is not empty after a customer leaves, the service for the leaving customer is completed by one server with probability of $1 - p_0$. Thus, the service time distribution function is

$$\begin{aligned} X(t) &= P\{X \le t\} \\ &= (1 - e^{-\mu t})(1 - p_0) + (1 - e^{-2\mu q t}) p_0 \\ &= (1 - e^{-\mu t}) \frac{\lambda}{2\mu q - \lambda q + \lambda} \\ &+ (1 - e^{-2\mu q t}) \frac{2\mu q - \lambda q}{2\mu q - \lambda q + \lambda}, t \ge 0, \end{aligned} \tag{11}$$

then the steady-state service time is

$$\begin{aligned} E(X) &= \int_0^{\infty} t dX(t) \\ &= \frac{1}{\mu} \frac{\lambda}{2\mu q - \lambda q + \lambda} + \frac{1}{2\mu q} \frac{2\mu q - \lambda q}{2\mu q - \lambda q + \lambda} \\ &= \frac{2\mu + \lambda}{[(2\mu - \lambda)q + \lambda]2\mu}, \end{aligned} \tag{12}$$

comparing Eq. (12) with Eq. (4), Little's formula holds.

Finally, the steady-state sojourn time denoted by $\overline{W}$ is

$$\begin{aligned} \overline{W} &= \overline{W}_q + E(X) \\ &= \frac{\lambda^2}{[(2\mu - \lambda)q + \lambda]2\mu(2\mu - \lambda)} \\ &+ \frac{2\mu + \lambda}{[(2\mu - \lambda)q + \lambda]2\mu} \\ &= \frac{2\mu}{[(2\mu - \lambda)q + \lambda](2\mu - \lambda)}, \end{aligned} \tag{13}$$

comparing Eq. (13) with Eq. (5), Little's formula holds.

Remark 3. Letting $q = \frac{1}{2}$, Eq. (13) reduces to

$$\overline{W} = \frac{4\mu}{4\mu^2 - \lambda^2}, \tag{14}$$

and Eq. (14) is the steady-state sojourn time of an arbitrary customer in the classical $M/M/2$ queue system [14].

C. Distribution function of sojourn time

Since $W = W_q + X$, and W_q is independent of X, then we can use the convolution formula of distribution function

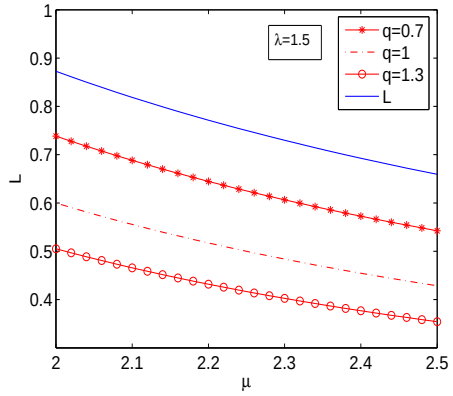

Figure 1. The steady-state queue length versus μ .

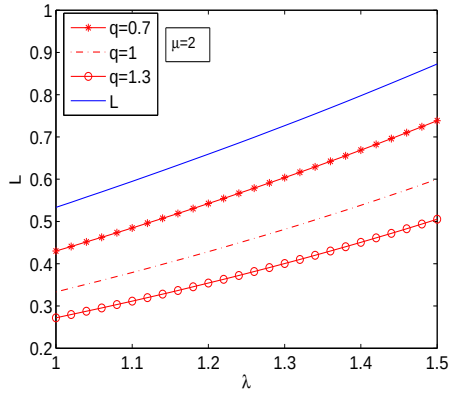

Figure 2. The steady-state queue length versus λ .

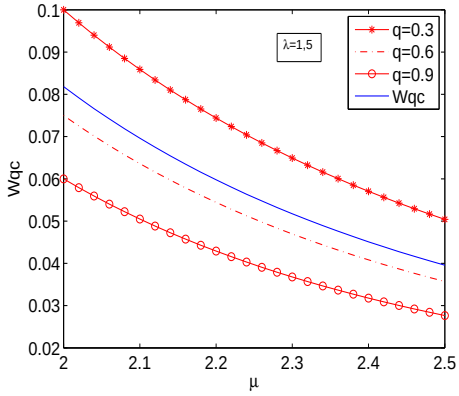

Figure 3. The steady-state waiting time versus μ .

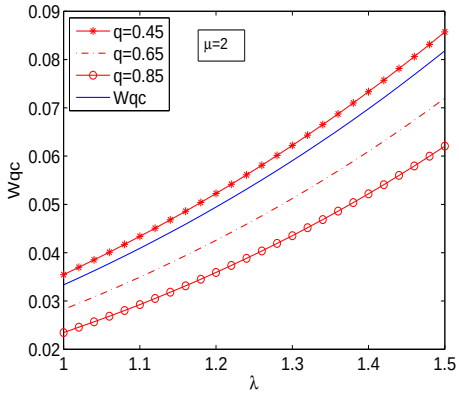

Figure 4. The steady-state waiting time versus λ .

to derive the distribution function of sojourn time as follows:

$$
\begin{aligned}
W(t) &= \int_0^t W_q(t-x)dX(x) \\
&= \int_0^t [1 - \tfrac{\lambda^2}{(2\mu-\lambda)2\mu q+2\mu\lambda} e^{-(2\mu-\lambda)(t-x)}] \\
&\quad d[(1-e^{-\mu x})\tfrac{\lambda}{2\mu q-\lambda q+\lambda} + (1-e^{-2\mu q x})\tfrac{2\mu q-\lambda q}{2\mu q-\lambda q+\lambda}] \\
&= \begin{cases}
\frac{1}{1+q}(1-e^{-\mu t} - \frac{1}{2+2q}e^{-\mu t}\mu t) \\
\quad + \frac{q}{1+q}[1 + \frac{q}{(1+q)(1-2q)}e^{-\mu t} \\
\quad - \frac{1-2q^2}{(1+q)(1-2q)}e^{-2\mu q t})], \quad (\mu = \lambda,\ t \geq 0), \\
\frac{1-q}{q^2+1-q}[1 - e^{-\mu t} - \frac{(1-q)^2}{(q^2+1-q)(2q-1)} \\
\quad \cdot (e^{-\mu t} - e^{-2\mu q t})] \\
\quad + \frac{q^2}{q^2+1-q}[1 - e^{-2\mu q t}(1 + \frac{(1-q)^2}{q^2+1-q}2\mu q t)], \\
\quad (1-q = \frac{\lambda}{2\mu},\ t \geq 0), \\
\frac{\lambda}{2\mu q-\lambda q+\lambda}[1 - e^{-\mu t} \\
\quad - \frac{\lambda^2}{[(2\mu-\lambda)2q+2\lambda](\mu-\lambda)}e^{-\mu t}(1-e^{-(\mu-\lambda)t})] \\
\quad + \frac{2\mu q-\lambda q}{2\mu q-\lambda q+\lambda}[1 - e^{-2\mu q t} - \frac{\lambda^2 q}{[(2\mu-\lambda)q+\lambda][2\mu(1-q)-\lambda]} \\
\quad \cdot e^{-2\mu q t}(1 - e^{-[2\mu(1-q)-\lambda]t})], \\
\quad (\mu \neq \lambda,\ \ 1-q \neq \frac{\lambda}{2\mu},\ \ t \geq 0).
\end{cases}
\end{aligned}
\tag{15}
$$

When $\mu = \lambda$, from Eq. (13) or Eq. (15), we obtain the same result as follows:

$$\int_0^\infty t dW(t) = \frac{2}{(1+q)\mu} = \overline{W}.$$

V. Numerical Experiments

In this section, we explore the qualitative behavior of the model, and illustrate it by numerical examples.

Firstly, we use Eq. (5) and Eq. (6) to find the different performances of queueing length between the two models. The steady-state queue length of classical $M/M/2$ system is denoted by L, and

$$L = \frac{4\mu\lambda}{4\mu^2 - \lambda^2}.$$

In the numerical examples, the steady-state queue lengths of the new model $\overline{N}$ are calculated for three different values of $q(= 0.7, 1, 1.3)$ under the stead-state condition. In Figure 1, the steady-state queue lengths are depicted versus μ for given value of $\lambda = 1.5$. As μ increases, the queue lengths always decrease. On the other hand, in Figure 2, the steady-state queue lengths are depicted versus λ for given value of $\mu = 2$. As λ increases, the queue lengths always increase. The curves of L are the highest in Figure 1 and Figure 2, that is due to $q > \frac{1}{2}$. Furthermore, the queue lengths of the new model $\overline{N}$ become smaller as the values of q become larger.

Secondly, we use Eq. (9) and Eq. (10) to find the different performances of the steady-state waiting time between the two models. The steady-state waiting time of classical $M/M/2$ system is denoted by W_{qc}, and

$$W_{qc} = \frac{\lambda^2}{\mu(4\mu^2 - \lambda^2)}.$$

In Figure 3, the steady-state waiting times are depicted versus μ for given value of $\lambda = 1.5$. As μ increases, the steady-state waiting times always decrease. Especially, the curve of $\overline{W}_q$ for $q = 0.3$ is higher than the curve of W_{qc}, it is due to $q < \frac{1}{2}$. In Figure 4, the steady-state waiting times are depicted versus λ for given value of $\mu = 2$. As λ increases, the steady-state waiting times always increase. Similarly, the curve of $\overline{W}_q$ for $q = 0.45$ is higher than the curve of W_{qc}, it is also due to $q < \frac{1}{2}$. In addition, the steady-state waiting times of the model of this paper $\overline{W}_q$ become smaller as the values of q become larger in Figure 3 and Figure 4.

Following, we use Eq. (13) and Eq. (14) to find the different performances of the steady-state sojourn time between the two models. The steady-state sojourn time of classical $M/M/2$ system is denoted by W_c, and

$$W_c = \frac{4\mu}{4\mu^2 - \lambda^2}.$$

In Figure 5, the steady-state sojourn times are depicted versus μ for given value of $\lambda = 1.5$. As μ increases, the steady-state sojourn times always decrease. Especially, the curve of $\overline{W}$ for $q = 0.3$ is higher than the curve of W_c, it is due to $q < \frac{1}{2}$. In Figure 6, the steady-state sojourn times are depicted versus λ for given value of $\mu = 2$. As λ increases, the steady-state sojourn times always increase. Similarly, the curve of $\overline{W}$ for $q = 0.45$ is higher than the curve of W_c, it is also due to $q < \frac{1}{2}$. In addition, the steady-state sojourn times of the model of this paper $\overline{W}$ become smaller as the values of q become larger in Figure 5 and Figure 6.

Thirdly, the probabilities of the system is empty p_0 are depicted versus q for the different values of λ and the different values of μ in Figure 7. As q increases, the probabilities of p_0 invariably increase.

Following, the probabilities of a customer need not wait P_{NW} are depicted versus q for the different values of λ and the different values of μ in Figure 8. As q increases, the probabilities of P_{NW} invariably increase.

Furthermore, Figure 9 shows the joint effect of μ and λ on the steady-state queue length of the model of this paper for $q = 0.8$, Figure 10 shows the joint effect of μ and λ on the steady-state waiting time of the model of this paper for $q = 0.8$, and Figure 11 shows the joint effect of μ and λ on the steady-state sojourn time of the model of this paper for $q = 0.8$. In addition, the curved surface of Figure 9 coincide with the curves of Figure 1 and Figure 2, the curved surface of Figure 10 coincide with the curves of Figure 3 and Figure 4, and the curved surface of Figure 11 coincide with the curves of Figure 5 and Figure 6.

VI. Conclusions

In this paper, we have studied the $M/M/2$ queueing system where the two servers can service for one customer collectively at the same time. We obtained the steady-state results of queue length, waiting time, sojourn time and other important properties. We compared the model of this paper with the classical $M/M/2$ queueing system through numerical examples, and found that the model of this paper has significant advantages than the classical $M/M/2$ queue system ($q > \frac{1}{2}$). The TTO service pattern is common in the practical two-server queue systems, so the model of this paper can be used in practical production systems or service systems directly. Finally, for further research of the model of this paper someone can introduce the MTO service policy to the $M/M/N(N > 2)$ queueing system.

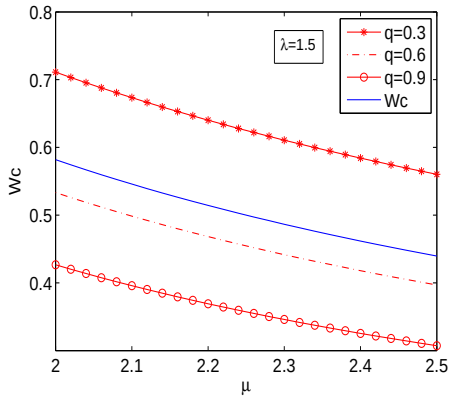

Figure 5. The steady-state sojourn time versus μ .

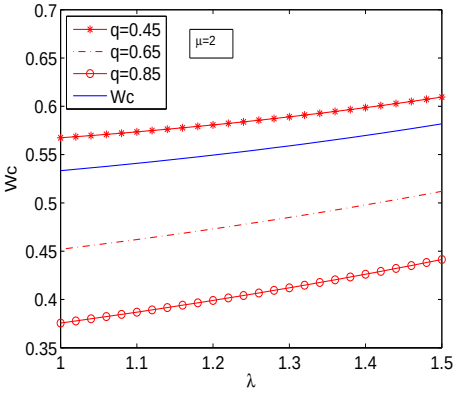

Figure 6. The steady-state sojourn time versus λ .

References

[1] Burnetas A and Economou A, Equilibrium customer strategies in a single server Markovian queue with setup times, *Queueing Systems*, 2007, **56**(3): 213-228.

[2] Economou A and Kanta S, Equilibrium balking strategies in the observable single-server queue with breakdowns and repairs, *Operations Research Letters*,2008, **36**(6): 696-699.

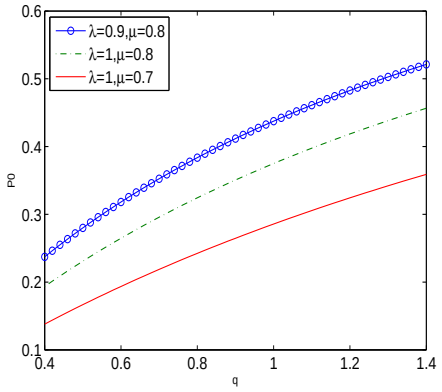

Figure 7. The probability of the system is empty versus q.

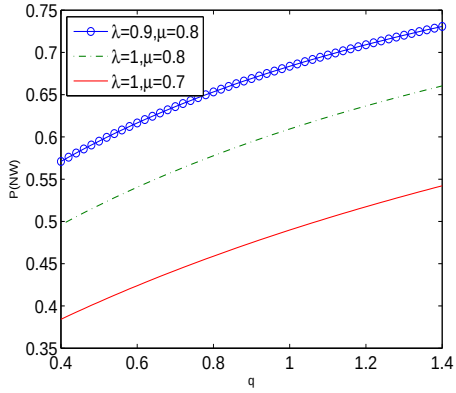

Figure 8. The probability of a customer need not to wait versus q.

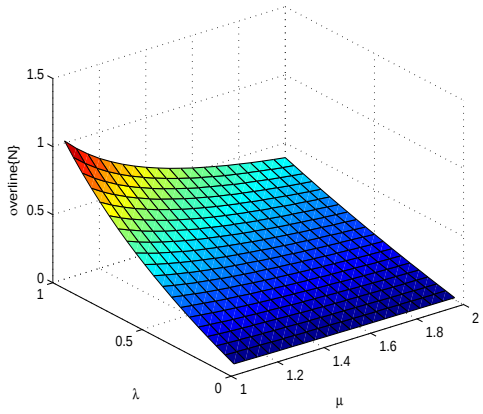

Figure 9. The steady-state queue length for different values of μ and λ (q=0.8).

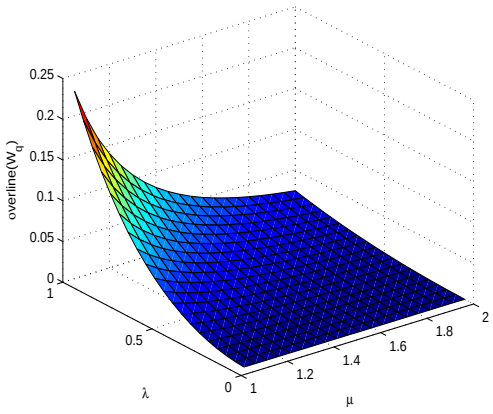

Figure 10. The steady-state waiting time for different values of μ and λ (q=0.8).

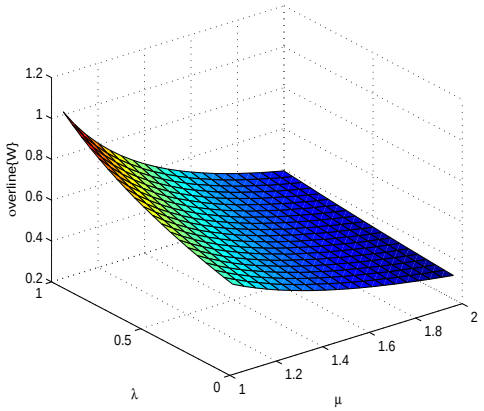

Figure 11. The steady-state sojourn time for different values of μ and λ (q=0.8).

[3] Guo P F and Hassin R, Strategic behavior and social optimization in Markovian vacation queues: The case of heterogeneous customers, *Operations Research Letters*, 2011, **222**(2): 278-286.

[4] Sun W and Li S Y and Li Q L, Equilibrium balking strategies of customers in Markovian queues with two-stage working vacations, *Applied Mathematics & Computation*, 2014, **248**: 195-214.

[5] Li L and Wang J T and Zhang F, Equilibrium customer strategies in Markovian queues with partial breakdowns, *Computers & Industrial Engineering*,2013, **66**(4): 751-757.

[6] Li X Y and Wang J T and Zhang F, New results on equilibrium balking strategies in the single-server queue with breakdowns and repairs, *Applied Mathematics & Computation*, 2014, **241**: 380-388.

[7] Li J T and Li T, An M/M/1 retrial queue with working vacation, orbit search and balking, *Engineering Letters*, 2019, **27**(1): 97-102.

[8] Lan S J and Tang Y H, The structure of departure process and optimal control strategy N^* for $Geo/G/1$ discrete-time queue with multiple server vacations and min(N, V)-policy, *Journal of Systems Science & Complexity*, 2017, **30**:1382-1402.

[9] Krishna K B, Rukmani R and Thangaraj V, On multiserver feedback retrial queue with finite buffer, *Applied Mathematical Modelling*, 2009, **33**(4): 2062-2083.

[10] Do T V, An efficient computation algorithm for a multiserver feedback retrial queue with a large queueing capacity, *Applied Mathematical Modelling*, 2010, **34**(8): 2272-2278.

[11] Baumann H and Sandmann W, Multi-server tandem queue with Markovian arrival process, phase-type service times, and finite buffers, *European Journal of Operational Research*, 2016, **256**(1): 187-195.

[12] Liu Z M and Yu S L, The $M/M/C$ queueing system in a random environment, *Journal of Mathematical Analysis & Applications*, 2016, **436**(1): 556-567.

[13] Nazir A, Accelerated anticor online portfolio selection on multi-core CPUs and GPU with OpenCL, *IAENG International Journal of Computer Science*, 2018, **45**(3): 390-402.

[14] Gross D, Shortle J F, Thompson J M. and Harris C M. Fundamentals of Queueing Theory, Fourth Edition. 2013.