J. Appl. Math. Comput. (2016) 50:117–138
DOI 10.1007/s12190-014-0862-6

ORIGINAL RESEARCH

Threshold-policy analysis of an M/M/1 queue with working vacations

Ji-hong Li · Bao-an Cheng

Received: 17 September 2014 / Published online: 9 January 2015

Abstract In this paper, we consider two M/M/1 queues with working vacations and two policies, m-policy and (m, N)-policy, respectively. The server begins to take the vacation when the number of customers is below m after a service. The server also works in a slow speed in the vacation rather that stoping work completely. We establish a system with two operation periods, higher speed and lower speed periods. First, we study pure m-policy where the server continues another vacation if a vacation is completed and there are less than m customers, otherwise he comes back to regular work. Another (m, N)-policy is the generalization of m-policy where if a vacation is completed and there are less than N customers, the server continues another vacation. Using the quasi birth–death process and matrix-geometric solution method, we give the distributions for the number of customers and some indices of the system, including expected sojourn time and state probabilities of the server. Finally, some numerical examples are presented to verify the validity of the model.

Keywords Working vacation · m-policy · (m, N)-policy · Matrix-geometric solution · Cost function

1 Introduction

In a service system which is composed of the service agents and many potential customers, the queue is very common and it controls and allocates the service ability for customers. When the service ability of the agent is uncomfortable with the number of customers, such as too idle or too tense, adjustment may occur and the service agent

J. Li (✉) · B. Cheng
Research Institute of Management and Decision & College of Economic and Management, Shanxi University, Taiyuan 030006, People's Republic of China
e-mail: lijh1982@sxu.edu.cn

should change its service rate or suggest other service scheme to increase its efficiency of the service, for example, establishing the high–low rate transferring policy, or some gates/thresholds to control the entrance of customers.

Working vacation policy is a high–low rate transferring policy and also a class of semi-vacation policy that was introduced by Servi and Finn [7] in 2002, where a customer is served at a lower rate during a vacation period. Such a vacation is different from the classical vacation queueing models. In the classical vacation queueing models, the server doesn't take the original work and possibly deals with the other tasks. Such policy may make the loss or dissatisfaction of the customers. For the working vacation policy, the server also can work in the lower rate. So, the working vacation is more reasonable and general than the classical vacation.

For working vacation models, Servi and Finn [7] studied an M/M/1 queue with working vacations, and obtained the probability generating function of the queue length and the LST of the waiting time, and applied results to performance analysis of gateway router in fiber communication networks. Subsequently, Kim et al.[3], Wu and Takagi [11] generalized results in [7] to an M/G/1 queue with working vacations. Baba [1] extended this study to a GI/M/1 queue with working vacation by the matrix-analysis method. Liu et al. [8] obtain the concise expressions for the queue length and waiting time for the M/M/1 queue with multiple working vacation and verify the stochastic decomposition structures of the queue length and waiting time. Li and Tian [4,5] obtained the expressions for the queue length and waiting time for two types of GI/Geo/1 queue with working vacations and verified the stochastic decomposition structures of the queue length and waiting time. Tian and Zhang [10], Zhang et al.[12] gave the threshold policy analysis for multi-server queue and M/G/1 queue with general vacations. For general queueing analysis, including the vacation policy, the readers are recommended to Gross and Harris [2] and Tian and Zhang [9].

In recent papers on the working vacations, the authors only concentrate on the vacation queues with exhaustive service and the server only takes the vacation when the system is empty. In this paper, we will consider the M/M/1 queue with the threshold policy and working vacations. Such model is different from the other models with working vacations. The server begins or ends to take the vacation at the certain point, i.e.the threshold. Such policy is also called by the non-exhaustive service. Meanwhile, the server also works in the vacation period at the lower rate rather than stoping working completely. The motivation for studying this kind of models can be presented both in practical aspect and theoretical aspect. Firstly, such model can be seen as the service system with high and low speed periods controlled by thresholds. When the number of customers or signals under one certain threshold in the system, the server can work slowly. Such policy will enable the cost of system to reduce. Such a model also is more practical than the classical threshold queue with vacation where the server can not work during the vacation period. Many practical problems present this character. In banks, when the number of customers or work is under some value, some counters will be closed to do other work. Under this case, the bank needs to consider what will happen based on the performance indices of the bank, such as the queue length or waiting time. So this kind of models has important practical background. In theoretical view, this kind of mod-

els will be more general than those M/M/1 queues considered before and the classical threshold queues with vacations are also the special examples of this kind of models.

The rest of this paper is organized as follows. In Sect. 2, we study the M/M/1 queue with m-policy, where the quasi birth and death process, the distribution of the queue length is presented. Section. 2.3 turns to the M/M/1 queue with (m, N)-policy. In Sect. 4, some numerical examples are presented to verify the validity of the model and the cost function is also established in M/M/1 queue with m-policy. Section. 5 concludes the results.

2 M/M/1 queue with m-policy

2.1 Model formulation and quasi birth and death process (QBD)

Consider a classical M/M/1 queue with an arrival rate λ and service rate μ_b. At the instant of a service completion, the server begins a vacation of random length at the instant when the queue length is below m and vacation duration V follows an exponential distribution with parameter θ. During a vacation, the original customers or arriving customers in a vacation period can be served at a mean rate of μ_v. When a vacation ends, if the number of customers in the queue is less than m, another vacation is taken; Otherwise, the server switches service rate from μ_v to μ_b. This service discipline is a m-threshold policy with working vacation. Evidently, this model is a non-exhaustive service queue and the server can begin to take the vacation when there are customers in the system. Many service systems are the special cases of this model and when $m = 1$, this model becomes the general M/M/1 queue with multiple working vacations which was considered by Servi and Finn [7] and Liu et al. [8].

We assume that inter-arrival times, service times, and working vacation times are mutually independent. In addition, the service discipline is first in first out(FIFO).

Let $Q_1(t)$ be the number of customers in the system at time t and let

$$J_1(t) = \begin{cases} 0, & \text{the system is in a working vacation period at time } t, \\ 1, & \text{the system is in a regular busy period at time } t. \end{cases}$$

then $\{Q_1(t), J_1(t)\}$ is a QBD with the state space

$$\Omega = \Big\{(k, 0) : 0 \le k \le m-1\Big\} \bigcup \Big\{(k, j) : k \ge m, \quad j = 0, 1\Big\}.$$

Evidently, when the number of customers is less than m, the server only stays in vacation period.

Using the lexicographical sequence for the states, the infinitesimal generator can be written as

$$\mathbf{Q_1} = \begin{bmatrix} -\lambda & C_0 & & & & & & \\ B_1 & A_1 & C_1 & & & & & \\ & \ddots & \ddots & \ddots & & & & \\ & & B_{m-1} & A_{m-1} & C_{m-1} & & & \\ & & & B_m & A & C & & \\ & & & & B & A & C & \\ & & & & & \ddots & \ddots & \ddots \end{bmatrix},$$

where

$$B_k = \mu_v, 1 \le k \le m-1, B_m = (\mu_v, \mu_b)^T;$$
$$C_k = \lambda, 0 \le k \le m-2, \quad C_{m-1} = (\lambda, 0); \quad A_k = -(\lambda + \mu_v), 1 \le k \le m-1;$$
$$C = \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}, \quad B = \begin{bmatrix} \mu_v & 0 \\ 0 & \mu_b \end{bmatrix}, \quad A = \begin{bmatrix} -(\lambda + \theta + \mu_v) & \theta \\ 0 & -(\lambda + \mu_b) \end{bmatrix}.$$

To analyze this QBD process, it is necessary to solve for the minimal non-negative solution of the matrix quadratic equation

$$R^2 B + RA + C = 0. \tag{1}$$

and this solution is called the rate matrix and denoted by R. Obviously, we have

Lemma 1 *If $\rho = \lambda(\mu_b)^{-1} < 1$, the matrix equation (1) has the minimal non-negative solution*

$$R = \begin{bmatrix} r & \dfrac{\theta\, r}{\mu_b(1-r)} \\ 0 & \rho \end{bmatrix}, \tag{2}$$

where

$$r = \frac{1}{2\mu_v}\left(\lambda + \theta + \mu_v - \sqrt{(\lambda + \theta + \mu_v)^2 - 4\lambda\mu_v}\right)$$

and $0 < r < 1$.

Proof Because the matrices A, B, C of (1) are all upper triangular, we can assume that R has the same structure as

$$R = \begin{bmatrix} r_{11} & r_{12} \\ 0 & r_{22} \end{bmatrix}.$$

Substituting R^2 and R into (1) gives the following set of equations:

$$\begin{cases} \mu_v r_{11}^2 - (\lambda \ + \theta + \mu_v)\, r_{11} + \lambda = 0, \\ \mu_b r_{22}^2 - (\lambda \ + \mu_b)\, r_{22} + \lambda = 0, \\ \mu_b\, r_{12}(r_{11} + r_{22}) + \theta\, r_{11} - (\lambda + \mu_b) r_{12} = 0. \end{cases} \tag{3}$$

To obtain the minimal non-negative solution of (1), taking $r_{22} = \rho$ (the other root is $r_{22} = 1$) in the second equation and $r_{11} = r$ (the other root is greater than 1) in the first equation of (3). Using the elementary method (discriminant of quadratic equation), we can prove that $0 < r < 1$. Substituting r and ρ into the last equation of (3), we get the expression for r_{12}. Thus, we can have the results in Lemma 1.

Lemma 2 *r satisfies the following relationship*

$$\lambda + \theta + \mu_v(1 - r) = \mu_v + \frac{\theta}{1 - r} = \frac{\lambda}{r}. \tag{4}$$

Proof From the Lemma 1, r satisfies the first equation of (3). Dividing two ends of this equation by r, we give

$$r\mu_v - (\lambda + \theta + \mu_v) + \frac{\lambda}{r} = 0.$$

Hence, we have

$$\lambda + \theta + \mu_v(1 - r) = \frac{\lambda}{r}.$$

Transferring λ from left to right of the above relation, dividing two ends of this equation by $1 - r$, we get the other equation of (4)

$$\mu_v + \frac{\theta}{1 - r} = \frac{\lambda}{r}.$$

Theorem 1 *The QBD process $\{Q_1(t), J_1(t)\}$ is positive recurrent if and only if $\rho < 1$.*

Proof Based on Theorem 3.1.1 of Neuts [6], the QBD process $\{Q_1(t), J_1(t)\}$ is positive recurrent if and only if the spectral radius SP (R) of the rate matrix R is less than 1, and set of equations

$$(x_{00}, x_{10}, \cdots, x_{m-1,0}, x_{m0}, x_{m1})B[R] = 0 \tag{5}$$

has positive solution, where

$$B[R] = \begin{bmatrix} -\lambda & C_0 & & & & & \\ B_1 & A_1 & C_1 & & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & B_{m-1} & A_{m-1} & C_{m-1} & \\ & & & & B_m & RB + A \end{bmatrix},$$

where

$$RB + A = \begin{bmatrix} -\dfrac{\lambda}{r} & \dfrac{\lambda}{r} - \mu_v \\ 0 & -\mu_b \end{bmatrix}.$$

With (4), we can easily verify $B[R]$ is an infinitesimal generator. Substituting $B[R]$ into the above relation, we obtain the set of equations

$$\begin{cases} -\lambda x_{00} + \mu_v x_{10} = 0; \\ \lambda x_{k-1,0} - (\lambda + \mu_v) x_{k0} + \mu_v x_{k+1,0} = 0, \quad 1 \le k \le m-2; \\ \lambda x_{m-2,0} - (\lambda + \mu_v) x_{m-1,0} + \mu_v x_{m,0} + \mu_b x_{m1} = 0, \quad k = m-1; \\ \lambda x_{m-1,0} - \dfrac{\lambda}{r} x_{m0} = 0; \\ \left(\dfrac{\lambda}{r} - \mu_v\right) x_{m0} - \mu_b x_{m1} = 0. \end{cases} \tag{6}$$

From the first and second equations in (6), we easily get

$$x_{k0} = x_{00} \left(\frac{\lambda}{\mu_v}\right)^k, \quad 0 \le k \le m-1.$$

Then, from the other equations and (4),

$$\begin{cases} x_{m0} = r x_{m-1,0} = x_{00} r \left(\dfrac{\lambda}{\mu_v}\right)^{m-1}; \\ x_{m1} = \dfrac{1}{\mu_b}\left(\dfrac{\lambda}{r} - \mu_v\right) x_{m0} = x_{00} \left(\dfrac{\lambda}{\mu_v}\right)^{m-1} \dfrac{\theta r}{\mu_b (1-r)}, \end{cases} \tag{7}$$

where x_{00} is a random real number, so Eq. (5) has positive solution. Thus, the QBD process $\{Q_1(t), J_1(t)\}$ is positive recurrent if and only if SP $(R) = \max(r, \rho) < 1$.

2.2 Queue length distribution

If $\rho < 1$, let (Q_1, J_1) be the stationary limit of the QBD process $\{Q_1(t), J_1(t)\}$. Introduce

$$\pi_k = \pi_{k0}, \quad 0 \le k \le m-1; \quad \pi_k = (\pi_{k0}, \pi_{k1}), \quad k \ge m$$
$$\pi_{kj} = P\{Q = k, J = j\} = \lim_{t \to \infty} P\{Q(t) = k, J(t) = j\}, \quad (k, j) \in \Omega.$$

Theorem 2 *If $\rho < 1$, the stationary probability distribution of (Q_1, J_1) is*

$$\begin{cases} \pi_{k0} = K \left(\dfrac{\lambda}{\mu_v}\right)^k, \quad 0 \le k \le m-1, \\ \pi_{k0} = K \left(\dfrac{\lambda}{\mu_v}\right)^{m-1} r^{k-m+1}, \quad k \ge m; \\ \pi_{k1} = K \left(\dfrac{\lambda}{\mu_v}\right)^{m-1} \dfrac{\theta r}{\mu_b (1-r)} \displaystyle\sum_{j=0}^{k-m} r^j \rho^{k-m-j}, \quad k \ge m, \end{cases} \tag{8}$$

where

$$K=\left[\sum_{k=0}^{m-1}\left(\frac{\lambda}{\mu_v}\right)^k+\left(\frac{\lambda}{\mu_v}\right)^{m-1}\frac{r}{1-r}+\left(\frac{\lambda}{\mu_v}\right)^{m-1}\frac{\theta r}{\mu_b(1-r)}\frac{1}{1-r}\frac{1}{1-\rho}\right]^{-1}.$$

Proof With the matrix-geometric solution method(see in [8]), we have

$$\pi_k=(\pi_{k0},\pi_{k1})=(\pi_{m0},\pi_{m1})R^{k-m},\quad k\geq m. \tag{9}$$

and $(\pi_{00},\pi_{10},\ldots,\pi_{m-1,0},\pi_{m0},\pi_{m1})$ satisfies the set of equations

$$(\pi_{00},\pi_{10},\ldots,\pi_{m-1,0},\pi_{m0},\pi_{m1})B[R]=0. \tag{10}$$

We have obtained the expressions for $\pi_{k0}, 0\leq k\leq m$ and π_{m1} in Theorem 1. Thus, for $k\geq m$, note that

$$R^k=\begin{bmatrix} r^k & \dfrac{\theta r}{\mu_b(1-r)}\displaystyle\sum_{j=0}^{k-1}r^j\rho^{k-1-j} \\ 0 & \rho^k \end{bmatrix},\qquad k\geq 1.$$

With (7), substituting (π_{m0},π_{m1}) and R^{k-m} into (9), we obtain (8). Finally, the constant factor K can be determined by the normalization condition.

Further, we can obtain the distribution for the number of customers Q_1

$$P\{Q_1=k\}=\pi_{k0}=K\left(\frac{\lambda}{\mu_v}\right)^k,\quad 0\leq k\leq m-1;$$

$$P\{Q_1=k\}=\pi_{k0}+\pi_{k1}=K\left(\frac{\lambda}{\mu_v}\right)^{m-1}\left[r^{k-m+1}+\frac{\theta r}{\mu_b(1-r)}\sum_{j=0}^{k-m}r^j\rho^{k-m-j}\right],\quad k\geq m.$$

After some computation, the generating function of Q_1 is as follows

$$Q_1(z)=K\frac{1-(\rho_0 z)^m}{1-\rho_0 z}+K{\rho_0}^{m-1}z^m\left[\frac{r}{1-rz}+\frac{\theta r}{\mu_b(1-r)}\frac{1}{1-rz}\frac{1}{1-\rho z}\right],\quad \rho_0=\frac{\lambda}{\mu_v}$$

Thus,

$$E(Q_1)=K\frac{1-\rho_0^m}{1-\rho_0}\left[\frac{\rho_0}{1-\rho_0}-m\frac{\rho_0^m}{1-\rho_0^m}\right]+K\rho_0^{m-1}\frac{r}{1-r}\left[m+\frac{r}{1-r}\right]+K\rho_0\frac{\theta r}{\mu_b(1-r)}\frac{1}{1-r}\frac{1}{1-\rho}\left[m+\frac{r}{1-r}+\frac{\rho}{1-\rho}\right].$$

Meanwhile, we can easily obtain the state probabilities of a server in steady-state.

$$\begin{aligned} P_v &= P\{J_1 = 0\} = \sum_{k=0}^{\infty} \pi_{k0} = K\left[\frac{1-\rho_0^m}{1-\rho_0} + \rho_0^{m-1}\frac{r}{1-r}\right], \\ P_b &= P\{J_1 = 1\} = \sum_{k=m}^{\infty} \pi_{k1} = K\rho_0^{m-1}\frac{\theta r}{\mu_b(1-r)}\frac{1}{1-r}\frac{1}{1-\rho}. \end{aligned} \tag{11}$$

Remark 1 Many models studied before are the special examples of the model we consider above.

When $m = 1$, i.e., the server only begins the vacation when the system becomes empty, we can obtain the results of M/M/1 queue with working vacations (see Liu et al. [8]).

When $\mu_v = 0$, i.e., the server does't take service during the vacation period, our model becomes the classical M/M/1 queue with vacations and m-policy. Meanwhile, if $\theta = 0$, $\mu_v = \mu_b$, the model becomes the classical M/M/1 queue without vacation.

2.3 Conditional queue length and sojourn time

Note the expressions for $Q_1^{(m)}$ and $S^{(m)}$ below:

$$\begin{aligned} Q_1^{(m)} &= \{Q_1 - m | Q_1 \geq m, J = 1\}; \\ S_m^b &= \{S | Q_1 \geq m, J = 1\}. \end{aligned}$$

$Q_1^{(m)}$ represents the number of customers in the system except for m customers, and S_m^b represents the sojourn time when the server is in the normal working level.

Firstly, we discuss the conditional number of waiting customers.

Theorem 3 *If $\rho < 1$ and $\mu_b > \mu_v$, the conditional stationary queue length $Q_1^{(m)}$ can be decomposed into the sum of three independent random variables: $Q_1^{(m)} = Q_0 + Q_{1d}$, where Q_0 is the stationary queue length of a classical M/M/1 queue without vacation, and follows a geometric distribution with parameter $1-\rho$; Additional queue length Q_{1d} follows geometric distribution with parameter $1-r$.*

Proof Conditional probability that the server is busy and there are more than or equal to m customers in the system

$$P\{Q_1 \geq m, J = 1\} = P\{J = 1\} = \sum_{k=m}^{\infty} \pi_{k1} = K\rho_0^{m-1}\frac{\theta r}{\mu_b(1-r)}\frac{1}{1-r}\frac{1}{1-\rho}.$$

So, for $k \geq 0$

$$P\{Q_1^{(m)} = k\} = P\{Q_1 = k + m | Q_1 \geq m, J = 1\} = \frac{\pi_{k+m,1}}{P\{Q \geq m, J = 1\}}$$
$$= (1-\rho)(1-r)\sum_{j=0}^{k} r^j \rho^{k-j}, \quad k \geq 0.$$

Thus, we easily obtain the probability generating function of $Q_1^{(m)}$ as follows

$$Q_1^{(m)}(z) = \sum_{k=0}^{\infty} z^k P\{Q_1^{(m)} = k\} = \frac{1-\rho}{1-\rho z}\frac{1-r}{1-rz} = Q_0(z)Q_{1d}(z).$$

With the conditional stochastic decomposition structure in Theorem 3, we can easily get the expected number of customers when the server is in the normal busy period.

$$E(Q_1^{(m)}) = \frac{\rho}{1-\rho} + E(Q_{1d}) = \frac{\rho}{1-\rho} + \frac{r}{1-r}.$$

Now, we analyze the conditional sojourn time of each customer when the server is busy and there are more than or equal to m customers in the system as we denote above.

Lemma 3 (i) *The LST of the conditional sojourn time when the server is busy is given in*

$$S_m^{*b}(s) = \left(1 - \frac{s}{\lambda}\right)^m \frac{\mu - \lambda}{\mu - \lambda + s}\frac{\lambda(1-r)}{\lambda(1-r) + rs} \tag{12}$$

(ii) *The conditional expected sojourn time when the server is busy can be expressed by*

$$E(S_m^b) = \frac{E(Q_1^{(m)}) + m}{\lambda}. \tag{13}$$

Proof From the memoryless of exponential distribution and the Little-formula, if a customer departs when the server is busy and there are more than or equal to m customers in the system, the remaining customers should be those who arrive during his sojourn time, that means

$$S_m^{*b}(\lambda(1-z)) = z^m Q_1^{(m)}(z)$$

then, we have

$$S_m^{*b}(s) = \left(1 - \frac{s}{\lambda}\right)^m Q_1^{(m)}\left(1 - \frac{s}{\lambda}\right) = \left(1 - \frac{s}{\lambda}\right)^m \frac{\mu - \lambda}{\mu - \lambda + s}\frac{\lambda(1-r)}{\lambda(1-r) + rs}$$

Further, from the expression for $Q_1^{(m)}$, the relation that $Q_1^{(m)} + m = \{Q_1 | Q_1 \geq m, J = 1\}$ exists, then from the Little law, the conditional expected sojourn time satisfies

$$E(Q_1^{(m)}) + m = \lambda S_m^{*b}.$$

So the conditional expected sojourn time is given as Eq. (13).

Similarly, denote Q_1^v and S_m^v as the conditional queue length and sojourn time when the server is in the vacation period, i.e.,

$$Q_1^v = \{Q_1 | J = 0\}; \qquad S_m^v = \{S | J = 0\}.$$

Firstly, we can compute the probability generating function of Q_1^v as follows

$$Q_1^v(z) = \sum_{k=0}^{\infty} z^k P\{Q_1^v = k\} = \sum_{k=0}^{\infty} z^k \frac{\pi_{k0}}{P\{J=0\}} = \frac{\dfrac{1-(\rho_0 z)^m}{1-\rho_0 z} + (\rho_0 z)^{m-1}\dfrac{rz}{1-rz}}{\dfrac{1-\rho_0^m}{1-\rho_0} + \rho_0^{m-1}\dfrac{r}{1-r}}$$

Then, the expected number of customers when the server is in the vacation period is given by

$$E(Q_1^v) = \frac{\dfrac{1-\rho_0^m}{1-\rho_0}\dfrac{\rho_0}{1-\rho_0} + \rho_0^{m-1}\dfrac{r}{1-r}\left[m + \dfrac{r}{1-r}\right]}{\dfrac{1-\rho_0^m}{1-\rho_0} + \rho_0^{m-1}\dfrac{r}{1-r}}.$$

Lemma 4 (i) *The LST of the conditional sojourn time under the vacation period is given in*

$$S_m^{*v}(s) = \frac{\dfrac{1-\left(\rho_0 - \dfrac{s}{\mu_v}\right)^m}{1-\left(\rho_0 - \dfrac{s}{\mu_v}\right)} + \left(\rho_0 - \dfrac{s}{\mu_v}\right)^{m-1}\dfrac{r(\lambda - s)}{\lambda - r(\lambda - s)}}{\dfrac{1-\rho_0^m}{1-\rho_0} + \rho_0^{m-1}\dfrac{r}{1-r}} \tag{14}$$

(ii) *The conditional expected sojourn time under the vacation period can be expressed by*

$$E(S_m^v) = \frac{E(Q_1^v)}{\lambda}. \tag{15}$$

Proof From the memoryless of exponential distribution and the Little-formula, if a customer departs when the server is in vacation period, the remaining customers should be those who arrive during his sojourn time, which means

$$S_m^{*v}(\lambda(1-z)) = Q_1^v(z)$$

then, we have

$$S_m^{*v}(s) = Q_1^v\left(1 - \frac{s}{\lambda}\right)$$

from which, the Eq. (14) is obtained. Further, the conditional expected sojourn time satisfies

$$E(Q_1^v) = \lambda E(S_m^v).$$

Then the conditional expected sojourn time is given as Eq. (15).

Theorem 4 *For an arbitrary customer who arrives to the system, the Laplace transform and mean of his sojourn time should be*

$$\begin{aligned} S_m^*(s) &= P_v S_m * v(s) + P_b S_m^{*b}(s) = Q_1\left(1 - \frac{s}{\lambda}\right), \\ E(S_m) &= P_v E(S_m^v) + P_b E(S_m^b) = E(Q_1)/\lambda, \end{aligned} \tag{16}$$

where $S_m^{*v}(s)$, $S_m^{*b}(s)$, $E(S_m^v)$, $E(S_m^b)$ *are given in Eqs.* (12)–(15)*, respectively.*

3 M/M/1 queue with (m, N)-policy

3.1 QBD model

Consider a classical M/M/1 queue with arrival rate λ and service rate μ_b (see Gross and Harris [3]). After a service, the server begins a vacation of random length at the instant when the queue length is below m and vacation duration V follows an exponential distribution with parameter θ. During a vacation, the original customers or arriving customers in a vacation period can be served at a mean rate of μ_v. When a vacation ends, if the number of customers in the queue is less than N , another vacation is taken; Otherwise, the server switches service rate from μ_v to μ_b, and a regular busy period starts. This service discipline is a two-threshold policy with working vacation. Evidently, this model is an non-exhaustive service queue and the server can begin to take the vacation when there are customers in system. Many service systems are the special cases of this model and when $m = N$, this model becomes the M/M/1 queue with m-policy in Sect. 2.

Let $Q_2(t)$ be the number of customers in system at time t and let

$$J_2(t) = \begin{cases} 0, & \text{the system is in a working vacation period at time } t, \\ 1, & \text{the system is in a regular busy period at time } t. \end{cases}$$

then $\{Q_2(t), J_2(t)\}$ is a QBD with the state space

$$\Omega = \Big\{(k, 0) : 0 \le k \le m - 1\Big\} \bigcup \Big\{(k, j) : k \ge m, \quad j = 0, 1\Big\}.$$

Using the lexicographical sequence for the states, the infinitesimal generator can be written as

$$Q_2 = \begin{bmatrix} -\lambda & C_0 & & & & & & & & & \\ B_1 & A_1 & C_1 & & & & & & & & \\ & \ddots & \ddots & \ddots & & & & & & & \\ & & B_{m-1} & A_{m-1} & C_{m-1} & & & & & & \\ & & & B_m & A_m & C & & & & & \\ & & & \ddots & \ddots & \ddots & & & & & \\ & & & & B_{N-1} & A_{N-1} & C & & & & \\ & & & & & B_N & A & C & & & \\ & & & & & & B & A & C & & \\ & & & & & & & \ddots & \ddots & \ddots & \end{bmatrix},$$

where

$$\begin{aligned}
&B_k = \mu_v, 1 \le k \le m-1, B_m = (\mu_v, \mu_b)^T;\\
&C_k = \lambda, 0 \le k \le m-2, C_{m-1} = (\lambda, 0); A_k = -(\lambda + \mu_v), 1 \le k \le m-1;\\
&B_k = \begin{bmatrix} \mu_v & 0 \\ 0 & \mu_b \end{bmatrix}, m+1 \le k \le N-1;\\
&A_k = \begin{bmatrix} -(\lambda+\mu_v) & 0 \\ 0 & -(\lambda+\mu_b) \end{bmatrix}, m \le k \le N-1.\\
&C = \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}, B = \begin{bmatrix} \mu_v & 0 \\ 0 & \mu_b \end{bmatrix} A = \begin{bmatrix} -(\lambda+\theta+\mu_v) & \theta \\ 0 & -(\lambda+\mu_b) \end{bmatrix}.
\end{aligned}$$

To analyze this QBD process, it is necessary to solve for the minimal non-negative solution of the matrix quadratic equation

$$R^2 B + RA + C = 0. \tag{17}$$

Because that B, A and C, R has the same expression (2) with that in M/M/1 queue with m-policy, we have

Theorem 5 *The QBD process* $\{Q_2(t), J_2(t)\}$ *is positive recurrent if and only if* $\rho < 1$.

Proof Based on the theorem 3.1.1 of Neuts [6], the QBD process $\{Q_2(t), J_2(t)\}$ is positive recurrent if and only if the spectral radius SP (R) of the rate matrix R is less than 1, and set of equations

$$(x_0, x_{10}, \ldots, x_{m-1,0}, x_{m0}, x_{m1}, \ldots, x_{N0}, x_{N1}) B[R] = 0 \tag{18}$$

has positive solution, where

$$B[R] = \begin{bmatrix} -\lambda & C_0 & & & & & & & \\ B_1 & A_1 & C_1 & & & & & & \\ & \ddots & \ddots & \ddots & & & & & \\ & & B_{m-1} & A_{m-1} & C_{m-1} & & & & \\ & & & B_m & A_m & C & & & \\ & & & \ddots & \ddots & \ddots & & & \\ & & & & B_{N-1} & A_{N-1} & C & & \\ & & & & & B_N & RB+A & & \end{bmatrix}$$

and

$$RB + A = \begin{bmatrix} -\frac{\lambda}{r} & \frac{\lambda}{r} - \mu_v \\ 0 & -\mu_b \end{bmatrix}.$$

With (4), we can easily verify $B[R]$ is an infinitesimal generator, that (18) has positive solution. Thus, the QBD process $\{Q_2(t), J_2(t)\}$ is positive recurrent if and only if $\mathrm{SP}(R) = \max(r, \rho) < 1$.

3.2 Queue length distribution

If $\rho < 1$, let (Q_2, J_2) be the stationary limit of the QBD process $\{Q_2(t), J_2(t)\}$. Let

$$\begin{aligned} &\pi_k = \pi_{k0}, \quad 0 \le k \le m-1; \pi_k = (\pi_{k0}, \pi_{k1}), \quad k \ge m \\ &\pi_{kj} = P\{Q_v = k, J = j\} = \lim_{t\to\infty} P\{Q_v(t) = k, J(t) = j\}, \quad (k, j) \in \Omega. \end{aligned}$$

For convenience, let

$$\begin{aligned} &\psi(k) = 1 + (1-r) \sum_{j=1}^{N-1-k} \left(\frac{\mu_v}{\lambda}\right)^j, \quad 1 \le k \le N-1; \\ &\psi(N) = r. \end{aligned} \tag{19}$$

Theorem 6 *If $\rho < 1$, the stationary probability distribution of (Q_2, J_2) is*

$$\begin{cases} \pi_{k0} = K(\frac{\lambda}{\mu_v})^k, \quad 0 \le k \le m-1, \\ \pi_{k0} = K\dfrac{\psi(k)}{\psi(m-1)}\left(\dfrac{\lambda}{\mu_v}\right)^{m-1}, \quad m \le k \le N; \\ \pi_{k0} = K\beta_{N0} r^{k-N}, \quad k \ge N; \\ \pi_{k1} = K\dfrac{1}{\psi(m-1)}\dfrac{\theta r}{\mu_b(1-r)}\left(\dfrac{\lambda}{\mu_v}\right)^{m-1}\dfrac{1-\rho^{k-m+1}}{1-\rho}, \quad m \le k \le N; \\ \pi_{k1} = K\beta_{N0}\dfrac{\theta r}{\mu_b(1-r)}\displaystyle\sum_{j=0}^{k-N-1} r^j \rho^{k-N-1-j} + K\beta_{N1}\rho^{k-N}, \quad k \ge N+1. \end{cases} \tag{20}$$

where

$$\beta_{N0} = \frac{r}{\psi(m-1)}\left(\frac{\lambda}{\mu_v}\right)^{m-1};$$
$$\beta_{N1} = \frac{1}{\psi(m-1)}\frac{\theta r}{\mu_b(1-r)}\left(\frac{\lambda}{\mu_v}\right)^{m-1}\frac{1-\rho^{N-m+1}}{1-\rho}.$$

And, K can be achieved by the normalization condition.

Proof With the matrix-geometric solution method, we have

$$\pi_k = (\pi_{k0}, \pi_{k1}) = (\pi_{N0}, \pi_{N1})R^{k-N}, \quad k \geq N. \tag{21}$$

and $(\pi_{00}, \pi_{10}, \ldots, \pi_{m-1,0}, \pi_{m0}, \pi_{m1}, \ldots, \pi_{N0}, \pi_{N1})$ satisfies the set of equations

$$(\pi_{00}, \pi_{10}, \ldots, \pi_{m-1,0}, \pi_{m0}, \pi_{m1}, \ldots, \pi_{N0}, \pi_{N1})B[R] = 0 \tag{22}$$

Substituting $B[R]$ into the above relation, we obtain the set of equations

$$\begin{cases} -\lambda\pi_{00} + \mu_v\pi_{10} = 0, \\ \lambda\pi_{k-1,0} - (\lambda+\mu_v)\pi_{k0} + \mu_v\pi_{k+1,0} = 0, \quad 1 \leq k \leq m-2; \\ \lambda\pi_{m-2,0} - (\lambda+\mu_v)\pi_{m-1,0} + \mu_v\pi_{m,0} + \mu_b\pi_{m1} = 0, \quad k = m-1; \\ -(\lambda+\mu_b)\pi_{m1} + \mu_v\pi_{m+1,1} = 0, \qquad k = m; \\ \lambda\pi_{k-1,1} - (\lambda+\mu_b)\pi_{k1} + \mu_b\pi_{k+1,1} = 0, \quad m+1 \leq k \leq N-1; \\ \lambda\pi_{k-1,0} - (\lambda+\mu_v)\pi_{k0} + \mu_v\pi_{k+1,0} = 0, \quad m \leq k \leq N-1; \\ \lambda\pi_{N-1,0} - \dfrac{\lambda}{r}\pi_{N0} = 0; \\ \lambda\pi_{N-1,1} + (\dfrac{\lambda}{r} - \mu_v)\pi_{N0} - \mu_b\pi_{N1} = 0. \end{cases} \tag{23}$$

Assume that every equation in (23) can be expressed by (23–1) to (23–8), respectively. Taking $\pi_{00} = K$, from (23–1) to (23–2), we get

$$\pi_{k0} = K\left(\frac{\lambda}{\mu_v}\right)^k, \quad 1 \leq k \leq m-1$$

from (23–6),

$$\pi_{k0} = \sum_{j=0}^{k-m}\left(\frac{\lambda}{\mu_v}\right)^k\pi_{m0} - \sum_{j=1}^{k-m}\left(\frac{\lambda}{\mu_v}\right)^k\pi_{m-1,0}; \quad m \leq k \leq N;$$

and from (23–7) and the above equation, we get

$$\pi_{m0} = \frac{1+(1-r)\sum_{j=1}^{N-m-1}\left(\frac{\mu_v}{\lambda}\right)^j}{1+(1-r)\sum_{j=1}^{N-m}\left(\frac{\mu_v}{\lambda}\right)^j}\pi_{m-1,0} = \frac{\psi(m)}{\psi(m-1)}\pi_{m-1,0},$$
$$\pi_{k0} = \frac{1+(1-r)\sum_{j=1}^{N-k-1}\left(\frac{\mu_v}{\lambda}\right)^j}{1+(1-r)\sum_{j=1}^{N-m}\left(\frac{\mu_v}{\lambda}\right)^j}\pi_{m-1,0} = \frac{\psi(k)}{\psi(m-1)}\pi_{m-1,0}; \quad m \le k \le N \tag{24}$$

From (23–4), (23–5) and (23–8), we can verify step by step

$$\begin{aligned}\pi_{m1} &= \frac{\theta r}{\mu_b(1-r)}\frac{1}{\psi(m-1)}\pi_{m-1,0},\\ \pi_{k1} &= \sum_{k=0}^{k-m}\left(\frac{\lambda}{\mu_b}\right)^k\frac{\theta r}{\mu_b(1-r)}\frac{1}{\psi(m-1)}\pi_{m-1,0}, \quad m \le k \le N;\end{aligned} \tag{25}$$

Substituting $\pi_{m-1,0}$ and we can get the results for $1 \le k \le N$.

For $k \ge N$, note that

$$R^k = \begin{bmatrix} r^k & \frac{\theta r}{\mu_b(1-r)}\sum_{j=0}^{k-1} r^j\rho^{k-1-j} \\ 0 & \rho^k \end{bmatrix}, \qquad k \ge 1.$$

Substituting (π_{N0}, π_{N1}) and R^{k-N} into (21) , then with (23) and (24), we obtain (20). Finally, the constant factor K can be determined by the normalization condition.

Further, we can obtain the distribution of the number of customers Q_2:

$$\begin{aligned}P\{Q_2 = k\} &= \pi_{k0} = K\left(\frac{\lambda}{\mu_v}\right)^k, \quad 0 \le k \le m-1;\\ P\{Q_2 = k\} &= \pi_{k0} + \pi_{k1} = K\frac{1}{\psi(m-1)}\left(\frac{\lambda}{\mu_v}\right)^{m-1}\\ &\quad \times\left[\psi(k) + \frac{\theta r}{\mu_b(1-r)}\frac{1-\rho^{k-m+1}}{1-\rho}\right],\\ &\quad m \le k \le N-1;\\ P\{Q_2 = k\} &= \pi_{k0} + \pi_{k1}\\ &= K\left[\beta_{N0}r^{k-N} + \beta_{N0}\frac{\theta r}{\mu_b(1-r)}\sum_{j=0}^{k-N-1} r^j\rho^{k-N-1-j} + \beta_{N1}\rho^{k-N}\right],\\ &\quad k \ge N.\end{aligned}$$

Meanwhile, we can easily obtain the state probabilities of a server in steady-state.

$$
\begin{aligned}
P\{J_2=0\} &= \sum_{k=0}^{\infty}\pi_{k0} \\
&= K\left[\sum_{k=0}^{m-1}\left(\frac{\lambda}{\mu_v}\right)^k + \sum_{k=m}^{N-1}\frac{\psi(k)}{\psi(m-1)}\left(\frac{\lambda}{\mu_v}\right)^{m-1} + \sum_{k=N}^{+\infty}\beta_{N0}r^{k-N}\right] \\
&= K\left[\frac{1-\left(\frac{\lambda}{\mu_v}\right)^m}{1-\frac{\lambda}{\mu_v}} + \frac{1}{\psi(m-1)}\left(\frac{\lambda}{\mu_v}\right)^{m-1}\sum_{k=m}^{N-1}\psi(k) + \frac{\beta_{N0}}{1-r}\right] \\
P\{J_2=1\} &= \sum_{k=m}^{\infty}\pi_{k1} = K\frac{1}{\psi(m-1)}\left(\frac{\lambda}{\mu_v}\right)^{m-1}\frac{\theta r}{\mu_b(1-r)}\sum_{k=m}^{N}\frac{1-\rho^{k-m+1}}{1-\rho} \\
&\quad + K\sum_{k=N+1}^{+\infty}\left(\beta_{N0}\frac{\theta r}{\mu_b(1-r)}\sum_{j=0}^{k-N-1}r^j\rho^{k-N-1-j} + \beta_{N1}\rho^{k-N}\right) \\
&= K\frac{1}{\psi(m-1)}\left(\frac{\lambda}{\mu_v}\right)^{m-1}\frac{\theta r}{\mu_b(1-r)}\frac{(N-m)(1-\rho)-\rho(1-\rho^{N-m})}{1-\rho^2} \\
&\quad + K\left[\beta_{N0}\frac{\theta r}{\mu_b(1-r)}\frac{1}{1-r}\frac{1}{1-\rho} + \beta_{N1}\frac{\rho}{1-\rho}\right].
\end{aligned}
\tag{26}
$$

3.3 Conditional queue length and sojourn time

Now, we give conditional stochastic decomposition structures of the stationary length of waiting customers when the server is busy and there are more than or equal to N customers in the system, denoted by Q_2^N.

We can have the expression for Q_2^N below

$$Q_2^N = \{Q_2 - N | Q_2 \geq N, J = 1\};$$

Firstly, we discuss the conditional number of waiting customers.

Theorem 7 *If $\rho < 1$ and $\mu_b > \mu_v$, the conditional stationary queue length Q_2^N can be decomposed into the sum of two independent random variables: $Q_2^N = Q_0 + Q_{2d}$, where Q_0 is the stationary queue length of a classical M/M/1 queue without vacation, follows a geometric distribution with parameter $1-\rho$; Additional queue length Q_{2d} has a modified geometric distribution*

$$
\begin{aligned}
P\{Q_{2d}=0\} &= \frac{\beta_{N1}}{\delta}, \\
P\{Q_{2d}=k\} &= \left(1-\frac{\beta_{N1}}{\delta}\right)(1-r)r^k, \qquad k \geq 1.
\end{aligned}
\tag{27}
$$

where

$$\delta = \beta_{N1} + \beta_{N0}\frac{\theta r}{\mu_b(1-r)^2}.$$

Proof Conditional probability that the server is busy and there are more than or equal to N customers in the system

$$\begin{aligned} P\{Q_2 \geq N, J=1\} &= \sum_{k=N}^{\infty} \pi_{k1} \\ &= K\left[\beta_{N0}\frac{\theta r}{\mu_b(1-r)}\sum_{k=N+1}^{+\infty}\sum_{j=0}^{k-N-1} r^j \rho^{k-N-1-j}\beta_{N1}\sum_{k=N}^{+\infty}\rho^{k-N}\right] \\ &= K\left[\beta_{N1}\frac{1}{1-\rho} + \beta_{N0}\frac{\theta r}{\mu_b(1-r)^2}\frac{1}{1-\rho}\right] = K\frac{1}{1-\rho}\delta. \end{aligned}$$

So, for $k \geq 0$

$$\begin{aligned} P\{Q_2^N = k\} = P\{Q = k+N | Q \geq N, J=1\} &= \frac{\pi_{k+N,1}}{P\{Q_2 \geq N, J=1\}} \\ &= \frac{1-\rho}{\delta}\left[\beta_{N1}\rho^k + \beta_{N0}\frac{\theta r}{\mu_b(1-r)}\sum_{j=0}^{k-1} r^j \rho^{k-1-j}\right] \end{aligned}$$

And, the probability generating function of Q^N is as follows

$$\begin{aligned} Q_2^N(z) &= \sum_{k=0}^{\infty} z^k P\{Q_2^N = k\} \\ &= \frac{1-\rho}{\delta}\sum_{k=0}^{\infty} z^k\left[\beta_{N1}\rho^k + \beta_{N0}\frac{\theta r}{\mu_b(1-r)}\sum_{j=0}^{k-1} r^j\rho^{k-1-j}\right] \\ &= \frac{1-\rho}{\delta}\left[\beta_{N1}\frac{1}{1-\rho z} + \beta_{N0}\frac{\theta r}{\mu_b(1-r)}\frac{z}{1-rz}\frac{1}{1-\rho z}\right] \\ &= \frac{1-\rho}{1-\rho z}\frac{1}{\delta}\left[\beta_{N1} + \beta_{N0}\frac{\theta r}{\mu_b(1-r)^2}\frac{(1-r)z}{1-rz}\right] \\ &= \frac{1-\rho}{1-\rho z}Q_d(z) = Q_0(z)Q_{2d}(z). \end{aligned}$$

From the equation $Q_{2d}(z)$, we can get the result.

Equation (27) indicates that the additional delay Q_{2d} can be written as the mixture of two random variables: $Q_{2d} = q_0 X_0 + q_1 X_1$, where $q_0 = \dfrac{\beta_{N1}}{\delta}$, $q_1 = \dfrac{\beta_{N0}}{\delta}\dfrac{\theta r}{\mu_b(1-r)^2}$, and $X_0 \equiv 0$, X_1 follows a geometric distribution with parameter $(1-r)$ on the set $\{1, 2, \ldots\}$.

With the conditional stochastic decomposition structure in Theorem 7, we can easily get means

$$E(Q_{2d}) = \frac{\beta_{N0}}{\delta}\frac{\theta r}{\mu_b(1-r)^3},$$
$$E(Q_2^N) = \frac{\rho}{1-\rho} + E(Q_{2d}) = \frac{\rho}{1-\rho} + \frac{\beta_{N0}}{\delta}\frac{\theta r}{\mu_b(1-r)^3}.$$

Denote the conditional sojourn time by $E(S_2^N)$, we have

$$E(S_2^N) = \frac{E(Q_1^N) + N}{\lambda}.$$

Similar to analysis in Sect. 3, the LSTs of the conditional sojourn times when the server is busy and vacation can be computed by

$$S_{mN}^{*b}(s) = \sum_{k=m}^{\infty} \frac{\pi_{k1}}{P\{J=1\}}\left(1-\frac{s}{\lambda}\right)^k; \qquad S_{mN}^{*v}(s) = \sum_{k=0}^{\infty} \frac{\pi_{k0}}{P\{J=0\}}\left(1-\frac{s}{\lambda}\right)^k.$$

The LST of the sojourn time of an arbitrary customer can be concluded that

$$S_{mN}^{*}(s) = P\{J=1\}S_{mN}^{*b}(s) + P\{J=0\}S_{mN}^{*v}(s) = \sum_{k=0}^{\infty} P\{Q_2=k\}\left(1-\frac{s}{\lambda}\right)^k,$$

4 Performance analysis

In the above analysis, we obtain some performance measures, such as the mean queue length, server's state probability and conditional waiting time in the steady state. The working vacation policy enables the system to operate flexibly and the queue length and waiting time may decrease. Thus, our model should be reasonable to analyze the practical problems. For example, consider an ATM networks, where cell arrivals in a switched virtual channel(SVC) form a poisson process with parameter λ, cell transmission time is an exponential distributed random variable with rate μ_b. When there are less than certain value m cells, we set a period of working vacation, during which arriving cells can be transmitted at a lower rate $\mu_v(\mu_v < \mu_b)$ immediately in order to save the operating cost. The policy of working vacation takes over cell transmission and save switching cost together, therefore, our model is fitter for practical situation than others.

In Table 1, in a SVC, some special performance measures are presented when $\rho = 0.67$ and $\theta = 0.25$ in two cases, where $E(Q_{11})(E(Q_{12}))$, $P_1\{J_1 = 1\}(P_2\{J_1 = 1\})$, $E(S_1)(E(S_2))$ represent the mean number of cells, the state probability of the SVC in the normal period and the processing time with the lower transmission rate $\mu_v = 0.25(\mu_v = 0.5)$, respectively. Evidently, with the increase of m, the state probability of the SVC in the normal period decreases, but the expected mean number

Table 1 Numerical results when $\rho = 0.67$ and $\theta = 0.25$

m	$E(Q_{11})$	$P_1\{J_1 = 1\}$	$E(S_1)$	$E(Q_{12})$	$P_2\{J_1 = 1\}$	$E(S_2)$
1	2.8054	0.5746	8.0334	2.1930	0.4824	6.9649
2	2.9115	0.4397	8.8435	2.2101	0.3339	7.5774
3	2.8930	0.3252	8.9685	2.1961	0.2284	7.5961
4	2.7996	0.2339	8.6634	2.1685	0.1550	7.3265
5	2.6733	0.1646	8.1471	2.1377	0.1046	6.9508
6	2.5427	0.1139	7.5727	2.1086	0.0703	6.5672
7	2.4237	0.0779	7.0299	2.0836	0.0471	6.2222
8	2.3231	0.0529	6.5603	2.0632	0.0315	5.9326

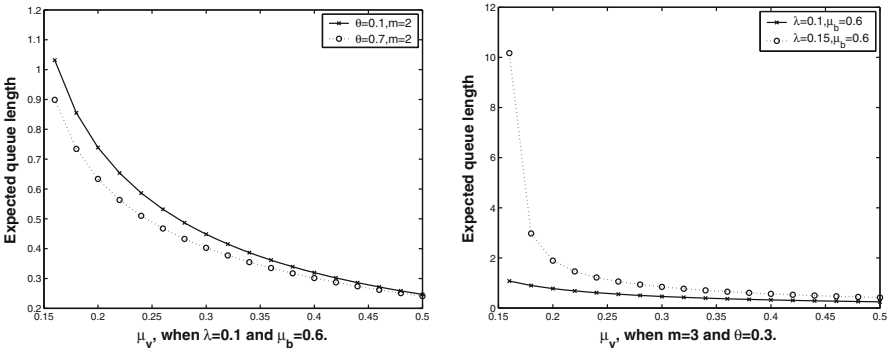

Fig. 1 The *Curve* of expected queue *length* with the change of μ_v

and processing time of cells may not have the decreasing/increasing property. With the increase of the value of m, the expected mean number and processing time of cells may increase first, when m increases to one value, the the expected mean number or processing time of cells begins to decrease. This may be caused by the fact that the SVC also can provide low transmission below m level. When the threshold value m is small($m > 1$), the transfers of two periods should be more frequent to induce more crowd of the cells than that in no threshold case($m = 1$). But when the threshold is increased to one proper value($m > 1$), more cells will be transmitted by the slow rate during the vacation period and certainly the expected mean number and processing time of cells will also decrease. This also demonstrates that the connection of the threshold policy and working vacation will increase the efficiency of the system.

For the model, the different systems may have the different parameters in the practical problems. Certainly, the change of parameters, such as the lower service rate and vacation rate in the system, also may influence the performance measures in the model. So, we present numerical examples in some situations to explain that our model represents some practical problems reasonably well.

According to the expression for $E(Q_1)$, we show the effect of μ_v on the queue length when two parameters of the system are fixed in two situations (see Fig. 1). Evidently, along with the increase of the μ_v, i.e., the service rate in the vacation

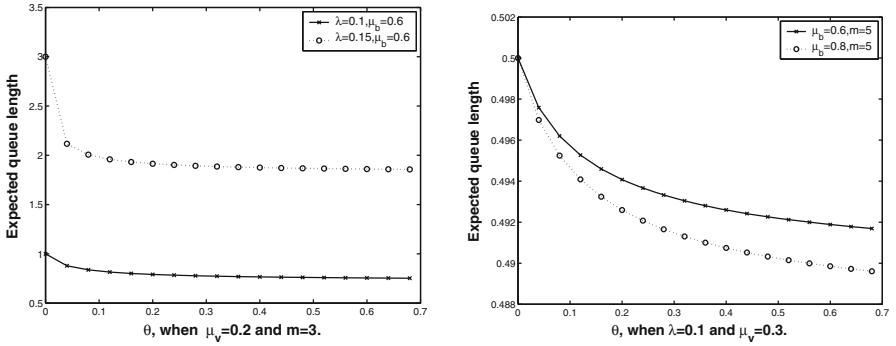

Fig. 2 The *Curve* of expected queue *length* with the change of θ

period, the number of the customers in steady state decreases. And, we also find that the vacation rate and arrival rate have small and large effect on the queue length, respectively. Meanwhile, we only show the trend of the certain range of μ_v, and if the value of μ_v is too small or approaches μ_b, it is not worth to take service in the vacation period. Meanwhile, with the increase of θ, the expected queue length also decreases (see Fig. 2). Such change trends are consistent with the practical situations which can be simulated by the model we consider.

In this model, we set a threshold m and N, and under the thresholds, the server will work at the lower rate μ_v. Thus, the system is a model with two service periods: the higher speed and lower speed periods. Such policy will decrease the service cost, but with the lower service rate, the waiting time and the queue length will increase to make the cost of system rise correspondingly. Thus, we must consider the vacation service rate to minimize the system cost.

The cost of the system is considered. Assume c_w represent the unit time cost of every waiting customer, and c_1 and c_2 are the service costs every unit time during the normal working level and vacation period, respectively. Thus, we can establish the cost function $Z(m, \mu_v)$ per time:

$$Min : Z(m, \mu_v) = c_w E(Q_1) + c_1 \mu_b P\{J_1 = 1\} + c_2 \mu_v P\{J_1 = 0\}$$

where $E(Q_1)$, $P\{J_1 = 1\}$ and $P\{J_1 = 0\}$ have been obtained in sections above.

The optimal m^* and μ_v^* to minimize $Z(m, \mu_v)$ should be found. First, we consider the optimal m. When μ_v is constant, m^* satisfies

$$Z(m^*) \leq Z(m^* + 1), \quad Z(m^*) \leq Z(m^* - 1).$$

By the Boundary analysis method(BAM), the minimal m^* can be given step by step. The basic steps can be showed as follows: Take $m = k (k \geq 1)$, if $Z(k) \leq Z(k+1)$, $Z(k) \leq Z(k-1)$, the optimal threshold $m^* = k$, and we obtain the minimal threshold; otherwise, take $m = k + 1$, continue the same process.

In theory, we should obtain the optimal threshold in this process, but in Fig. 3, we obverse that with the increase of the value m, the system cost may always decrease so

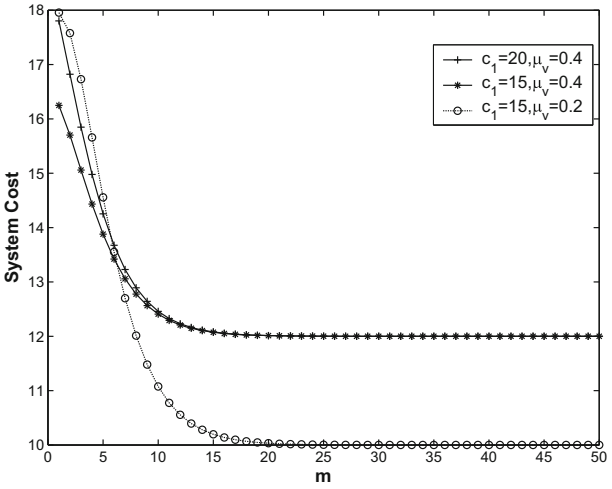

Fig. 3 The *Curve* of system costs with the change of m

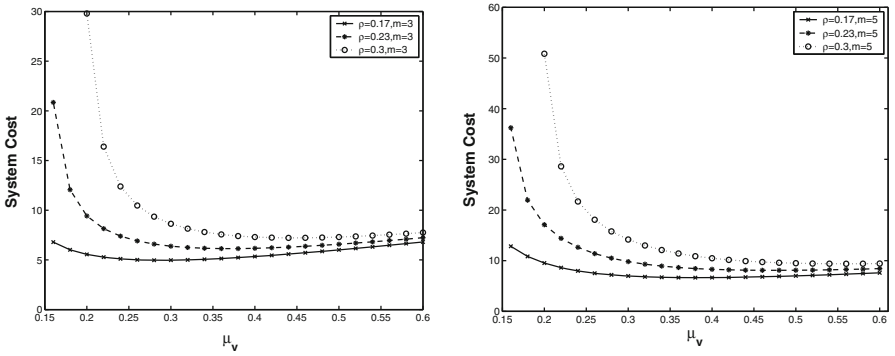

Fig. 4 The *Curve* of system costs with the change of μ_v

that no optimal threshold can be found, but the decreasing trend becomes not evident when m increases to one certain value. This can be explained in practice and if there are not leavings/balkings of the customers and once they arrive at the system, they will be waiting until their service completion, and the service agent controls the whole service process. Under this condition, the larger the threshold is, the smaller the cost is, but when the threshold achieves one certain value, most customers will be served by the slow rate and the effect of the normal service cost c_1 on the system cost will decrease. This also will cause the unwillingness and leaving/balking of the customers. And we will consider this phenomenon in our further research.

If m is given, the optimal vacation service rate μ_v^* can also be found in special situations. In Fig. 4, the system costs when $c_w = 8, c_1 = 20, c_2 = 10$, are presented and from the trend of the curve, the optimal vacation service rate μ_v^* exists ($0 < \mu_v^* < 0.6$).

5 Conclusion

In this paper, we consider the M/M/1 queue with two threshold-policies and working vacations. In fact, we establish the system with lower and higher speed operation periods. Many performance measures, including the state probability of the server and corresponding expected conditional queue length and sojourn time are obtained. With those results, we can further optimize for the (m, N) and the engineers can set up the reasonable thresholds to make the cost of the system lowest or profit highest. But there are some works which the paper cannot give more analysis, for example, the practical sojourn time or its distribution. The service process in two threshold-policy system is so complex that we can not conduct the specific waiting time analysis. This may be the weakness of the model, but we give the sojourn time analysis under two threshold policies which also can give some guides for the practice and further research.

This paper only consider the system indices for the M/M/1 queue with two threshold-policies and working vacations. As we stated in Sect. 4, the larger threshold may induce the customers to leave/balk for other service agents. In further research, we may consider whether the customers' behavior can be analyzed because the behavior may be complex under some information levels if one or two-threshold policy is established under working vacations.

Acknowledgments This research is supported by National Natural Science Foundation of China (No.71301091) and Program for the Outstanding Innovative Teams of Higher Learning Institutions of Shanxi (OIT)

References

1. Baba, Y.: Analysis of a GI/M/1 queue with multiple working vacations. Oper. Res. Lett. **33**, 201–209 (2005)
2. Gross, D., Harris, C.: Fundamentals of Queueing Theory, 2nd edn. Wiley, New York (1985)
3. Kim, J., Choi, D., Chae, K.: Analysis of queue-length distribution of the M/G/1 queue with working vacations. In: International conference on statistics and related fields, Honolulu, 2003
4. Li, J., Tian, N.: The discrete-time GI/Geo/1 queue with working vacations and vacation interruption. Appl. Math. Comput. **185**, 1–10 (2007)
5. Li, J., Tian, N., Liu, W.: Discrete-time GI/Geo/1 queue with working vacations. Queueing Syst. **56**, 53–63 (2007)
6. Neuts, M.: Matrix-Geometric Solutions in Stochastic Models. Johns Hopkins University Press, Baltimore (1981)
7. Servi, L., Finn, S.: M/M/1 queue with working vacations(M/M/1/WV). Perform. Eval. **50**, 41–52 (2002)
8. Liu, W., Xu, X., Tian, N.: Stochastic decompositions in the M/M/1 queue with working vacations. Oper. Res. Lett. **35**, 595–600 (2007)
9. Tian, N., Zhang, Z.G.: Vacation Queueing Models: Theory and Applications. Springer, New York (2006)
10. Tian, N.S., Zhang, Z.G.: A two threshold vacation policy in multiserver queueing systems. Eur. J. Oper. Res. **168**(1), 153–163 (2006)
11. Wu, D., Takagim, H.: M/G/1 queue with multiple working vacations. In: Proceedings of the Queueing Symposium, Stochastic Models and the Applications, pp. 51–60. Kakegawa (2003)
12. Zhang, Z.G., Vichson, R.G., van Eengie, M.J.A.: Optimal two threshold policies in an M/G/1 queue with two vacation types. Perform. Eval. **29**, 63–80 (1997)