RAIRO-Oper. Res. 51 (2017) 931–944
DOI: 10.1051/ro/2017002

RAIRO Operations Research
www.rairo-ro.org

FLUID QUEUE DRIVEN BY A MULTI-SERVER QUEUE WITH MULTIPLE VACATIONS AND VACATION INTERRUPTION

Senlin yu[1], Zaiming liu[1] and Jinbiao wu[1]

Abstract. This paper studies a fluid queue driven by a multi-server queue with multiple working vacations and vacation interruption. The stationary distribution of the background environment is obtained after some manipulation. A system of differential equations satisfied by the fluid queue is presented, by which we gain the matrix-geometric structure of the Laplace transform of the stationary buffer content. Furthermore, we derive the explicit expression of the mean buffer content. Finally, the numerical example is employed to illustrate our results.

Mathematics Subject Classification. 60K25, 68M20.

Received April 25, 2016. Accepted December 24, 2016.

1. Introduction

In several queueing systems, *e.g.* production lines, the customer-related processes (arrivals and services) typically evolve at a much faster time-scale than the machine-related processes (vacations, inspection times, repairs etc.). It is reasonable to use the fluid queues to analyze these systems. A fluid queue is an input-output system, where the customers are modeled as a continuous fluid that enters and leaves a storage device, called a buffer, according to rates that depend on underlying stochastic process that is related to the state of the machine (on, off, under repair or preventive maintenance etc.). Such fluid queues are motivated as modeling tools of, for example, high-speed communication networks, transportation systems and production-inventory systems. The common characteristic of these application areas is that the customer-units are processed very fast in comparison with changes in the server status. Indeed, the on-off alterations of the independent sources that occur in fluid queues can be seen as vacations/failures of the server, the gate-keeper or the administrator of the system.

The first systematic study of the field can be found in Virtamo and Norros [12], they presented a spectral-decomposition method. Li and Zhao [8] considered a fluid queue driven by a QBD process with either finitely-many levels or infinitely-many levels. The authors proposed the method of the RG-factorization, which is a unified approach for studying fluid queues driven by a Markov chain with block-structured transition property. Mao *et al.* [9] analyzed a fluid model driven by an M/M/1 vacation queue. Xu *et al.* [13] studied the fluid queue driven by an M/M/1 queue with vacation interruption. Recently, Baek *et al.* [4] analyzed the MAP-modulated fluid flow queueing system with multiple vacations. Ammar [2, 3] obtained the analytical expressions of the

Keywords. M/M/c queue, fluid queue, working vacations, buffer content, vacation interruption.

[1] School of Mathematics and Statistics, Central South University, Changsha, Hunan 410083, P.R. China.
yusenlin@csu.edu.cn; math_lzm@csu.edu.cn; wujinbiao@csu.edu

Article published by EDP Sciences

stationary distribution of the buffer content by using modified Bessel functions. Economou and Manou [6] discussed the equilibrium strategies in an observable fluid queue with an alternating service process.

In many situations, the server in the background queueing system may become unavailable for a random period of time to perform a secondary task when the queue becomes empty. A better modeling assumption would be to assume that the server works at a slower rate during vacation periods in comparison to that of a regular working period. Such models are classified as queues subject to working vacations. In some practical situations, the server can stop the vacation due to some reasons, such as the number of customers achieves a certain value in the vacation period. The server may take service in the working vacation period with a lower rate and must return to normal service level at times, or the expense of waiting customers in the vacation period will be very high. The vacation interruption policy is more reasonable to the queueing system with vacations. Li *et al.* [7] studied the equilibrium strategies in the M/M/1 queue with working vacations and vacation interruption.

Many researchers have presented the background process as an alternating restitution process, dealing with the successive idle and busy periods of the driving queue. Markov modulated fluid queues are a particular class of fluid models useful for modeling many physical phenomena and they often allow tractable analysis. Certain interesting real-world applications of Markov modulated fluid queues can be found in Aggarwal *et al.* [1] and Yan and Kularni [14]. More recently, Bosman and Nunez−Queija [5] considered a tandem fluid queue to evaluate the performance of streaming media over an unreliable network.

The conventional fluid queues have been successfully applied to various real-world systems. However, wider applications were restricted due to the requirement that the system (server) has to begin to process the fluid as soon as the fluid level turns to be positive. This limited range of applications due to the lack of server control can be found. In a production-inventory system, the server needs to delay the production until a certain amount of raw material is accumulated. From the engineering point of view, delaying the production is meaningful if the setup cost is very high. By delaying the production, the production cycle becomes longer and it results in a low average setup cost per time unit. Similar example can be found in a computer and communication network, in which the system needs a maintenance period whenever there is no workload to be processed. The study of fluid model driven by vacation queues provides greater flexibility to the design and control of input and output rates of fluid flow thereby adapting the fluid queues to wider application background.

Due to the inherent complexity of real-world systems, the multi-server background environment is more reasonable and practicable. This paper presents an analytical solution for the fluid queue driven by a multi-server queue with multiple working vacations and vacation interruption in stationary regime. When the background queueing system is empty, the servers will take a working vacation synchronously. During the working vacation period, when a service is completed and there are customers in the system, the servers resume to a regular busy period. It is assumed that the fluid in the buffer is increasing when there are customers in the driving queue. Besides, the buffer content decreases at a constant rate δ_0 when the driving queue is empty. The system of equations governing the process is explicitly solved using the matrix-geometric methodology. The stationary distribution of the buffer content is thereby obtained in the Laplace domain. Closed-form expressions help to gain a deeper insight into the model.

A practical example of the fluid queue arises from health care scenario. We consider a medical service system staffed with a chief physician (called main server) and a physician assistant (called substitute server). The physician assistant only provides service to the patients when the chief physician is on vacation (called working vacation) and the service rate of the physician assistant is usually lower than that of the chief physician. When the chief physician finds no customer in the waiting line, he will need to rest from his work, *i.e.*, go on a vacation. During the chief physician is on vacation, the physician assistant will serve the patients, if any, and after his service completion if there are patients in the waiting line, the chief physician will come back from his vacation no matter his vacation ends or not, *i.e.*, vacation interruption happens. Meanwhile, if there is no patient when a working vacation ends, the chief physician begins another vacation, otherwise, the chief physician takes over the physician assistant. After receiving service from the chief physician or the physician assistant, the patients need to be in hospital for some time. For simplicity, we assume that the patients are discharged from hospital at a constant rate. The number of patients in the hospital under continuous review can be viewed as a fluid process that fluctuates according to the state of the physician. The input rate to the buffer is determined by the status of the physician. Such scenario can be modeled as a fluid queue driven by a single-server queue with multiple working vacations and vacation interruption. In addition, there are many physicians in the hospital and thus the multi-server modulating queue is more practicable.

The external environment is a level-dependent QBD process and the method used here is neither based on spectral analysis nor on the use of Bessel functions as done before, we present a direct approach in this paper. This method is attractive due to the fact that it results in the matrix-geometric structure of the Laplace transform of the stationary buffer content. The paper is organized as follows. The stationary distribution of the background environment is given in Section 2. In Section 3, we set up a system of differential equations satisfied by the fluid queue. In Section 4, the Laplace transform of the stationary distribution of the buffer content is obtained. Numerical examples and conclusions are given in Section 5.

2. The background queueing system

We study an M/M/c queue with arrival intensity λ and exponential service of rate μ_b for each of all c servers. The c servers take a working vacation of random length V synchronously when the queue becomes empty, V is exponentially distributed with parameter η. During the working vacation period, customers are served at a lower rate $\mu_v < \mu_b$. In a working vacation, it is assumed that, when a service is completed and there are customers in the system, the vacation is interrupted and the servers resume to a regular busy period. Moreover, if a vacation is completed and there are customers in the system, the c servers begin to serve the customers with rate μ_b and a busy period begins. Otherwise, another working vacation is taken. This policy is called the multiple working vacations and vacation interruption, which has been appropriately analyzed in the literature.

We assume that interarrival times, service times and working vacation times are mutually independent. In addition, the service discipline is first come first served.

The state of the system can be represented by a pair $(L(t), I(t))$, where $L(t)$ denotes the queue length at time t and $I(t)$ represents the state of the system (1: normal working state; 0: working vacation). Clearly, the stochastic process $\{(L(t), I(t)), t \geq 0\}$ is a quasi birth-and-death (QBD) process with state space

$$\Omega = \{(0,0)\} \bigcup \{(k,i), k \geq 1, i = 0,1\}.$$

Using the lexicographical sequence for the states, the q-matrix of the QBD process $\{(L(t), I(t)), t \geq 0\}$ is given by

$$\mathbf{Q} = \begin{pmatrix} -\lambda & \mathbf{C}_{01} & & & & & & \\ \mathbf{B}_{10} & \mathbf{A}_1 & \mathbf{C} & & & & & \\ & \mathbf{B}_2 & \mathbf{A}_2 & \mathbf{C} & & & & \\ & & \ddots & \ddots & \ddots & & & \\ & & & \mathbf{B}_{c-1} & \mathbf{A}_{c-1} & \mathbf{C} & & \\ & & & & \mathbf{B} & \mathbf{A} & \mathbf{C} & \\ & & & & & \ddots & \ddots & \ddots \end{pmatrix},$$

where

$$\mathbf{C}_{01} = (\lambda, 0), \quad \mathbf{B}_{10} = (\mu_v, \mu_b)^T, \quad \mathbf{B}_k = \begin{pmatrix} 0 & k\mu_v \\ 0 & k\mu_b \end{pmatrix}, \ 2 \leq k \leq c-1,$$

$$\mathbf{A}_k = \begin{pmatrix} -(\lambda + k\mu_v + \eta) & \eta \\ 0 & -(\lambda + k\mu_b) \end{pmatrix}, \ 1 \leq k \leq c-1, \quad \mathbf{B} = \begin{pmatrix} 0 & c\mu_v \\ 0 & c\mu_b \end{pmatrix},$$

$$\mathbf{A}=\begin{pmatrix}-(\lambda+c\mu_v+\eta) & \eta\\ 0 & -(\lambda+c\mu_b)\end{pmatrix},\quad \mathbf{C}=\begin{pmatrix}\lambda & 0\\ 0 & \lambda\end{pmatrix}.$$

Theorem 2.1. *If the system workload $\rho=\frac{\lambda}{c\mu_b}<1$, the matrix equation $\boldsymbol{R}^2\boldsymbol{B}+\boldsymbol{R}\boldsymbol{A}+\boldsymbol{C}=\boldsymbol{0}$ has the minimal non-negative solution*

$$\boldsymbol{R}=\begin{pmatrix} r & \dfrac{r(c\mu_v r+\eta)}{c\mu_b(1-r)}\\ 0 & \rho\end{pmatrix},$$

where $r=\frac{\lambda}{\lambda+c\mu_v+\eta}$. The solution $\boldsymbol{R}$ is called the rate matrix.

Proof. Since $\mathbf{B}$, $\mathbf{A}$ and $\mathbf{C}$ are all upper triangular matrices, we can assume that the solution $\mathbf{R}$ has the same structure as

$$\mathbf{R}=\begin{pmatrix}R_{11} & R_{12}\\ 0 & R_{22}\end{pmatrix}.$$

Substitution of $\mathbf{R}$ into the matrix equation $\mathbf{R}^2\mathbf{B}+\mathbf{R}\mathbf{A}+\mathbf{C}=\mathbf{0}$ gives

$$-(\lambda+c\mu_v+\eta)R_{11}+\lambda=0,$$

$$c\mu_b R_{22}^2-(\lambda+c\mu_b)R_{22}+\lambda=0,$$

$$c\mu_v R_{11}^2+c\mu_b(R_{11}R_{12}+R_{12}R_{22})+\eta R_{11}-(\lambda+c\mu_b)R_{12}=0.$$

Clearly, $R_{11}=\frac{\lambda}{\lambda+c\mu_v+\eta}=r$. To get the minimal non-negative solution, we take $R_{22}=\frac{\lambda}{c\mu_b}$ in the second equation. Substituting $R_{11}=r$ and $R_{22}=\frac{\lambda}{c\mu_b}$ into the third equation, the theorem is proved. □

We introduce the following relations:

$$c\mu_v r+\eta=(\lambda+\eta)(1-r),$$

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

After some simplifications, we arrive at

$$\mathbf{RB}+\mathbf{A}=\begin{pmatrix}-(\lambda+c\mu_v+\eta) & \lambda+\eta\\ 0 & -c\mu_b\end{pmatrix}.$$

We focus our attention on the stationary distribution of the process $\{(L(t),I(t)),t\geq 0\}$. Let p_{ki} be the stationary probability of the QBD process $\{(L(t),I(t)),t\geq 0\}$ in state (k,i), that is,

$$p_{ki}=\lim_{t\to\infty}P\{L(t)=k,I(t)=i\},\ (k,i)\in\Omega.$$

Let $\mathbf{x}=(p_{00},p_{10},p_{11},\ldots,p_{c0},p_{c1})$, we need to solve the matrix equation $\mathbf{x}B[\mathbf{R}]=\mathbf{0}$, where

$$B[\mathbf{R}]=\begin{pmatrix}-\lambda & \mathbf{C}_{01} & & & & \\ \mathbf{B}_{10} & \mathbf{A}_1 & \mathbf{C} & & & \\ & \mathbf{B}_2 & \mathbf{A}_2 & \mathbf{C} & & \\ & & \ddots & \ddots & \ddots & \\ & & & \mathbf{B}_{c-1} & \mathbf{A}_{c-1} & \mathbf{C}\\ & & & & \mathbf{B} & \mathbf{RB}+\mathbf{A}\end{pmatrix}.$$

We introduce a series of recursive relations:

$$\varphi_0=1,\qquad \varphi_k=\frac{\lambda+\eta+(c-k)\mu_v}{\lambda}\varphi_{k-1},\ 1\leq k\leq c-1.$$

Theorem 2.2. *If $\rho < 1$, the stationary distribution of the process $\{(L(t), I(t)), t \geq 0\}$ is given by*

$$p_{k0} = \begin{cases} M\varphi_{c-1-k}, \ 0 \leq k \leq c-1, \\ Mr, \ k = c, \\ p_{c0}r^{k-c}, \ k > c, \end{cases}$$

$$p_{k1} = \begin{cases} \dfrac{M\lambda}{\mu_b}\varphi_{c-1} - \dfrac{M\mu_v}{\mu_b}\varphi_{c-2}, \ k = 1, \\ p_{c1}\rho^{k-c} + \dfrac{r(c\mu_v r + \eta)p_{c0}}{c\mu_b(1-r)} \displaystyle\sum_{b=0}^{k-c-1} r^b \rho^{k-c-1-b}, \ k > c, \end{cases}$$

when $2 \leq k \leq c$,

$$p_{k1} = \frac{M}{k!}\Bigg\{ \left(\frac{\lambda}{\mu_b}\right)^{k-1} \left(\frac{\lambda}{\mu_b}\varphi_{c-1} - \frac{\mu_v}{\mu_b}\varphi_{c-2}\right) + \frac{(\lambda+\eta)r}{\lambda} \sum_{j=1}^{k-1} \left(\frac{\lambda}{\mu_b}\right)^j (k-j)!$$
$$+ \frac{\eta}{\lambda}\sum_{j=1}^{k-1} \left(\frac{\lambda}{\mu_b}\right)^j (k-j)! \sum_{v=k+1-j}^{c-1} \varphi_{c-1-v} + \frac{\mu_v}{\lambda}\sum_{j=1}^{k-1} \left(\frac{\lambda}{\mu_b}\right)^j (k-j)! \sum_{b=k+2-j}^{c} b\varphi_{c-1-b}\Bigg\}.$$

Herein $M = p_{c-1,0}$, the constant M can be determined by the normalization condition $\sum_{k=0}^{\infty} p_{k0} + \sum_{k=1}^{\infty} p_{k1} = 1$.

Proof. The matrix equation $\mathbf{x}B[\mathbf{R}] = \mathbf{0}$ can be written as

$$\begin{cases} -\lambda p_{00} + \mu_v p_{10} + \mu_b p_{11} = 0, \\ \eta p_{10} - (\lambda + \mu_b)p_{11} + 2\mu_b p_{21} + 2\mu_v p_{20} = 0, \\ \lambda p_{k-1,0} - (\lambda + \eta + k\mu_v)p_{k0} = 0, \ 1 \leq k \leq c-1, \\ \lambda p_{k-1,1} + \eta p_{k0} - (\lambda + k\mu_b)p_{k1} + (k+1)\mu_b p_{k+1,1} + (k+1)\mu_v p_{k+1,0} = 0, \ 2 \leq k \leq c-1, \\ \lambda p_{c-1,0} - (\lambda + \eta + c\mu_v)p_{c0} = 0, \\ c\mu_b p_{c1} - \lambda p_{c-1,1} = (\lambda + \eta)p_{c0}, \end{cases} \tag{2.1}$$

the six equations are named by (2.1.1)–(2.1.6), respectively.

Let $p_{c-1,0} = M$, from (2.1.5), we have $p_{c0} = rM$. Substitution of this relation into equation (2.1.3) gives

$$p_{k0} = M\varphi_{c-1-k}, \ 0 \leq k \leq c-1.$$

The fact that $p_{11} = \frac{M\lambda}{\mu_b}\varphi_{c-1} - \frac{M\mu_v}{\mu_b}\varphi_{c-2}$ follows from (2.1.1). Combining (2.1.4) and (2.1.6) yields

$$k\mu_b p_{k1} = \lambda p_{k-1,1} + (\lambda + \eta)p_{c0} + \eta \sum_{v=k}^{c-1} p_{v0} + \sum_{b=k+1}^{c} b\mu_v p_{b0}, \ 2 \leq k \leq c-1.$$

After some manipulation, we arrive at

$$p_{k1} = \frac{1}{k!}\Bigg\{ \left(\frac{\lambda}{\mu_b}\right)^{k-1} p_{11} + \frac{(\lambda+\eta)p_{c0}}{\lambda} \sum_{j=1}^{k-1} \left(\frac{\lambda}{\mu_b}\right)^j (k-j)! + \frac{\eta}{\lambda}\sum_{j=1}^{k-1} \left(\frac{\lambda}{\mu_b}\right)^j (k-j)! \sum_{v=k+1-j}^{c-1} p_{v0}$$
$$+ \frac{\mu_v}{\lambda}\sum_{j=1}^{k-1} \left(\frac{\lambda}{\mu_b}\right)^j (k-j)! \sum_{b=k+2-j}^{c} bp_{b0}\Bigg\}, \ 2 \leq k \leq c.$$

When $k > c$, based on matrix-geometric solution method, we have

$$(p_{k0}, p_{k1}) = (p_{c0}, p_{c1})\mathbf{R}^{k-c} = (p_{c0}, p_{c1}) \begin{pmatrix} r^{k-c} & \frac{r(c\mu_v r+\eta)}{c\mu_b(1-r)} \sum\limits_{b=0}^{k-c-1} r^b \rho^{k-c-1-b} \\ 0 & \rho^{k-c} \end{pmatrix}.$$

Finally, the constant $M = p_{c-1,0}$ follows from the normalization condition $\sum_{k=0}^{\infty} p_{k0} + \sum_{k=1}^{\infty} p_{k1} = 1$. □

Clearly, the QBD process $\{(L(t), I(t)), t \geq 0\}$ is stable since the spectral radius of the rate matrix $\mathbf{R}$ satisfies $SP(\mathbf{R}) = \max(r, \rho) < 1$ and the matrix equation $\mathbf{x}B[\mathbf{R}] = \mathbf{0}$ has a positive solution, as shown in Theorems 2.1 and 2.2.

3. Model description

This section deals with the stationary analysis of the fluid queue driven by a multi-server queue with multiple working vacations and vacation interruption. Let $C(t)$ be the buffer content (the amount of fluid in the buffer) at time t, it is a non-negative random variable. We assume that the fluid input rate and the fluid output rate are modulated by the background environment $\{(L(t), I(t)), t \geq 0\}$, that is,

$$\nu[L(t), I(t), C(t)] = \frac{\mathrm{d}C(t)}{\mathrm{d}t} = \begin{cases} \delta, & (L(t), I(t)) = (0,0),\ C(t) > 0, \\ 0, & (L(t), I(t)) = (0,0),\ C(t) = 0, \\ \delta_1, & (L(t), I(t)) = (k,0),\ k \geq 1, \\ \delta_2, & (L(t), I(t)) = (k,1),\ k \geq 1, \end{cases}$$

where $\delta < 0$ and $0 < \delta_1 < \delta_2$. It is assumed that the buffer capacity is infinite. The buffer content is decreasing at a rate of $|\delta|$ when the driving queue is empty, while the buffer content is increasing at a rate of δ_1 (δ_2) when there are customers in the driving system and the driving queue stays in a working vacation period (busy period). The buffer content cannot decrease whenever the buffer is empty.

Since the change of the process $\{C(t), t \geq 0\}$ depends only on its rate, which in turn changes according to the QBD process $\{(L(t), I(t)), t \geq 0\}$, it is clear that $\{(L(t), I(t), C(t)), t \geq 0\}$ is a Markov process.

As the fluid level varies dynamically, it is necessary that the net effective rate of the fluid remains negative to ensure the stability of the process in a long run. If the background queueing process $\{(L(t), I(t)), t \geq 0\}$ is stable, let

$$d = \delta p_{00} + \delta_1 \sum_{k=1}^{\infty} p_{k0} + \delta_2 \sum_{k=1}^{\infty} p_{k1},$$

the quantity d is referred to as the mean drift of the fluid queue. When the buffer is infinite, the stochastic process $\{(L(t), I(t), C(t)), t \geq 0\}$ is stable if the mean drift $d < 0$ and $\rho < 1$. We assume throughout the analysis that these stability conditions are satisfied.

Define the joint probability distribution functions of the process $\{(L(t), I(t), C(t)), t \geq 0\}$ at time t as

$$F_{ki}(t, x) = P\{L(t) = k, I(t) = i, C(t) \leq x\},\ \ x \geq 0,\ \ (k, i) \in \Omega.$$

When the process $\{(L(t), I(t), C(t)), t \geq 0\}$ is stable, its stationary random vector is given by (L, I, C), let

$$F_{ki}(x) = P\{L = k, I = i, C \leq x\} = \lim_{t\to\infty} F_{ki}(t, x),\ \ x \geq 0,\ \ (k, i) \in \Omega.$$

The stationary distribution function of the buffer content is given by

$$F(x) = \lim_{t\to\infty} P(C(t) \leq x) = F_{00}(x) + \sum_{k=1}^{\infty} F_{k0}(x) + \sum_{k=1}^{\infty} F_{k1}(x).$$

To simplify matters, we introduce the vectors $F_k(x) = (F_{k0}(x), F_{k1}(x))$, $k \geq 1$. As shown in Mitra [10], the stationary joint distribution functions of the fluid buffer content satisfy a set of differential equations. Using the standard probability arguments, we get the following system of differential equations:

$$\begin{cases} \delta \dfrac{\mathrm{d}F_{00}(x)}{\mathrm{d}x} = -\lambda F_{00}(x) + \mu_v F_{10}(x) + \mu_b F_{11}(x), \\ \delta_1 \dfrac{\mathrm{d}F_{k0}(x)}{\mathrm{d}x} = -(\lambda + k\mu_v + \eta)F_{k0}(x) + \lambda F_{k-1,0}(x), \ 1 \leq k \leq c-1, \\ \delta_2 \dfrac{\mathrm{d}F_{11}(x)}{\mathrm{d}x} = -(\lambda + \mu_b)F_{11}(x) + \eta F_{10}(x) + 2\mu_b F_{21}(x) + 2\mu_v F_{20}(x), \\ \delta_2 \dfrac{\mathrm{d}F_{k1}(x)}{\mathrm{d}x} = -(\lambda + k\mu_b)F_{k1}(x) + \lambda F_{k-1,1}(x) + \eta F_{k0}(x) + m(x), \ 2 \leq k \leq c-1, \\ \delta_1 \dfrac{\mathrm{d}F_{k0}(x)}{\mathrm{d}x} = -(\lambda + c\mu_v + \eta)F_{k0}(x) + \lambda F_{k-1,0}(x), \ k \geq c, \\ \delta_2 \dfrac{\mathrm{d}F_{k1}(x)}{\mathrm{d}x} = -(\lambda + c\mu_b)F_{k1}(x) + \lambda F_{k-1,1}(x) + \eta F_{k0}(x) + c\mu_b F_{k+1,1}(x) + c\mu_v F_{k+1,0}(x), \ k \geq c, \end{cases}$$

where $m(x) = (k+1)\mu_b F_{k+1,1}(x) + (k+1)\mu_v F_{k+1,0}(x)$.

The boundary conditions are given by

$$F_{00}(0) = a, \qquad F_{ki}(0) = 0, \ k \geq 1, \ i = 0, 1,$$

where a is referred to as the stationary probability of empty buffer content, the value of a will be given later. Since we make an assumption that the content of the buffer is increasing when there are customers in the background queueing system, it is impossible to have the buffer empty when the modulating process is in any of the states other than $(0, 0)$. However, when the background queueing system is empty, the buffer content decreases at rate δ and hence with some positive probability, it is possible that the content of the buffer is empty. Therefore the boundary conditions are valid.

These differential equations are equivalent to the following matrix equation:

$$\frac{\mathrm{d}}{\mathrm{d}x}(F_{00}(x), F_1(x), F_2(x), \ldots)\mathbf{H} = (F_{00}(x), F_1(x), F_2(x), \ldots)\mathbf{Q}, \tag{3.1}$$

where $\mathbf{H} = \mathrm{diag}(\delta, \delta_1, \delta_2, \delta_1, \delta_2, \ldots)$, $\mathbf{Q}$ is the infinitesimal generator of the QBD process $\{(L(t), I(t)), t \geq 0\}$.

Clearly, these differential equations are resistant to analytical solutions. Let us now examine things from another angle, we define the Laplace transform of the functions $F_{ki}(x)$ as

$$\hat{F}_{ki}(s) = \int_0^{+\infty} \mathrm{e}^{-sx} F_{ki}(x)\mathrm{d}x, \ s \geq 0, \ (k, i) \in \Omega,$$

and $\hat{F}_k(s) = (\hat{F}_{k0}(s), \hat{F}_{k1}(s))$, $k \geq 1$. Since the joint distribution functions $F_{ki}(x)$ $((k, i) \in \Omega)$ are non-negative, the variable s is real and $s \geq 0$, see Rozovsky [11]. The use of Laplace transforms for real arguments is very common in Applied Probability.

The Laplace transform of the stationary distribution of the buffer content is given by

$$\hat{F}(s) = \int_0^{+\infty} \mathrm{e}^{-sx} F(x)\mathrm{d}x, \ s \geq 0.$$

Taking the Laplace transform on both sides of (3.1), then

$$(\hat{F}_{00}(s), \hat{F}_1(s), \hat{F}_2(s), \ldots)(\mathbf{Q} - s\mathbf{H}) = (-a\delta, 0, 0, \ldots), \tag{3.2}$$

the right hand side of (3.2) comes from the boundary conditions.

4. Stationary analysis

The stationary distribution of the fluid queue plays an important role in our analysis. In this section, the expressions for the Laplace transform of the joint stationary distribution of the fluid queue are proved to be of matrix factorial form. Based on this fact, we can get the analytical expression for the Laplace-Stieltjes transform of the stationary distribution of the buffer content.

For $s \geq 0$, we introduce the matrix

$$\mathbf{A}(s) = \begin{pmatrix} -(\lambda + c\mu_v + \eta + s\delta_1) & \eta \\ 0 & -(\lambda + c\mu_b + s\delta_2) \end{pmatrix}.$$

Theorem 4.1. *For $s \geq 0$, the matrix equation*

$$\boldsymbol{R}^2(s)\boldsymbol{B} + \boldsymbol{R}(s)\boldsymbol{A}(s) + \boldsymbol{C} = \boldsymbol{0}$$

has the minimal non-negative solution

$$\boldsymbol{R}(s) = \begin{pmatrix} r(s) & \dfrac{r(s)(c\mu_v r(s) + \eta)}{c\mu_b(h_1(s) - r(s))} \\ 0 & h(s) \end{pmatrix},$$

the expressions of $r(s)$, $h(s)$ and $h_1(s)$ will be given in the proof.

Proof. The proof is similar to that exhibited in Theorem 2.1. Since $\mathbf{B}$, $\mathbf{A}(s)$ and $\mathbf{C}$ are all upper triangular matrices, we can assume that the solution $\mathbf{R}(s)$ has the same structure as

$$\mathbf{R}(s) = \begin{pmatrix} R_{11}(s) & R_{12}(s) \\ 0 & R_{22}(s) \end{pmatrix}.$$

In an analogous fashion, we find

$$-(\lambda + c\mu_v + \eta + s\delta_1)R_{11}(s) + \lambda = 0,$$

$$c\mu_v R_{11}^2(s) + c\mu_b(R_{11}(s)R_{12}(s) + R_{12}(s)R_{22}(s)) + \eta R_{11}(s) - (\lambda + c\mu_b + s\delta_2)R_{12}(s) = 0, \tag{4.1}$$

$$c\mu_b R_{22}^2(s) - (\lambda + c\mu_b + s\delta_2)R_{22}(s) + \lambda = 0. \tag{4.2}$$

It is clear that $R_{11}(s) = r(s) = \frac{\lambda}{\lambda + c\mu_v + \eta + s\delta_1}$ and $0 < r(s) < 1$. The quadratic equation (4.2) possesses two roots $h(s)$ and $h_1(s)$:

$$h(s)(h_1(s)) = \frac{1}{2c\mu_b}\left(\lambda + c\mu_b + s\delta_2 \mp \sqrt{(\lambda + c\mu_b + s\delta_2)^2 - 4c\mu_b\lambda}\right).$$

Since $\rho = \frac{\lambda}{c\mu_b} < 1$, then $(c\mu_b - \lambda + s\delta_2)^2 < (\lambda + c\mu_b + s\delta_2)^2 - 4c\mu_b\lambda < (\lambda + c\mu_b + s\delta_2)^2$, $s \geq 0$. Substituting this inequality into the expressions of $h(s)$ and $h_1(s)$, we find $0 < h(s) < 1$ and $h_1(s) \geq 1$, $s \geq 0$. In addition, $h(0) = \rho$, $h_1(0) = 1$ and $h(s) < h_1(s)$.

To get the minimal non-negative solution, we take the minimal solution to the equation (4.2), that is, $R_{22}(s) = h(s)$. Note that $h(s) + h_1(s) = \frac{\lambda + c\mu_b + s\delta_2}{c\mu_b}$, we can get the expression of $R_{12}(s)$ through setting $R_{11}(s) = r(s)$ and $R_{22}(s) = h(s)$ in (4.1), then

$$R_{12}(s) = \frac{r(s)(c\mu_v r(s) + \eta)}{c\mu_b(h_1(s) - r(s))}.$$

This completes the proof. Note that $\mathbf{R}(0) = \mathbf{R}$, the root $\mathbf{R}(s)$ plays an important role in the following analysis. □

In the following theorem, we will show that the Laplace transform of the joint stationary distribution $\{\hat{F}_k(s), k \geq c\}$ are of matrix factorial form. We introduce a series of recursive relations:

$$\psi_0(s) = 1, \quad \psi_k(s) = \frac{\lambda + \eta + (c-k)\mu_v + \delta_1 s}{\lambda}\psi_{k-1}(s),\ 1 \leq k \leq c-1.$$

Theorem 4.2. *If $d < 0$ and $\rho < 1$, then the expressions of $\hat{F}_{00}(s)$ and $\{\hat{F}_k(s), k \geq 1\}$ are satisfied with the following relations*

$$\begin{cases} \hat{F}_{k0}(s) = M_0(s)\psi_{c-1-k}(s), \quad 0 \leq k \leq c-1, \\ (\hat{F}_{k0}(s), \hat{F}_{k1}(s)) = (M_0(s), M_1(s))\boldsymbol{R}^{k-c+1}(s), \quad k \geq c, \end{cases} \tag{4.3}$$

where $(M_0(s), M_1(s)) = (\hat{F}_{c-1,0}(s), \hat{F}_{c-1,1}(s))$. Moreover, $\{\hat{F}_{k1}(s), 1 \leq k \leq c-2\}$ can be expressed by $M_0(s)$ and $M_1(s)$, the related computation is omitted since we only need to calculate $\sum_{k=1}^{c-2}\hat{F}_{k1}(s)$.

Proof. Clearly, the matrix equation (3.2) can be equivalently written as the following system of equations:

$$\begin{cases} -(\lambda + s\delta)\hat{F}_{00}(s) + \mu_v\hat{F}_{10}(s) + \mu_b\hat{F}_{11}(s) = -a\delta, \\ \eta\hat{F}_{10}(s) - (\lambda + \mu_b + \delta_2 s)\hat{F}_{11}(s) + 2\mu_b\hat{F}_{21}(s) + 2\mu_v\hat{F}_{20}(s) = 0, \\ \lambda\hat{F}_{k-1,0}(s) - (\lambda + k\mu_v + \eta + \delta_1 s)\hat{F}_{k0}(s) = 0,\ 1 \leq k \leq c-1, \\ \lambda\hat{F}_{k-1,1}(s) + \eta\hat{F}_{k0}(s) - (\lambda + k\mu_b + \delta_2 s)\hat{F}_{k1}(s) + m(s) = 0,\ 2 \leq k \leq c-1, \\ (\hat{F}_{k-1,0}(s), \hat{F}_{k-1,1}(s))\mathbf{C} + (\hat{F}_{k0}(s), \hat{F}_{k1}(s))\mathbf{A}(s) + (\hat{F}_{k+1,0}(s), \hat{F}_{k+1,1}(s))\mathbf{B} = (0,0),\ k \geq c, \end{cases} \tag{4.4}$$

where $m(s) = (k+1)\mu_b\hat{F}_{k+1,1}(s) + (k+1)\mu_v\hat{F}_{k+1,0}(s)$. The five equations in (4.4) are named by (4.4.1)−(4.4.5), respectively.

The stochastic process $\{(L(t), I(t), C(t)), t \geq 0\}$ has a unique joint stationary distribution $\{F_{ki}(x), (k,i) \in \Omega\}$, thus there exists a unique solution to the above system of equations. We only need to verify that the expressions in Theorem 4.2 are satisfied with (4.4).

For $k \geq c$, substitution of $(\hat{F}_{k0}(s), \hat{F}_{k1}(s)) = (\hat{F}_{c-1,0}(s), \hat{F}_{c-1,1}(s))\mathbf{R}^{k-c+1}(s)$ into (4.4.5) gives

$$\begin{aligned} &(\hat{F}_{k-1,0}(s), \hat{F}_{k-1,1}(s))\mathbf{C} + (\hat{F}_{k0}(s), \hat{F}_{k1}(s))\mathbf{A}(s) + (\hat{F}_{k+1,0}(s), \hat{F}_{k+1,1}(s))\mathbf{B} \\ &= (\hat{F}_{c-1,0}(s), \hat{F}_{c-1,1}(s))\mathbf{R}^{k-c}(s)[\mathbf{R}^2(s)\mathbf{B} + \mathbf{R}(s)\mathbf{A}(s) + \mathbf{C}] = (0,0). \end{aligned}$$

Let $(\hat{F}_{c-1,0}(s), \hat{F}_{c-1,1}(s)) = (M_0(s), M_1(s))$, we find

$$(\hat{F}_{k0}(s), \hat{F}_{k1}(s)) = (M_0(s), M_1(s))\mathbf{R}^{k-c+1}(s).$$

When $k = c$, the second branch of (4.3) reduces to

$$\hat{F}_{c0}(s) = M_0(s)r(s), \quad \hat{F}_{c1}(s) = M_0(s)\frac{r(s)(c\mu_v r(s) + \eta)}{c\mu_b(h_1(s) - r(s))} + M_1(s)h(s). \tag{4.5}$$

From (4.4.3), we can express $\hat{F}_{k0}(s)$ $(0 \leq k \leq c-1)$ in terms of $\hat{F}_{c-1,0}(s)$, that is,

$$\hat{F}_{k0}(s) = M_0(s)\psi_{c-1-k}(s),\ 0 \leq k \leq c-1.$$

In addition, we do not pursue the expressions of $\{\hat{F}_{k1}(s),\ 1 \leq k \leq c-2\}$ since we only need to calculate $\sum_{k=1}^{c-2}\hat{F}_{k1}(s)$. The expressions of $M_0(s)$ and $M_1(s)$ can be derived from equations (4.4.1) and (4.4.2). □

The computation is fairly delicate and lengthy. To get the Laplace transform of the stationary buffer content, it only remains to calculate $\sum_{k=1}^{c-2}\hat{F}_{k1}(s)$. Taking sum on both sides of (4.4.4) from 2 to $c-1$, then

$$\lambda\sum_{k=2}^{c-1}\hat{F}_{k-1,1}(s)+\eta\sum_{k=2}^{c-1}\hat{F}_{k0}(s)-\sum_{k=2}^{c-1}(\lambda+k\mu_b+\delta_2 s)\hat{F}_{k1}(s)+\sum_{k=2}^{c-1}(k+1)\mu_b\hat{F}_{k+1,1}(s)+\sum_{k=2}^{c-1}(k+1)\mu_v\hat{F}_{k+1,0}(s)=0.$$

After some manipulation, we find

$$\eta\sum_{k=2}^{c-1}\hat{F}_{k0}(s)+\sum_{k=3}^{c}k\mu_v\hat{F}_{k0}(s)-\delta_2 s\sum_{k=2}^{c-1}\hat{F}_{k1}(s)+\lambda\hat{F}_{11}(s)-\lambda\hat{F}_{c-1,1}(s)-2\mu_b\hat{F}_{21}(s)+c\mu_b\hat{F}_{c1}(s)=0. \tag{4.6}$$

Summing up (4.6) and (4.4.2), we can get the expression of $\sum_{k=1}^{c-1}\hat{F}_{k1}(s)$, that is,

$$\delta_2 s\sum_{k=1}^{c-1}\hat{F}_{k1}(s)=\eta\sum_{k=1}^{c-1}\hat{F}_{k0}(s)+\sum_{k=2}^{c}k\mu_v\hat{F}_{k0}(s)-\mu_b\hat{F}_{11}(s)-\lambda\hat{F}_{c-1,1}(s)+c\mu_b\hat{F}_{c1}(s).$$

Clearly, $\hat{F}_{11}(s)$ can be eliminated from (4.4.1), that is,

$$\hat{F}_{11}(s)=\frac{-a\delta+(\lambda+s\delta)M_0(s)\psi_{c-1}(s)-\mu_v M_0(s)\psi_{c-2}(s)}{\mu_b}\cdot$$

Note that $\hat{F}_{c0}(s)=M_0(s)r(s)$, the expression of $\sum_{k=1}^{c-2}\hat{F}_{k1}(s)$ is given by

$$\begin{aligned}\sum_{k=1}^{c-2}\hat{F}_{k1}(s)=\frac{M_0(s)}{\delta_2 s}\Bigg[&\eta\sum_{k=1}^{c-1}\psi_{c-1-k}(s)-(\lambda+s\delta)\psi_{c-1}(s)+\mu_v\psi_{c-2}(s)+\frac{r(s)(c\mu_v r(s)+\eta)}{h_1(s)-r(s)}\\&+\sum_{k=2}^{c-1}k\mu_v\psi_{c-1-k}(s)+c\mu_v r(s)\Bigg]+\frac{M_1(s)[c\mu_b h(s)-(\lambda+\delta_2 s)]}{\delta_2 s}+\frac{a\delta}{\delta_2 s}\cdot\end{aligned}$$

We continue by examining the Laplace transform of the stationary buffer content, based on the total probability formula, we have

$$\begin{aligned}\hat{F}(s)&=\int_0^{+\infty}\mathrm{e}^{-sx}F(x)\mathrm{d}x=\sum_{k=0}^{c-2}\hat{F}_{k0}(s)+\sum_{k=1}^{c-2}\hat{F}_{k1}(s)+\sum_{k=c-1}^{\infty}(\hat{F}_{k0}(s),\hat{F}_{k1}(s))\mathbf{e}\\&=M_0(s)\sum_{k=0}^{c-2}\psi_{c-1-k}(s)+\sum_{k=1}^{c-2}\hat{F}_{k1}(s)+(M_0(s),M_1(s))(\mathbf{I}-\mathbf{R}(s))^{-1}\mathbf{e}.\end{aligned}$$

We have arrived at a critical point. Note that $\mathbf{R}(s)$ is an upper triangular matrix and the spectral radius $SP[\mathbf{R}(s)]=\max(r(s),h(s))<1$, thus the matrix $\mathbf{I}-\mathbf{R}(s)$ is invertible, in addition,

$$(\mathbf{I}-\mathbf{R}(s))^{-1}=\begin{pmatrix}\dfrac{1}{1-r(s)} & \dfrac{r(s)(c\mu_v r(s)+\eta)}{c\mu_b(h_1(s)-r(s))(1-r(s))(1-h(s))}\\ 0 & \frac{1}{1-h(s)}\end{pmatrix}.$$

After some simplifications, the Laplace transform of the stationary buffer content is given by

$$
\begin{aligned}
\hat{F}(s) ={}& M_0(s)\sum_{k=0}^{c-2}\psi_{c-1-k}(s) + \frac{M_0(s)}{1-r(s)} + \frac{M_0(s)r(s)(c\mu_v r(s)+\eta)}{c\mu_b(h_1(s)-r(s))(1-r(s))(1-h(s))} + \frac{a\delta}{\delta_2 s}\\
&+ \frac{M_0(s)}{\delta_2 s}\Bigg[\eta\sum_{k=1}^{c-1}\psi_{c-1-k}(s) - (\lambda+\delta s)\psi_{c-1}(s) + \mu_v\psi_{c-2}(s) + \frac{r(s)(c\mu_v r(s)+\eta)}{h_1(s)-r(s)}\\
&+ \sum_{k=2}^{c-1}k\mu_v\psi_{c-1-k}(s) + c\mu_v r(s)\Bigg] + M_1(s)\left[\frac{c\mu_b h(s)-\lambda}{\delta_2 s} + \frac{h(s)}{1-h(s)}\right]\cdot
\end{aligned}
$$

We are in a position to get the expectation of the stationary buffer content. To achieve this goal, we define the Laplace-Stieltjes transform of the joint probability distribution function $F_{ki}(x)$ as

$$f^*_{ki}(s) = \int_0^{+\infty} \mathrm{e}^{-sx}\mathrm{d}F_{ki}(x),\ \ (k,i)\in\Omega.$$

Let $f^*(s)$ be the Laplace-Stieltjes transform of the stationary distribution of the buffer content, it follows that

$$f^*(s) = \int_0^{+\infty} \mathrm{e}^{-sx}\mathrm{d}F(x) = s\hat{F}(s).$$

We can now evaluate the explicit expression of a from the normalization condition $\lim\limits_{s\to 0} f^*(s) = \lim\limits_{s\to 0} s\hat{F}(s) = 1$, the constant $a = F_{00}(0)$ is called the stationary probability of empty buffer content, then

$$a = \frac{\delta_2}{\delta} + \frac{M_0(0)}{\delta}\left[\lambda\psi_{c-1}(0) - \mu_b\psi_{c-2}(0) - \eta\sum_{k=0}^{c-2}\psi_k(0) - \sum_{k=2}^{c-1}k\mu_v\psi_{c-1-k}(0) - \frac{c\mu_v r + r\eta}{1-r}\right], \tag{4.7}$$

where $M_0(0) = M_0(s)|_{s=0}$ and $\psi_k(0) = \psi_k(s)|_{s=0}$, $1 \le k \le c-1$.

It remains to calculate $E(C)$, the expectation of the stationary buffer content. From Theorem 4.1, we have

$$h(s)+h_1(s) = \frac{\lambda + c\mu_b + \delta_2 s}{c\mu_b},\quad h(s)h_1(s) = \frac{\lambda}{c\mu_b},\quad r(s) = \frac{\lambda}{\lambda + c\mu_v + \eta + s\delta_1}\cdot$$

Note that $h(0) = \rho$ and $h_1(0) = 1$, then

$$h'(0) = \frac{-\delta_2\rho}{c\mu_b(1-\rho)},\ \ h_1'(0) = \frac{\delta_2}{c\mu_b(1-\rho)},\ \ r'(0) = \frac{-\delta_1 r^2}{\lambda}\cdot$$

It is well known that $E(C) = -\lim\limits_{s\to 0}\frac{\mathrm{d}}{\mathrm{d}s}f^*(s) = -\lim\limits_{s\to 0}\frac{\mathrm{d}}{\mathrm{d}s}s\hat{F}(s)$, after some manipulation, we eventually arrive at

$$
\begin{aligned}
E(C) ={}& \frac{M_0(0)}{\delta_2}\Bigg[\delta\psi_{c-1}(0) + \lambda\psi'_{c-1}(0) - \mu_v\psi'_{c-2}(0) - \delta_2\sum_{k=1}^{c-1}\psi_k(0) - \sum_{k=2}^{c-1}k\mu_v\psi'_{c-1-k}(0) - c\mu_v r'(0)\\
&- \frac{r'(0)(1-r)(\eta+2rc\mu_v) - r(\eta+c\mu_v r)(h_1'(0)-r'(0))}{(1-r)^2} - \frac{\delta_2}{1-r} - \frac{\delta_2 r(c\mu_v r+\eta)}{(c\mu_b-\lambda)(1-r)^2}\\
&- \eta\sum_{k=0}^{c-2}\psi'_k(0)\Bigg] + \frac{M_0'(0)}{\delta_2}\Bigg[\lambda\psi_{c-1}(0) - \mu_v\psi_{c-2}(0) - \eta\sum_{k=0}^{c-2}\psi_k(0) - \sum_{k=2}^{c-1}k\mu_v\psi_{c-1-k}(0) - c\mu_v r\\
&- \frac{r(\eta+c\mu_v r)}{1-r}\Bigg] - M_1(0)\Bigg[\frac{c\mu_b h'(0)}{\delta_2} + \frac{\rho}{1-\rho}\Bigg],
\end{aligned}
$$

where $M_i'(0) = M_i'(s)|_{s=0}$, $i = 1, 2$, $\psi_k'(0) = \psi_k'(s)|_{s=0}$, $1 \le k \le c-1$.

The present paper aims to exhibit the art of computation. Comparing to the classical single-server driving queues discussed in the literature, the computations become much more involved if the underlying stochastic environment is a multi-server vacation queue. The multi-server driving queue is an essential generalization and thus these results are not surprising.

5. Numerical examples and conclusions

In the fluid queues, the fluid input rate and the fluid output rate depend on the background queueing process that is related to the state of the server (working vacations, under repair and preventive maintenance etc.). Consider a production-inventory model operating in a random environment. The inventory level increases when the production rate exceeds the demand rate and decreases otherwise. The inventory level under continuous review can be viewed as a fluid process that fluctuates according to the evolution of the underlying background environment. For example, consider a machine shop with c servers. When the servers are busy, items are produced continuously at a rate r for each server and if they are idle, there is no production. However, for all practical reasons, the servers might decide to take a working vacation for a random period of time to perform a secondary task. The servers can be substituted by laymen to meet the ongoing demands during the working vacation period. By adopting the working vacation policy, the production cycle becomes larger and it results in a low average setup cost per time unit. During the working vacation period, items are produced continuously at a rate r' for each server and $r' < r$. Further, by offering service at a reduced rate, the servers may interrupt the vacation and continue the busy period due to certain unforeseen reasons, like a sudden increase in the demand. The demand rate is assumed to vary from time to time at the rate δ_t independent of the state of the underlying queueing process. The level of inventory thus oscillates among $cr - \delta_t$, $cr' - \delta_t$ and $-\delta_t$ depending on the state of the background queueing system. This model reflects situations in which the production and demand rates undergo recurring changes in a stochastic fashion, and can be modeled as Markovian. Such scenario can be approximated by the fluid queue driven by a multi-server queue with multiple working vacations and vacation interruption.

Fluid queues are quite useful as approximate models for certain queueing and inventory systems where the flow consists of discrete entities, but the individual units of traffic have less impact on the performance of the system. We present a numerical example to illustrate our results.

We consider a set of parameters as follows: $\lambda = 1$, $c = 3$, $\mu_b = 1$, $\mu_v = 0.2$, $\eta = 0.2$, $\delta_2 = 1$, $\delta_1 = 0.5$ and $\delta = -4$. Note that the stability conditions must be satisfied, *i.e.*, $\rho < 1$ and $d < 0$. Clearly, we have $\psi_1(s) = 1.6 + 0.5s$, $\psi_2(s) = 2.24 + 1.5s + 0.25s^2$, $h(s) = \frac{2}{3} + \frac{1}{6}s - \frac{1}{6}\sqrt{4 + 8s + s^2}$ and $r(s) = \frac{10}{18+5s}$. Thus $h(0) = \frac{1}{3}$, $h'(0) = -\frac{1}{6}$, $h_1'(0) = \frac{1}{2}$ and $r'(0) = -\frac{25}{162}$.

From Theorem 4.2, we have $\hat{F}_{00}(s) = (2.24+1.5s+0.25s^2)M_0(s)$, $\hat{F}_{10}(s) = (1.6+0.5s)M_0(s)$, $\hat{F}_{20}(s) = M_0(s)$, $\hat{F}_{30}(s) = \frac{10}{18+5s}M_0(s)$ and $\hat{F}_{21}(s) = M_1(s)$. Based on equation (4.5), we have

$$\hat{F}_{31}(s) = \frac{(192 + 20s)M_0(s)}{(18 + 5s)^2[4 + s + \sqrt{4 + 8s + s^2}] - 1280 - 300s} + M_1(s)h(s).$$

Similarly, from (4.4.4), we find

$$\hat{F}_{11}(s) = \left(1 + \frac{1}{2}s + \frac{1}{2}\sqrt{4 + 8s + s^2}\right) M_1(s) - \left(0.2 + \frac{6}{18 + 5s} + \frac{576 + 60s}{f(s)}\right) M_0(s),$$

where $f(s) = (18 + 5s)^2[4 + s + \sqrt{4 + 8s + s^2}] - 1280 - 300s$.

Now, we present the expression of $\hat{F}'_{11}(s)$,

$$\begin{aligned}\hat{F}'_{11}(s) = &-\left[\frac{-30}{(18+5s)^2} + \frac{60f(s)-(576+60s)g(s)}{f^2(s)}\right]M_0(s) + \left[\frac{1}{2} + \frac{8+2s}{4\sqrt{4+8s+s^2}}\right]M_1(s) \\ &-\left[0.2 + \frac{6}{18+5s} + \frac{576+60s}{f(s)}\right]M'_0(s) + \left[1 + \frac{1}{2}s + \frac{1}{2}\sqrt{4+8s+s^2}\right]M'_1(s),\end{aligned}$$

where

$$g(s) = (180+50s)\left(4+s+\sqrt{4+8s+s^2}\right) + (18+5s)^2\left(1+\frac{8+2s}{2\sqrt{4+8s+s^2}}\right) - 300.$$

The expressions of $M_0(s)$ and $M_1(s)$ can be derived from equations (4.4.1) and (4.4.2), then

$$(4s-1)\hat{F}_{00}(s) + 0.2\hat{F}_{10}(s) + \hat{F}_{11}(s) = 4a, \tag{5.1}$$

$$0.2\hat{F}_{10}(s) - (2+s)\hat{F}_{11}(s) + 2\hat{F}_{21}(s) + 0.4\hat{F}_{20}(s) = 0. \tag{5.2}$$

We do not pursue the expressions of $M_0(s)$ and $M_1(s)$ since we only need to calculate $M_i(0)$ and $M'_i(0)$, $i = 0, 1$.

For equations (5.1) and (5.2), by letting $s = 0$, we obtain

$$-3.12M_0(0) + 2M_1(0) = 4a, \quad 3.66M_0(0) - 2M_1(0) = 0.$$

Hence, $M_0(0) = 7.4074a$ and $M_1(0) = 13.5556a$, where a is the stationary probability of empty buffer content. Substitution of $M_0(0)$ into (4.7) gives $a = 0.1824$. Thus $M_0(0) = 1.3514$ and $M_1(0) = 2.4730$.

In an analogous fashion, by differentiating equations (5.1) and (5.2) with respect to s and substituting $s = 0$, we arrive at

$$\begin{gathered}4\hat{F}_{00}(0) - \hat{F}'_{00}(0) + 0.2\hat{F}'_{10}(0) + F'_{11}(0) = 0, \\ 0.2\hat{F}'_{10}(0) - \hat{F}_{11}(0) - 2\hat{F}'_{11}(0) + 2\hat{F}'_{21}(0) + 0.4\hat{F}'_{20}(0) = 0.\end{gathered}$$

That is,

$$\begin{gathered}8.935M_0(0) + 1.5M_1(0) - 3.44M'_0(0) + 2M'_1(0) = 0, \\ -1.45M_0(0) - 5M_1(0) + 2.72M'_0(0) - 2M'_1(0) = 0,\end{gathered}$$

we find $M'_0(0) = 2.0273$ and $M'_1(0) = -4.4054$.

The expressions of $\hat{F}(s)$, $(\mathbf{I} - \mathbf{R}(s))^{-1}$ and $f^*(s)$ are not given here. Finally, after some manipulation, we eventually arrive at $E(C) = 15.4776$.

It is clear that the numerical computation of $E(C)$ is rather complicated when c is large. We do not pursue the effect of several system parameters on the mean buffer content due to its complexity. If the background environment is a multi-server vacation queue, the computations become much more involved.

Markov modulated fluid queues are a class of fluid models wherein the rates at which the content of the fluid buffer varies are modulated by the background Markov process. This paper studies a fluid queue driven by a multi-server queue with multiple working vacations and vacation interruption. Comparing to the classical fluid model driven by a single-server queueing system, the multi-server driving queue is an essential generalization. The differential equations satisfied by the fluid queue are presented, by which we gain the matrix-geometric structure of the Laplace transform of the stationary buffer content. Closed-form solutions help to gain a deeper insight into the model. The expectation of the stationary buffer content is given. Finally, we present a numerical example to illustrate our results.

Acknowledgements. The authors thank the reviewers for their valuable comments and acknowledge that this research was supported by the National Natural Science Foundation of China (No. 11671404) and the Fundamental Research Funds for the Central Universities of Central South University (No. 2016zzts014).

References

[1] V. Aggarwal, N. Gautam, S.R.T. Kumara and M. Greaves, Stochastic fluid flow models for determining optimal switching thresholds. *Perform. Eval.* **59** (2005) 19–46.

[2] S.I. Ammar, Fluid queue driven by an M/M/1 disasters queue. *Int. J. Comput. Math.* **91** (2014) 1497–1506.

[3] S.I. Ammar, Analysis of an M/M/1 driven fluid queue with multiple exponential vacations. *Appl. Math. Comput.* **227** (2014) 329–334.

[4] J.W. Baek, H.W. Lee, S.W. Lee and S. Ahn, A MAP-modulated fluid flow model with multiple vacations. *Ann. Oper. Res.* **202** (2013) 19–34.

[5] J.W. Bosman and R. Numez-Queija, A spectral theory approach for extreme value analysis in a tandem of fluid queues. *Queueing Syst.* **78** (2014) 121–154.

[6] A. Economou and A. Manou, Strategic behavior in an observable fluid queue with an alternating service process. *Eur. J. Oper. Res.* **254** (2016) 148–160.

[7] K. Li, J. Wang and Y. Ren, Equilibrium joining strategies in M/M/1 queues with working vacation and vacation interruptions. *RAIRO-Oper. Res.* **50** (2016) 451–471.

[8] Q. Li and Y. Zhao, Block-structured fluid queues driven by QBD processes. *Stoch. Anal. Appl.* **23** (2005) 1087–1112.

[9] B. Mao, F. Wang and N. Tian, Fluid model driven by an M/G/1 queue with multiple exponential vacations. *Appl. Math. Comput.* **218** (2011) 4041–4048.

[10] D. Mitra, Stochastic theory of a fluid model of producers and consumers coupled by a buffer. *Adv. Appl. Probab.* **20** (1988) 646–676.

[11] L. Rozovsky, Remarks on a link between the Laplace transform and distribution function of a non-negative random variable. *Stat. Probabil. Lett.* **79** (2009) 1501–1508.

[12] J. Virtamo and I. Norros, Fluid queue driven by an M/M/1 queue. *Queueing Syst.* **16** (1994) 373–386.

[13] X. Xu, H. Guo, Y. Zhao and J. Geng, The fluid model driven by the M/M/1 queue with working vacations and vacation interruption. *J. Comput. Inf. Syst.* **18** (2012) 4041–4048.

[14] K. Yan and V.G. Kularni, Optimal inventory policies under stochastic production and demand rates. *Stoch. Models* **24** (2008) 173–190.