ANALYSIS OF A BATCH SERVICE MULTI-SERVER POLLING SYSTEM WITH DYNAMIC SERVICE CONTROL

TAO JIANG*
College of Economics and Management,
Shandong University of Science and Technology
Qingdao, Shandong, 266590, China

LIWEI LIU
Department of Mathematics, School of Science,
Nanjing University of Science and Technology
Nanjing, Jiangsu, 210094, China

(Communicated by Wuyi Yue)

ABSTRACT. This paper considers a multi-server polling system with batch service of an unlimited size, i.e., the so called “Israeli queue” with multi-server, where the service rate of each server switches between a low and a high value depending on the number of groups standing in front of the servers upon its service completion. By means of matrix geometric method and LU-type RG factorization of the infinitesimal generator in irreducible QBD process, the explicit closed-form of rate matrix $R$ and the steady state distribution of the queue length are respectively derived. In terms of the results, some stationary performance measures are obtained. In addition, some numerical examples are presented.

1. Introduction. The so called “Israeli queue” is a special polling system where arriving customers form groups and the service at each group is performed in unlimited size batches. Actually, this queue discipline represents a real situation of a physical waiting line for buying tickets to a movie or a show. New arrivals see only the group leader, i.e., the first customer that originates the group. If a new arrival knows a group leader standing in line, he will join his group. Otherwise, if the new arrival does not know any of the group leaders, he will create a new group. As the leader arrives at the checkout counter, he purchases tickets for the whole customers in his group. Further, the purchasing process is assumed to be independent of the number of tickets purchased.

This model was first introduced and studied by van der Wal and Yechiali [20] so as to analyze a computer tape-reading problem in a system where large amounts of information are stored on tapes. For further studied on this topic, one can refer to ([1], [2]) and so on. Recently, some authors gave more complicated researches on this queue. For example, Perel and Yechiali [13] studied a single-server preemptive priority queueing system with two classes of customers in which the high priority

2010 Mathematics Subject Classification. Primary: 60K25; Secondary: 90B22.

Key words and phrases. Israeli queue, multi-server, dynamic service control, matrix-analytic approach, LU-type RG-factorization.

* Corresponding author: Tao Jiang.
(class-1) customers form a classical infinite-buffer $M/M/1$ queue, while the low priority (class-2) customers form the so-called “Israeli Queue” with batch service and a finite number of groups. In [13], they employed probability generating functions (PGFs) to analyze the model and obtained various performance measures, such as the mean number of low priority groups in the system, the covariance between the number of high and low priority customers, sojourn times of a class-2 group leader and an arbitrary class-2 customer, and the mean size of a class-2 group. Then, Perel and Yechiali [14] extended the Israeli queue model to the case with an infinite number of queues, and analyzed the $M/M/1$, $M/M/c$, and $M/M/1/N$ type queues. After that, they ([15, 16]) further studied the Israeli queue with retrial and the Israeli queue with a general group-joining policy, respectively. In [15], by both PGFs and matrix geometric methods, they calculated analytically various performance measures. In [16], they studied the queue with state dependent group joining policy and derived results for various performance measures, such as steady state distribution of the number of groups in the system, the sojourn times of group leader and an arbitrary customer, the mean size of a group, and lengths of busy periods.

Queueing models with different service rates were studied by various authors in the past. The idea of these models almost made the change of the service rate depending on the environment of the system, such as queues in random environment, queues with breakdown and working breakdown or models with vacations and working vacations. For these queueing systems, we can refer to ([12, 18, 17, 3, 8]). There also exists some literature about the queue models with different service rates mainly focus on the change of service rate depending on the queue size rather than on the environment. For example, the queueing systems with $N$-policy and $T$-policy. For the concrete introduction, the interested readers are referred to ([7, 6, 21, 11]). Recently, Zhang, Wang and Do [23] studied a single server queueing system under $T$-policy to control the service rate and obtained some properties of the key performance measures. Tirdad et al. [19] dealt with optimal control points of $M(t)/M/c/c$ queues with periodic arrival rates and two different levels of servers. Dimitrakopoulos and Burnetas [4] considered the problem of customer equilibrium strategies and optimal strategies in an $M/M/1$ queue with dynamic service control, where the service rate switches between a low and a high value depending on system congestion.

Motivated by applications of the Israeli queue and the queueing system with varying service capacity, especially in transportation and telecommunication networks, in this paper, we study a multi-server Israeli queue with varying service capacity, where the service rate of each server is determined by the number of groups standing in front of the servers upon service completion. We obtain the explicit expressions of the rate matrix $R$ and the stationary state probabilities. Moreover, we also give some numerical examples for the system to demonstrate our results.

The rest of the paper is organized as follows. Model description is given in section 2. Section 3 and section 4 are devoted to giving the expression of rate matrix $R$ and the explicit formulas of stationary state probabilities, respectively. Some system performances in terms of the stationary state probabilities are presented in section 5. Section 6 presents some figures to show the effects of various parameters on the system performance measures. Section 7 is the conclusion.

2. **Model description.** The queueing model under consideration is a multi-server Israeli queue with dynamic service control where the service rate of each server is
determined by the number of groups standing in front of the servers upon service completion. Customers arrive at the system according to a Poisson process with rate $\lambda$, while service times of a batch with fast or low service rate, independent of its size, follow exponential distributions with mean $1/\mu_1$ and $1/\mu_2$, respectively, where $\mu_1 > \mu_2$. A new arriving customer automatically searches for the leader of each group, and if he knows one, he joins his group and receives service in a batch mode together with the group leader. We assume that even if there is at least one vacant server among the $c$ servers, an arriving customer nevertheless searches for a group leader among the groups standing in front of the busy servers, see [14].

We further assume that the probability that an arriving customer knows a group leader standing in line is $\theta$, independent of the group’s size. Specifically, if there are $m$ groups stand in front of the servers, $m = 1, 2, \ldots, c-1$, then the probability that a new arrival joins the $k$th group is $(1-\theta)^{k-1}\theta$, for $1 \leq k \leq m$. When $m = 0, 1, \ldots, c-1$, the probability that an arriving customer creates a new group is $(1-\theta)^m$. However, if $c$ groups stand in front of the servers, i.e., all the servers are busy, and an arriving customer does not know any of the group leaders, he joins the waiting queue with probability $(1-\theta)^c$. Once a service completion, and the waiting queue is non-empty, the customer at the head of the waiting queue is allowed to go to the server immediately.

In our system, in order to analyze the system conveniently, we give further assumptions as follows.

1. We assume that, the servers are divided into two types, one is the fast server type (type 1) and the other one is slow server type (type 2). For each group leader, he first chooses one of the idle fast servers in type 1 to be served, and if all of the fast servers are busy, he then chooses the one of idle slow servers in type 2. This means that type 1 servers have the priority than type 2 servers to serve the customers, i.e., upon a new arrival, the priority is to allocate type 1 servers first and then type 2 servers. If a fast server (with a high service rate $\mu_1$) completes its service and the number of groups standing in front of the servers is less than $c$, the server changes its service rate to the low one $\mu_2$ with probability $p$ and becomes a slow server or retains the high service rate with probability $q = 1-p$. Otherwise, the server remains a fast server. If a slow server (with a low service rate $\mu_2$) completes its service and the number of groups standing in front of the servers is equal to $c$, the server switches its service rate to the high one and becomes a fast server. Otherwise, the server retains the low service rate $\mu_2$.

2. We assume that, at the instant of a service completion, the customer at the head of the waiting queue who is allowed to go to the server, does not search for the leader of each group, but rather joins the idle server and creates a new group to which new arrivals can join, see [15].

3. In order to ensure the priority of type 1 servers, we assume that, moving the customer groups from slow servers to fast servers are allowed at each service completion epoch of a fast server, i.e., at the instant of a busy fast server’s service completion, if the number of groups standing in front of the servers is less than $c$, and the fast server does not change its service rate to low one, then, one of the groups at slow servers (if any) will move to the fast server immediately, and continues to receive service at fast server. Conversely, at the instant of a busy fast server’s service completion, if the number of groups standing in front of the servers is equal to $c$, the fast server continues to serve
the customer who is from waiting queue. This assumption ensures that the priority of fast servers, i.e., if there are busy slow servers in the system, then all fast servers must be busy, similarly, if there are idle fast servers, then all slow servers must be idle. It does not appear the case that there are idle fast servers and busy slow servers together.

3. Rate matrix $R$. Let $N(t)$ be the number of groups standing in front of the servers at time $t$, $J(t)$ be the total number of type 2 servers (slow servers (idle and busy)) at time $t$, and $L(t)$ be the number of customers in the waiting queue at time $t$. The process of the system can be described by the pairs $\{(N(t), J(t), L(t)), t \geq 0\}$, and it forms a continuous-time three-dimension Markov chain with state space

$$\Omega = \{(n, j, 0) \cup (c, j, l), n = 0, 1, \ldots, c, j = 0, 1, \ldots, c, l \geq 1\}.$$ 

In order to avoid confusion, according to the priority of type 1 servers, we give an explicit explanation on the state.

$(n, j, 0), 0 \leq n, j \leq c$ denotes that there are $n$ groups standing in front of servers, the number of slow servers is $j$ and the waiting queue is empty. It indicates that there are $c - j$ fast servers. If $n < c - j$, then the number of busy fast servers is $n$, the number of idle fast server is $c - j - n$, all $j$ slow servers are idle. If $n \geq c - j$, then all $c - j$ fast servers are busy, the number of busy slow servers is $n - (c - j)$ and the number of idle slow servers is $j - (n - (c - j))$.

$(c, j, l), 0 \leq j \leq c, l \geq 1$ denotes that there are $c$ groups standing in front of servers, the number of slow servers is $j$ and $l$ customers are in the waiting queue. It indicates that all servers are occupied, all slow servers and fast servers are busy, i.e., the number of busy slow servers is $j$ and the number of busy fast servers is $c - j$. By referring to the continuous-time Markov process, we can obtain the state-transition-rate matrix as follows

$$Q = \begin{bmatrix}
A_0 & B_0 \\
C_1 & A_1 & B_1 \\
& A_2 & B_2 \\
& & \ddots & \ddots \\
& & & A_c & B_c \\
& & & C_{c+1} & A_c & B_c \\
& & & & C_{c+1} & A_c & B_c \\
& & & & & \ddots & \ddots & \ddots
\end{bmatrix},$$

where all partitioned matrices are square ones with $(c + 1) \times (c + 1)$ orders and for $j = 1, 2, \ldots, c$,

$$C_j = \begin{bmatrix}
d_j & f_j \\
d_j & f_j \\
& \ddots \\
& & d_j & f_j \\
& & d_{j-1} + \mu_2 & f_{j-1} \\
& & \ddots & \ddots \\
& & & d_1 + (j - 1) \mu_2 & f_1 \\
& & & & f_{j-1}
\end{bmatrix}^{(c + 1 - j)},$$

where

$$d_j = jq\mu_1, f_j = jp\mu_1, j = 1, 2, \ldots, c,$$
MULTI-SERVER POLLING SYSTEM WITH DYNAMIC SERVICE CONTROL

\[
C_{c+1} = \begin{bmatrix}
  c\mu_1 & (c-1)\mu_1 \\
  \mu_2 & 2\mu_2 \\
  (c-2)\mu_1 & 3\mu_2 \\
  \cdot & \cdot \\
  \cdot & \cdot \\
  (c-3)\mu_3 & (c-1)\mu_2 & \mu_1 & c\mu_2 & 0
\end{bmatrix},
\]

\[
B_j = \text{diag}(\lambda(1-\theta)^j, \ldots, \lambda(1-\theta)^j), j = 0, 1, \ldots, c,
\]

\[
A_j = \text{diag}(a_{j,1}, a_{j,2}, \ldots, a_{j,c+1}), j = 0, 1, \ldots, c,
\]

with diagonal elements

\[
a_{j,k} = -[\lambda(1-\theta)^j + \min(j, c+1-k)\mu_1 - \min(c+1-j-k, 0)\mu_2], k = 1, 2, \ldots, c+1.
\]

Note that \(Q\) is the infinitesimal generator of the QBD process. To analyze it effectively, an important matrix in analyzing the process is rate matrix \(R\), which is the minimal nonnegative solution of the equation

\[
B_c + RA_c + R^2C_{c+1} = 0. \quad (2)
\]

Next, we investigate the sufficient and necessary stability condition of our model. We will finish the proof based on Neuts [12].

**Theorem 3.1.** The system under consideration is stable if and only if

\[
\rho = \frac{\lambda(1-\theta)^c}{c\mu_1} < 1.
\]

**Proof.** First, define \(M_j(z) = (C_{c+1})_{jj}z^2 + (A_c)_{jj}z + (B_c)_{jj}, 1 \leq j \leq c+1\), where \(X_{jj}, 1 \leq j \leq c+1\) denotes the diagonal elements of matrix \(X\). Since \(M = C_{c+1} + A_c + B_c\) is reducible, then, based on theorem 1.4.1 of Neuts [12], the eigenvalues of \(R\) all lie in the interval \((0, 1)\) (i.e., \(\text{sp}(R) < 1\)) if and only if

\[
\frac{d}{dz} M_j(z)\big|_{z=1} > 0,
\]

for every index \(j, 1 \leq j \leq c+1\), for which \(M_{jj} = 0\) and \((B_c)_{jj} > 0\). From the structure of \(C_{c+1}, A_c, B_c\), we find that \(M_{11} = 0\). Then, \(\frac{d}{dz} M_j(z)\big|_{z=1} > 0\) translates into \(\lambda(1-\theta)^c < c\mu_1\), which leads to

\[
\rho = \frac{\lambda(1-\theta)^c}{c\mu_1} < 1.
\]

Then, the inequality is the necessary and sufficient condition for the system to be stable. \(\square\)

An explicit expression of \(R\) is presented in the following theorem.
Theorem 3.2. If $\rho = \frac{\lambda(1-\theta)^c}{c\mu_1} < 1$, the matrix equation (2) has a minimal non-negative solution

$$R = \begin{bmatrix} r_{1,1} & r_{2,1} & r_{3,1} & \cdots & r_{c,1} & r_{c+1,1} \\ r_{1,2} & r_{2,2} & r_{3,2} & \cdots & r_{c,2} & r_{c+1,2} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ r_{1,c} & r_{2,c} & r_{3,c} & \cdots & r_{c,c} & r_{c+1,c} \\ r_{c+1,1} & r_{c+1,2} & r_{c+1,3} & \cdots & r_{c+1,c} & r_{c+1,c+1} \end{bmatrix},$$

where

$$r_{1,1} = \frac{\lambda(1-\theta)^c + c\mu_1 - |\lambda(1-\theta)^c - c\mu_1|}{2c\mu_1} = \lambda(1-\theta)^c/c\mu_1,$$

$$r_{i,i} = \frac{\theta_i - \sqrt{\theta_i^2 - 4\lambda(1-\theta)^c(c-i+1)\mu_1}}{2(c-i+1)\mu_1},$$

$$\theta_i = \lambda(1-\theta)^c + (c-i+1)\mu_1 + (i-1)\mu_2, i = 2, \ldots, c,$$

$$r_{c+1,c+1} = \frac{\lambda(1-\theta)^c}{\lambda(1-\theta)^c + c\mu_2}, r_{i,j} = 0, i < j,$$

$$r_{i,j} = \sum_{k=j+1}^{i-1} r_{i,k}(r_{k,j}(c-j+1)\mu_1 + r_{k,j+1}j\mu_2) + r_{i,i}r_{i,j+1}j\mu_2,$$

$$i = 2, \ldots, c + 1, i > j.$$

Proof. Since $B_c$ and $A_c$ are diagonal matrices, and $C_{c+1}$ consists of main diagonal and the one below it (all other elements are 0), it follows that $R$ is a lower triangular matrix having the following form

$$R = \begin{bmatrix} r_{1,1} & r_{2,1} & r_{3,1} & \cdots & r_{c,1} & r_{c+1,1} \\ r_{1,2} & r_{2,2} & r_{3,2} & \cdots & r_{c,2} & r_{c+1,2} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ r_{1,c} & r_{2,c} & r_{3,c} & \cdots & r_{c,c} & r_{c+1,c} \\ r_{c+1,1} & r_{c+1,2} & r_{c+1,3} & \cdots & r_{c+1,c} & r_{c+1,c+1} \end{bmatrix}.$$

In the left-hand side of (2), we have

$$R^2C_{c+1} = \begin{bmatrix} \alpha_{1,1} & 0 & 0 & \cdots & 0 \\ \beta_{2,1} & \alpha_{2,2} & 0 & \cdots & 0 \\ \beta_{3,1} & \beta_{3,2} & \alpha_{3,3} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \beta_{c+1,1} & \beta_{c+1,2} & \beta_{c+1,3} & \cdots & \alpha_{c+1,c+1} \end{bmatrix}.$$
where

\[ R_{\text{c}} = \begin{pmatrix}
    d_{1,1} & 0 & 0 & \cdots & 0 \\
    d_{2,1} & d_{2,2} & 0 & \cdots & 0 \\
    d_{3,1} & d_{3,2} & d_{3,3} & \cdots & 0 \\
    \vdots & \vdots & \vdots & \ddots & \vdots \\
    d_{c+1,1} & d_{c+1,2} & d_{c+1,3} & \cdots & d_{c+1,c+1}
\end{pmatrix}, \]

where

\[ \alpha_{k,k} = r_{k,k}^2(c - k + 1)\mu_1, \quad k = 1, 2, \ldots, c, \quad \alpha_{c+1,c+1} = 0, \]

\[ \beta_{k,j} = \sum_{i=j}^{k} r_{k,i}r_{i,j}(c - j + 1)\mu_1 + \sum_{i=j+1}^{k} r_{k,i}r_{i,j+1}\mu_2, \quad k = 2, \ldots, c + 1, \quad j = 1, \ldots, k - 1, \]

\[ d_{i,j} = -(\lambda(1 - \theta)^c + (c - j + 1)\mu_1 + (j - 1)\mu_2)r_{i,j}, \quad i = 1, 2, \ldots, c + 1, \quad j = 1, 2, \ldots, i. \]

Substituting \( R^2C_{c+1}, R_{\text{c}}, B_{\text{c}} \) into (2), we have the following set of equations:

\[ (c - i + 1)\mu_1 r_{i}^2(i,i) - (\lambda(1 - \theta)^c + (c - i + 1)\mu_1 + (i - 1)\mu_2)r_{i,i} + \lambda(1 - \theta)^c = 0, \quad i = 1, 2, \ldots, c, \]

\[ -(\lambda(1 - \theta)^c + c\mu_2)r_{c+1,c+1} + \lambda(1 - \theta)^c = 0, \]

First, the unique solutions in the interval (0, 1) of quadratic equations (4) are given by

\[ r_{i,i} = \frac{\theta_i - \sqrt{\theta_i^2 - 4(\lambda(1 - \theta)^c + (c - i + 1)\mu_1)}}{2(c - i + 1)\mu_1}, \quad i = 2, \ldots, c, \]

where \( \theta_i = \lambda(1 - \theta)^c + (c - i + 1)\mu_1 + (i - 1)\mu_2. \) Then, from (5),

\[ r_{c+1,c+1} = \frac{\lambda(1 - \theta)^c}{\lambda(1 - \theta)^c + c\mu_2}. \]

Finally, if \( \rho = \lambda(1 - \theta)^c/c\mu_1 < 1, \) we have

\[ r_{1,1} = \frac{\lambda(1 - \theta)^c + c\mu_1 - |\lambda(1 - \theta)^c - c\mu_1|}{2c\mu_1} = \lambda(1 - \theta)^c/c\mu_1 = \rho < 1. \]

Once the diagonal elements of \( R \) are known, the elements below the diagonal can be computed recursively by solving the equation (6):

\[ r_{i,j} = \frac{\sum_{k=j+1}^{i-1} r_{i,k}(r_{k,j}(c - j + 1)\mu_1 + r_{k,j+1}\mu_2) + r_{i,i}r_{i,j+1}\mu_2}{\lambda(1 - \theta)^c + (c - j + 1)\mu_1 + (j - 1)\mu_2 - (r_{i,i} + r_{j,j})(c - j + 1)\mu_1}, \]

where \( i = 2, \ldots, c + 1, i > j, \) and \( r_{i,j} = 0 \) for \( i < j. \) So, we obtain the explicit expression of rate matrix \( R. \)

Some relevant papers ([13], [22]) showed how to explicitly calculate the entries of the matrix \( R. \) It is different from ([12], [9]) that the approach proposed by ([13], [22]) bypassed the need to numerically solve the entries of \( R \) with iterative algorithm and gave the closed-form formula of \( R. \) By referring to these papers, we can recursively calculate the other entries below the main diagonal of \( R \) in the following order.
We first compute the elements on the first diagonal below the main diagonal. More precisely, let \( j = i - 1 \) in (6), we get
\[
\lambda_{i,i-1} = \frac{r^2_{i-1}(i-1)\mu_2}{[\lambda(1-\theta)^c + (c-i+2)\mu_1 + (i-2)\mu_2] - (r_{i,i} + r_{i-1,i-1})(c-i+2)\mu_1}.
\]
We continue to calculate the entries on the next diagonal, that is, we assume \( j = i - 2 \), thus equation (6) becomes
\[
\lambda_{i,i-2} = \frac{r_{i,i-1}(i-1)\mu_1 + r_{i-1,i-1}(i-2)\mu_2 + r_{i,i}r_{i,i-1}(i-2)\mu_2}{[\lambda(1-\theta)^c + (c-i+3)\mu_1 + (i-3)\mu_2] - (r_{i,i} + r_{i-2,i-2})(c-i+3)\mu_1},
\]
Using the same recursion with \( j = i - 3, \cdots, 1 \), we can compute all the other entries of \( R \). In general, the \((i, i-l)\)th element of \( R \) is given by
\[
\lambda_{i,i-l} = \sum_{k=i-l+1}^{i-1} r_{i,k}(r_{k,j}(c-i+l+1)\mu_1 + r_{k,i-1}(i-l)\mu_2) + r_{i,i}r_{i,i-l+1}(i-l)\mu_2
\]
\[
= \frac{[\lambda(1-\theta)^c + (c-i+l+1)\mu_1 + (i-l-1)\mu_2] - (r_{i,i} + r_{i-1,i-1})(c-i+l+1)\mu_1}{[\lambda(1-\theta)^c + (c-i+3)\mu_1 + (i-3)\mu_2] - (r_{i,i} + r_{i-2,i-2})(c-i+3)\mu_1},
\]
where \( l = 1, \cdots, i-1 \), \( \sum (\cdot) = 0 \).

4. Steady state probabilities. If \( \rho = \frac{\lambda(1-\theta)^c}{\mu_1} < 1 \), this QBD process
\[
\{(N(t), J(t), L(t)), t \geq 0\}
\]
is positive recurrent, and it can be analyzed in steady state. In this section, we will construct the LU-type RG-factorization for the QBD process to obtain the steady state probabilities. Let
\[
\pi = [\pi_0, \pi_1, \cdots, \pi_c, \pi_{c+1}, \cdots,]
\]
denote the stationary probability vector of \( \{(N(t), J(t), L(t)), t \geq 0\} \), where
\[
\pi_k = (\pi_{k,0,0}, \pi_{k,1,0}, \cdots, \pi_{k,c,0}), 0 \leq k \leq c,
\]
\[
\pi_{c+k} = (\pi_{c,0,k}, \pi_{c,1,k}, \cdots, \pi_{c,c,k}), k \geq 1.
\]
\[
\pi_{k,i,0} = \lim_{t \to \infty} P(N(t) = k, J(t) = i, L(t) = 0), 0 \leq k \leq c, 0 \leq i \leq c,
\]
\[
\pi_{c,i,k} = \lim_{t \to \infty} P(N(t) = c, J(t) = i, L(t) = k), 0 \leq i \leq c, k \geq 1.
\]
As presented in section 1.5 of [12], the QBD process \( Q \) can be re-partitioned as follows:
\[
Q = \begin{bmatrix}
H_0 & H_1 & 0 & 0 & \cdots \\
H_2 & A_c & B_c & 0 & \cdots \\
& 0 & C_{c+1} & A_c & B_c & \cdots \\
& & 0 & C_{c+1} & A_c & \cdots \\
& & & \vdots & \vdots & \ddots & \ddots \\
& & & & 0 & 0 & \cdots & A_{c-2} & B_{c-2} \\
& & & & 0 & 0 & \cdots & 0 & C_{c+1} & A_{c-1} \\
& & & & & & & & \vdots & \vdots & \ddots & \ddots \\
& & & & & & & & & 0 & 0 & \cdots & A_{c-2} & B_{c-2} \\
& & & & & & & & & 0 & 0 & \cdots & 0 & C_{c+1} & A_{c-1}
\end{bmatrix}_{c(c+1) \times c(c+1)}
\]
Note that the matrix sequences \( R \) recurrent if and only if the Markov chain \( B \) is positive recurrent, therefore the Markov chain \( B[R] \) is transient, therefore the matrix \( B[R] \) is invertible for \( 0 \leq k \leq c-1 \). While the Markov chain \( U \) is positive recurrent if and only if the Markov chain \( B[R] \) is positive recurrent.

Based on the \( U \)-measure \( U_k, 0 \leq k \leq c \), we can respectively define the LU-type \( R \)-measures and \( G \)-measures as follows:

\[ R_k = C_k(-U_{k-1}^{-1}), 1 \leq k \leq c-1, G_k = (-U_k^{-1})B_k, 0 \leq k \leq c-1. \]

Note that the matrix sequences \( R_k, 1 \leq k \leq c-1 \) and \( G_k, 1 \leq k \leq c-1 \) are the unique nonnegative solution to the system of matrix equations

\[ R_{k+1}R_kB_{k-1} + R_{k+1}A_k + C_{k+1} = 0, 1 \leq k \leq c-1, \]

and

\[ B_k + A_kG_k + C_kG_{k+1}G_k = 0, 1 \leq k \leq c-1, \]

with the boundary conditions

\[ R_1 = C_1(-U_0^{-1}), G_0 = (-U_0^{-1})B_0. \]

Hence,

\[ R_{k+1} = -C_{k+1}[R_kB_{k-1} + A_k]^{-1}, 1 \leq k \leq c-1, \]

Define the \((c+1)(c+1) \times (c+1)(c+1)\) square matrix as follows:

\[ B[R] = \begin{bmatrix}
A_0 & B_0 & 0 & \cdots & 0 & 0 \\
C_1 & A_1 & B_1 & \cdots & 0 & 0 \\
0 & C_2 & A_2 & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & \cdots & A_{c-1} & B_{c-1} \\
0 & 0 & 0 & \cdots & C_c & RC_{c+1} + A_c
\end{bmatrix}, \]

where \( B[R] \) is an irreducible infinitesimal generator with finite dimensions. Then, we can solve the equilibrium distribution by first noting that \((c+1)(c+1)\)-dimensional vector \([\pi_0, \pi_1, \cdots, \pi_c]\) (i.e., the boundary state probability vector) satisfies the following equation:

\[ [\pi_0, \pi_1, \cdots, \pi_c]B[R] = 0, \]

that is,

\[ [\pi_0, \pi_1, \cdots, \pi_c] \begin{bmatrix}
H_0 & H_1 \\
H_2 & RC_{c+1} + A_c
\end{bmatrix} = 0. \]

Solving \( \pi_0, \pi_1, \cdots, \pi_c \) is key for deriving the steady state probabilities. Next, by using the censoring technique and referring to \([10]\), we construct the LU-type RG-factorization for \( B[R] \) to obtain the expressions of \( \pi_0, \pi_1, \cdots, \pi_c \).

First, we write the following matrix equations:

\[ H_1 = \begin{bmatrix}
0 \\
0 \\
\vdots \\
0
\end{bmatrix}, H_2 = \begin{bmatrix}
0 & 0 & \cdots & 0 & C_c
\end{bmatrix}_{c \times (c+1)}. \]
\[ G_k = -[A_k + C_kG_{k-1}]^{-1}B_k, \quad 1 \leq k \leq c - 1. \]

For the irreducible infinitesimal generator with finite dimensions \( B[R] \), we can easily obtain the following LU-type RG-factorization:

\[
B[R] = (I - R_L)U_D(I - G_U),
\]

where

\[
R_L = \begin{bmatrix}
0 & 0 & \cdots & 0 \\
R_1 & 0 & \cdots & 0 \\
\vdots & \ddots & \ddots & \vdots \\
R_{c-1} & 0 & \cdots & 0 \\
R_c & 0 & \cdots & 0
\end{bmatrix},
G_U = \begin{bmatrix}
0 & G_0 & \cdots & G_{c-1} \\
0 & 0 & \cdots & 0
\end{bmatrix},
U_D = \text{diag}(U_0, U_1, \ldots, U_c).
\]

Substituting LU-type RG-factorization of \( B[R] \) into \( [\pi_0, \pi_1, \cdots, \pi_c]B[R] = 0 \), we have

\[
(Y_0, Y_1, \cdots, Y_c) = (\pi_0, \pi_1, \cdots, \pi_c)(I - R_L).
\]

Using the properties of the \( U \)-measure, we have

\[
\pi_k U_k = 0, \quad 1 \leq k \leq c,
\]

and

\[
\begin{cases}
Y_k = \pi_k - \pi_{k+1}R_{k+1}, & 0 \leq k \leq c - 1, \\
Y_c = \pi_c.
\end{cases}
\]

As the matrix \( U_k \) is invertible for \( 0 \leq k \leq c - 1 \), we have \( Y_k = 0 \), i.e., \( \pi_k = \pi_{k+1}R_{k+1}, 0 \leq k \leq c - 1 \). In terms of \( \pi_c \), \( \pi_k \) can be further obtained by

\[
\pi_k = \pi_c \prod_{i=k+1}^{c} R_i, \quad 0 \leq k \leq c - 1.
\]

According to [12], for \( k \geq c + 1 \), we have

\[
\pi_k = \pi_c R^{k-c}, \quad k \geq c + 1.
\]

Solving the the conditions \( Y_c U_c = 0 \) and the normalization condition \( \sum_{k=0}^{c-1} \pi_k e + \pi_c(I - R)^{-1}e = 1 \), \( \pi_c \) can be obtained.

Then, the steady state probabilities can be obtained by the following theorem.

**Theorem 4.1.** If \( \rho = \frac{\lambda(1-\theta)c}{epu} < 1 \), the stationary state probabilities are given as follows:

\[
\begin{align*}
\pi_k &= (\pi_{k,0,0}, \pi_{k,1,0}, \cdots, \pi_{k,c,0}) = \pi_c \prod_{i=k+1}^{c} R_i, \quad 0 \leq k \leq c - 1, \\
\pi_k &= (\pi_{c,0,k-c}, \pi_{c,1,k-c}, \cdots, \pi_{c,c,k-c}) = \pi_c R^{k-c}, \quad k \geq c + 1,
\end{align*}
\]
where \( \pi_c \) satisfies the following set of equations

\[
\pi_c U_c = (0, 0, \cdots, 0),
\]

\[
\pi_c \sum_{k=0}^{c-1} \prod_{i=k+1}^{c} R_i e + \pi_c (I - R)^{-1} e = 1.
\]

Next, we give a simple analysis of the complexity. According to Elhafsi and Molle [5], for two \( c \times c \) matrices, the addition of the two matrices requires \( c \) operations, the multiplication of the two matrices requires \( c^2 \) operations, and the inversion of a \( c \times c \) matrix requires \( c^3 \) operations. With regard to the method proposed in this article, to compute the stationary probabilities, we have to compute \( R_i, i = 1, 2, \cdots, c \), which requires \( c \) matrix inversions, i.e., the asymptotic complexity is \( O(c^4) \).

5. Performance measures. Performance analysis is an important aspect of the queueing model. From the obtained steady state probabilities, we can derive the following performance measures.

The expected number of customers in the waiting queue

\[
E[L_q] = \pi_{c+1} \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 1 \end{bmatrix} + \pi_{c+2} \begin{bmatrix} 2 \\ 2 \\ \vdots \\ 2 \\ 2 \end{bmatrix} + \cdots = \pi_c Re + \pi_c 2Re + \cdots = \pi_c R(I - R)^{-2} e.
\]

The expected number of idle slow server

\[
E[ILS] = \pi_0 \begin{bmatrix} 0 \\ 1 \\ \vdots \\ c-1 \end{bmatrix} + \pi_1 \begin{bmatrix} 0 \\ 1 \\ \vdots \\ c-1 \end{bmatrix} + \cdots + \pi_{c-1} \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 1 \end{bmatrix}
\]

\[
= \pi_c \prod_{j=1}^{c} R_j \theta_1 + \pi_c \prod_{j=2}^{c} R_j \theta_2 + \cdots + \pi_c \prod_{j=c}^{c} R_j \theta_c = \pi_c \sum_{i=1}^{c} \left( \prod_{j=i}^{c} R_j \right) \theta_i,
\]

where

\[
\theta_1 = (0, 1, \cdots, c-1, c)^T, \theta_2 = (0, 1, \cdots, c-1, c-1)^T, \cdots, \theta_c = (0, 1, \cdots, 1)^T.
\]

The expected number of busy slow server

\[
E[BLS] = \pi_0 \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix} + \pi_1 \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} + \cdots + \pi_{c-1} \begin{bmatrix} 0 \\ 0 \\ \vdots \\ c-2 \\ c-1 \end{bmatrix} + \sum_{i=c}^{\infty} \pi_i \begin{bmatrix} 0 \\ 0 \\ \vdots \\ c-2 \\ 0 \end{bmatrix} + \pi_c (I - R)^{-1} \phi
\]

\[
= \pi_c \prod_{j=2}^{c} R_j \phi_1 + \cdots + \pi_c \prod_{j=c}^{c} R_j \phi_{c-1} + \pi_c (I - R)^{-1} \phi_c
\]
\[ e_{c} = \pi_{c} \left[ \sum_{i=1}^{c-1} \prod_{j=i+1}^{c} R_{j} \phi_{i} \right] + (I - R)^{-1} \phi_{c}, \]

where
\[ \phi_{1} = (0, 0, ..., 0, 1)^{T}, \phi_{2} = (0, 0, ..., 1, 2)^{T}, \cdots, \phi_{c} = (0, 1, ..., c - 1, c)^{T}. \]

The expected number of busy fast server
\[ E[IHS] = \pi_{0} \begin{bmatrix} c & c - 1 & \cdots & 0 \\ c & \cdots & 0 & 0 \end{bmatrix} + \pi_{1} \begin{bmatrix} c & \cdots & 0 \\ c - 1 & \cdots & 0 & 0 \end{bmatrix} + \cdots + \pi_{c-1} \begin{bmatrix} 1 \\ 0 & \cdots & 0 \end{bmatrix} \]
\[ = \pi_{c} \prod_{j=1}^{c} R_{j} g_{1} + \pi_{c} \prod_{j=2}^{c} R_{j} g_{2} + \cdots + \pi_{c} \prod_{j=c}^{c} R_{j} g_{c} = \pi_{c} \sum_{i=1}^{c} \left( \prod_{j=i}^{c} R_{j} \right) g_{i}, \]

where
\[ g_{1} = (c, c - 1, ..., 1, 0)^{T}, g_{2} = (c - 1, c - 2, ..., 0, c)^{T}, \cdots, g_{c} = (1, 0, ..., 0, 0)^{T}. \]



The expected number of busy fast server
\[ E[BHS] = \pi_{0} \begin{bmatrix} 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 \end{bmatrix} + \pi_{1} \begin{bmatrix} 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 1 & \cdots & 0 & 0 \end{bmatrix} + \pi_{c-1} \begin{bmatrix} c - 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 1 & \cdots & 0 & 0 \end{bmatrix} + \sum_{i=c}^{\infty} \pi_{i} \begin{bmatrix} 1 \\ \vdots & \ddots & \ddots & \vdots \\ 1 & \cdots & 0 & 0 \end{bmatrix} \]
\[ = \pi_{c} \prod_{j=2}^{c} R_{j} h_{1} + \cdots + \pi_{c} \prod_{j=c}^{c} R_{j} h_{c-1} + \pi_{c} (I - R)^{-1} h_{c} \]
\[ = \pi_{c} \left[ \sum_{i=1}^{c-1} \prod_{j=i+1}^{c} R_{j} h_{i} \right] + (I - R)^{-1} h_{c}, \]

where
\[ h_{1} = (1, 1, ..., 1, 0)^{T}, h_{2} = (2, 2, ..., 1, 0)^{T}, \cdots, h_{c} = (c, c - 1, ..., 1, 0)^{T}. \]

6. **Numerical examples.** In this section, we illustrate some numerical examples to study the impact of the parameters on the mean number of customers in the waiting line. Without loss of generality, we assume \( c = 2, 3, 4 \) and \( \mu_{2} = 1 \).

In Fig. 1, we plot the trend of the change for the mean number of customers in the waiting queue \( E[L_{q}] \) as the arrival rate \( \lambda \) increases from 1 to 3 for different values of \( c \). Clearly, from Fig. 1, we find that \( E[L_{q}] \) increases with the increase of \( \lambda \), which is identical to the intuitive expectations. It is also obvious that, if \( \lambda \) is fixed, the smaller \( c \) is, the larger \( E[L_{q}] \) becomes.

In Fig. 2, we pay attention to the curves of \( E[L_{q}] \) with the change of service rate \( \mu_{1} \). Fig. 2 indicates that an increase in \( \mu_{1} \) results in a decrease of \( E[L_{q}] \) for different values of \( c \). It is because that, higher service rate \( \mu_{1} \) values lead to more customers leaving the system. We also find that if \( \mu_{1} \) is fixed, the smaller \( c \) is, the larger \( E[L_{q}] \) becomes.
Figure 1. \( L_q \) versus \( \lambda \) \((p = 0.6, \theta = 0.2, \mu_1 = 3)\)

Figure 2. \( L_q \) versus \( \mu_1 \) \((\lambda = 3, p = 0.6, \theta = 0.2, )\)

Figure 3. \( L_q \) versus \( p \) \((\lambda = 3, \mu_1 = 3, \theta = 0.2, )\)
In Fig. 3, we plot the trend of the change for $E[L_q]$ as the probability $p$ increases from 0 to 1 for different values of $c$. From Fig. 3, we can easily find that $E[L_q]$ increases with the increase of $p$ for different values of $c$. Actually, it is in accordance with the fact. Higher probability $p$ values make the fast server more likely to be a slow one at the completion instant, which leads to more customers stranding in the system.

![Figure 4. $L_q$ versus $p$ ($\lambda = 3, \mu_1 = 3, p = 0.6,$)](image)

Fig. 4 shows the trend of the change for $E[L_q]$ as the joining group probability $\theta$ increases from 0 to 0.9 for different values of $c$. From Fig. 4, it is obvious that, with the increase of $\theta$, $E[L_q]$ decreases. It is reasonable that higher $\theta$ means higher probability that the new arrival joins the existing groups standing in front of the servers, which leads to the less customers standing in the waiting queue. We also find that for the fixed $\theta$, the bigger $\theta$ is, the smaller $E[L_q]$ becomes.

7. Conclusion. In this paper, we investigated a batch service multi-server polling system with dynamic service control in order to establish the theoretical foundations for applications and obtain the explicit computation formulas for the performance measures. In terms of matrix analytic method and LU-type RG factorization of the infinitesimal generator in irreducible QBD process, we respectively obtained the explicit expression of the rate matrix $R$ and the stationary state probabilities. Various performance measures are analytically calculated and numerical results are presented to study the effect of various parameters on the system performance measures. We hope that the results can be applied to more practical queueing systems.

Acknowledgments. The authors would like to thank the editor and the referees for the helpful suggestions and comments to improve the quality of this paper.

REFERENCES

[1] O. J. Boxma, Y. van der Wal and U. Yechiali, Polling with gated batch service, in: Proceedings of the Sixth International Conference on “Analysis of Manufacturing Systems”, Lunteren, Netherlands, 2007, 155-159.
[2] O.J. Boxma, Y. van der Wal and U. Yechiali, Polling with batch service, Stochastic Models, 24 (2008), 604–625.
[3] J. D. Cordeiro and J. P. Kharoufeh, The unreliable M/M/1 retrial queue in a random environment, *Stochastic Models*, **28** (2012), 29–48.

[4] Y. Dimitrakopoulos and A. N. Burnetas, Customer equilibrium and optimal strategies in an M/M/1 queue with dynamic service control, *European Journal of Operational Research*, **252** (2016), 477–486.

[5] E. H. Elhafsi and M. Molle, The solution to QBD processes with finite state space, *Stochastic Analysis and Applications*, **25** (2007), 763–779.

[6] D. P. Heyman, The T policy for the M/G/1 queue, *Management Science*, **23** (1977), 775–778.

[7] P. Jayachitra and A. J. Albert, Recent developments in queueing models under N-policy: A short survey, *International Journal of Mathematical Archive*, **5** (2014), 227–233.

[8] K. Kalidass and R. Kasturi, A queue with working breakdowns, *Computers and Industrial Engineering*, **63** (2012), 779–783.

[9] G. Latouche and V. Ramaswami, *Introduction to Matrix Analytic Methods in Stochastic Modeling*, SIAM, Philadelphia, 1999.

[10] Q. L. Li, *Constructive Computation in Stochastic Models with Applications: the RG-Factorizations*, Springer, Berlin and Tsinghua University Press, Beijing, 2010.

[11] Z. Ma, P. Wang and W. Yue, Performance analysis and optimization of a pseudo-fault Geo/Geo/1 repairable queueing system with N-policy, setup time and multiple working vacations, *Journal of Industrial and Management Optimization*, **13** (2017), 1467–1481.

[12] M. F. Neuts, *Matrix-Geometric Solutions in Stochastic Models: Algorithmic Approach*, Johns Hopkins University Press, Baltimore, 1981.

[13] N. Perel and U. Yechiali, The Israeli queue with priorities, *Stochastic Models*, **29** (2013), 353–379.

[14] N. Perel and U. Yechiali, The Israeli queue with infinite number of groups, *Probability in the Engineering and Informational Sciences*, **28** (2014), 1–19.

[15] N. Perel and U. Yechiali, The Israeli Queue with retrials, *Queueing Systems*, **78** (2014), 31–56.

[16] N. Perel and U. Yechiali, The Israeli Queue with a general group-joining policy, *Annals of Operations Research*, (2015), 1–34.

[17] L. D. Servi and S. G. Finn, M/M/1 queues with working vacations (M/M/1WV), *Performances Evaluation*, **50** (2002), 41–52.

[18] N. Tian and Z. G. Zhang, *Vacation Queueing Models: Theory and Applications*, Springer, New York, 2006.

[19] A. Tirdad, W. K. Grassmann and J. Tavakoli, Optimal policies of M(t)/M/c/c queues with two different levels of servers, *European Journal of Operational Research*, **249** (2016), 1124–1130.

[20] Y. van der Wal and U. Yechiali, Dynamic visit-order rules for batch-service polling, *Probability in the Engineering and Informational Sciences*, **17** (2003), 351–367.

[21] T. Y. Wang, K. H. Wang and W. L. Pearn, Optimization of the T-policy M/G/1 queue with server breakdowns and general start up times, *Journal of Computational and Applied Mathematics*, **228** (2009), 270–278.

[22] Z. G. Zhang and N. Tian, An analysis of queueing systems with multi-task servers, *European Journal of Operational Research*, **156** (2004), 375–389.

[23] X. Zhang, J. Wang and T. V. Do, Threshold properties of the M/M/1 queue under T-policy with applications, *Applied Mathematics and Computation*, **261** (2015), 284–301.

Received April 2016; revised September 2016.

E-mail address: jtao0728@163.com

E-mail address: lwliu@mail.njust.edu.cn