Double-Sided Queues and their Applications to Vaccine Inventory Management

Haoran Wu[1], Qi-Ming He[†], and Fatih Safa Erenay[†]

[1]School of Business, Sun Yat-sen University

[†]Department of Management Science and Engineering, University of Waterloo

Abstract

We consider a double-sided queueing model with batch Markovian arrival processes and finite discrete abandonment times, which arises in various stochastic systems such as perishable inventory systems and financial markets. Customers arrive at the system with a batch of orders to be matched by counterparts. While waiting to be matched, customers become impatient and may abandon the system without service. The abandonment time of a customer depends on its batch size and its position in the queue. First, we propose an approach to obtain the stationary joint distribution of age processes via the stationary analysis of a multi-layer Markov modulated fluid flow ($MMFF$) process. Second, using the stationary joint distribution of the age processes, we derive a number of queueing quantities related to matching rates, fill rates, sojourn times and queue length for both sides of the system. Last, we apply our model to analyze a vaccine inventory system and gain insight into the effect of uncertainty in supply and demand processes on the performance of the inventory system. It is observed that batch Markovian arrival processes are better choices for modelling the supply/demand process in systems with high uncertainty for more accurate performance quantities.

Keywords: Double-sided queues, Abandonment, Batch Markovian arrival process, Markov modulated fluid flow process, Vaccine inventory system

1 Introduction

Inspired by the perishable inventory system and crossing networks trading system, we develop a double-sided queueing model with batch Markovian arrivals and customer abandonment, and derive its queueing quantities. In the trading system, for example, orders of impatient customers (i.e., buyers and sellers) are

matched with counterparts in a first-come-first-matched discipline by the system. A customer with multiple-unit orders can be partially filled, and a customer with unmatched orders can abandon the system with a general (discrete) abandonment time. Due to the complicated matching behaviors of customers, the study of the double-sided queueing model is a challenging and interesting problem gaining increasing attention from both the industry and the research community in many fields, including i) *Perishable inventory system:* Perry and Stadje (1999) Bar-Lev et al. (2017); ii) *Financial market:* Afèche et al. (2014); iii) *Taxi-station system:* Kendall (1951); and iv) *Organ transplant system:* Zenios (1999), Boxma et al. (2011) and Akan et al. (2012).

More specifically, double-sided queuing models refer to queuing systems where there are two sides of customers (e.g., buyers and sellers in the trading system; organs and patients in the transplant system; and passengers and taxis in ride-hailing systems), which have been analyzed by various methods. Dobbie (1961) studied the transient queue length of a basic double-sided queueing model with Poisson arrivals for both sides without considering abandonment. This model was further studied by Giveen (1963) with time-dependent arrival rates, and the condition of limiting distribution was discussed. For double-sided queueing models with finite waiting rooms, the steady-state performance has been extensively investigated. For instance, Kashyap (1966) solved double-sided queueing models with a general renewal process and a Poisson arrival process using probability generating functions and the supplementary variable technique. Takahashi and Takahashi (2000) and Takahashi et al. (2000) used Phase-Type renewal processes and Markovian arrival processes in a double-sided queue and analyzed it by matrix-analytic methods. Conolly et al. (2002) studied an exponentially-driven double-sided queueing system with customer abandonment using Laplace transform. Liu et al. (2015) and Liu (2019) used the diffusion process to analyze a double-sided queue with general renewal processes. Afèche et al. (2014) considered a double-sided queueing model with batch arrivals and abandonment using the level crossing technique. Wu and He (2021) recently solved a double-sided queueing model with marked Markovian arrival processes and abandonment. In this paper, we analyze a double-sided queueing model with batch Markovian arrival processes for batches of orders and abandonment, and develop an efficient algorithm for computing queueing quantities at both batch (i.e., buyer or seller) level and order (i.e., single item) level.

The model studied in this paper is relevant to the analysis of the continuous review perishable inventory model without order setup cost or lead time. One pioneering work to connect queueing theory with continuous perishable inventory systems is done by Graves (1982), who used the virtual waiting time in an $M/M/1$ queue with impatient customers and an $M/D/1$ queue with finite buffer to model two continuous review perishable inventory systems. Later, a large number of work is done by Perry (e.g., Kaspi and Perry (1983);

Perry and Posner (1990); Perry and Stadje (1999, 2006)), thus this model is also named *Perry model* by a recent review paper (Karaesmen et al. (2011)). These papers mainly focus on the performance evaluation of systems with random input and output. However, these existing models assume that the arrival of perishable products follows a (possibly state dependent) Poisson process, which may be ineffective when the system is either highly uncertain or close to deterministic. In this paper, we generalize the model by considering batch Markovian arrival processes for both supply and demand arrivals.

To demonstrate the applicability of our model, we present two case studies related to the analysis and design of vaccine inventory systems in remote communities. In 2019, nearly 15% of infants worldwide did not receive an initial dose of the diphtheria-tetanus-pertussis (DTP3) vaccine or were only partially vaccinated based on a report by the World Health Organization (WHO (2020)). Therefore, performance analysis of the vaccine administration process to improve the effectiveness of childhood vaccination programs is important. One of the biggest challenges for doing so is to handle the high uncertainty in i) supply, ii) demand, and iii) storage. First, the uncertainty in supply may be caused by the randomness in resource and capacity limitations in warehouses as well as transportation because of the road conditions given that those remote communities are normally far away from the vaccine warehouse and perhaps only accessible by walking or river (UNICEF (2020)). Thus, the supply delivery time is quite unstable. Second, the demand for vaccines is hard to predict as it may be affected by a number of factors including weather, birth rates, and community perceptions (De Boeck et al. (2019)). There is plenty of anecdotal evidence that the arrivals of demand are non-stationary, but many papers assume stationary demand arrivals (Mofrad et al. (2018)). Third, the uncertainty of storage is mainly caused by unreliable storage equipment and vaccines' limited shelf life. Vaccines are temperature-sensitive products and have to be stored at low temperatures. Keeping vaccines inside the cold chain during transportation and storage is important, but cold chain equipment is often not available or unreliable due to equipment failure and power outages in low- and middle-income countries, especially in remote communities (Ren et al. (2009); Karp et al. (2015); Hsiao et al. (2017)). By using batch Markovian arrival processes ($BMAP$s) and general discrete distributions, the double-sided queueing model we study can capture the three types of uncertainties in the vaccine inventory/transportation system and can be a valuable tool for the design of that system.

In this paper, we apply the multi-layer Markov modulated fluid flow ($MMFF$) processes to analyze the double-sided queueing model with $BMAP$s and finite discrete abandonment times. Our main contributions are as follows:

1. A general double-sided queueing model is studied. i) $BMAP$ is a generalization of many commonly

used point processes including the Poisson process, Markov-modulated Poisson process and Phase-Type renewal process. $BMAP$s can approximate any stochastic arrival process (Asmussen and Koole (1993)). ii) With finite discrete abandonment times, our model covers many interesting abandonment scenarios, such as a) zero abandonment time (i.e., balking); b) constant abandonment time; c) infinite abandonment time (i.e., no abandonment), and d) mixtures of all those scenarios. Our model can approximate general continuous abandonment time distributions by discretization. iii) The abandonment time of a customer in our model depends on its batch size and its position in the queue, which is a fairly practical assumption for some stochastic models, such as the shelf life after opening in perishable inventory systems.

2. The queueing quantities for both sides at both batch and order levels are explicitly formulated by analyzing a multi-layer $MMFF$ process, which facilitates the development of an efficient algorithm to compute them. Some queueing quantities, such as the abandonment probability of the customer at the head of the queue, are not easy to obtain by other means.

3. There is great value in this paper to those studying continuous inventory review perishable inventory models. We investigate more general demand and supply arrival processes as well as abandonment time distributions. Those obtained quantities provide vital information for evaluating and designing inventory systems, such as the fill rates of supply and demand.

4. We apply our model to a vaccine inventory system and evaluate system performance. We consider various system settings and compare the performance under different levels of uncertainty of supply, demand and storage.

The remainder of the paper is organized as follows. We present the key results of multi-layer $MMFF$ processes in Section 2. In Section 3, the queueing model of interest is introduced. In Section 4, we show that a stochastic process related to the age of the buyer/seller at the head of the queue can be modeled as a multi-layer $MMFF$ process. In Section 5, the stationary distribution of the multi-layer $MMFF$ process is utilized to obtain various queueing quantities. In Section 6, we apply our model to analyze a vaccine inventory system. Section 7 concludes this paper.

2 Preliminaries

Multi-layer $MMFF$ processes were first introduced in Bean and O'Reilly (2008) and have been widely applied to many stochastic systems in risk analysis (e.g., Ahn et al. (2007); Badescu et al. (2007a,b); Badescu and

Landriault (2007, 2008, 2009); Badescu (2008)) and queueing theory (e.g., Horváth and Van Houdt (2012); Van Houdt (2012); Horváth (2015); He and Wu (2020); Wu and He (2021)). As a generalization of classical $MMFF$ processes, multi-layer $MMFF$ processes are fluid flow processes in which the change rate of the fluid level is modulated by layer-dependent continuous time Markov chains. A multi-layer $MMFF$ process is a two-dimensional process $\{(X(t), \phi(t)), t \geq 0\}$, where $X(t)$ is the fluid level and $\phi(t)$ is the underlying state at time t. The layers of a multi-layer $MMFF$ process are defined and connected as follows.

1. **Borders and Layers**:

   - There are $M+N+1$ constants such that $l_0 = -\infty < l_1 < ... < l_M = 0 < ... < l_{M+N} = \infty$, to be called *Borders*.

   - Borders form $M+N$ intervals (l_0, l_1), (l_1, l_2), ..., and (l_{M+N-1}, l_{M+N}), to be called *Layer* 1, 2, ..., and $M+N$, respectively.

2. **Generator and fluid rates within Layer n, i.e., $l_{n-1} < X(t) < l_n$, for $n = 1, ..., M+N$**:

   - $\{\phi(t), t \geq 0\}$ is a continuous time irreducible Markov chain on finite state space $\mathcal{S}^{(n)}$ with infinitesimal generator $Q^{(n)}$.

   - The fluid process $\{X(t), t \geq 0\}$ is controlled by $\phi(.)$ such that the value of $X(t)$ changes linearly at rate $c^{(n)}_{\phi(t)}$ at time t. The rate $c^{(n)}_i$ for fluid level change can be positive, negative, or zero. We put the rates into vector $\mathbf{c}^{(n)} = (c^{(n)}_i, i \in \mathcal{S}^{(n)})$. For convenience, the state space $\mathcal{S}^{(n)}$ has to be partitioned into three subsets according to the sign of $c^{(n)}_i$ as follows:

$$\mathcal{S}^{(n)}_+ = \{i \in \mathcal{S}^{(n)} : \ c^{(n)}_i > 0\}, \quad \mathcal{S}^{(n)}_- = \{i \in \mathcal{S}^{(n)} : \ c^{(n)}_i < 0\}, \quad \mathcal{S}^{(n)}_0 = \{i \in \mathcal{S}^{(n)} : \ c^{(n)}_i = 0\}. \tag{2.1}$$

   We further partition $\mathbf{c}^{(n)}$, according to the signs of its elements, and the infinitesimal generator $Q^{(n)}$ of the underlying Markov chain as

$$\mathbf{c}^{(n)} = (\mathbf{c}^{(n)}_+, \ \mathbf{c}^{(n)}_-, \ 0); \quad Q^{(n)} = \begin{matrix} \mathcal{S}^{(n)}_+ \\ \mathcal{S}^{(n)}_- \\ \mathcal{S}^{(n)}_0 \end{matrix} \begin{pmatrix} Q^{(n)}_{++} & Q^{(n)}_{+-} & Q^{(n)}_{+0} \\ Q^{(n)}_{-+} & Q^{(n)}_{--} & Q^{(n)}_{-0} \\ Q^{(n)}_{0+} & Q^{(n)}_{0-} & Q^{(n)}_{00} \end{pmatrix}. \tag{2.2}$$

   Given the generator $Q^{(n)}$ and fluid rate vector $\mathbf{c}^{(n)}$ in Layer n, the mean drift of the fluid in Layer n is

$$\mu^{(n)} = \boldsymbol{\alpha}^{(n)} \mathbf{c}^{(n)}, \tag{2.3}$$

where $\boldsymbol{\alpha}^{(n)}$ is the stationary distribution of $Q^{(n)}$. Note that we also put fluid change rate in diagonal matrices for notational convenience as

$$C_+^{(n)} = \text{diag}(\mathbf{c}_+^{(n)}); \quad C_-^{(n)} = -\text{diag}(\mathbf{c}_-^{(n)}). \tag{2.4}$$

3. **Transitions when reaching Border n, for $n = 1, ..., M+N-1$:**

   - If $X(t)$ reaches l_n from below, the process $\{\phi(t), t \geq 0\}$ can switch from $\mathcal{S}_+^{(n)}$ i) to $\mathcal{S}_-^{(n)}$ (i.e., $X(t)$ reflects back to Layer n) with transition probability matrix $P_{+b-}^{(n)}$; ii) to $\mathcal{S}_+^{(n+1)}$ (i.e., $X(t)$ passes Border n into Layer $(n+1)$) with transition probability matrix $P_{+b+}^{(n)}$; or iii) to $\mathcal{S}_b^{(n)}$ (i.e., $X(t)$ enters Border n) with transition probability matrix $P_{+bb}^{(n)}$. Note that $P_{+b-}^{(n)}\mathbf{e} + P_{+b+}^{(n)}\mathbf{e} + P_{+bb}^{(n)}\mathbf{e} = \mathbf{e}$, where $\mathbf{e}$ is the column vector of ones with an appropriate size.

   - If $X(t)$ reaches l_n from above, the process $\{\phi(t), t \geq 0\}$ can switch from $\mathcal{S}_-^{(n+1)}$ i) to $\mathcal{S}_+^{(n+1)}$ (i.e., $X(t)$ reflects back to Layer $(n+1)$) with transition probability matrix $P_{-b+}^{(n)}$; ii) to $\mathcal{S}_-^{(n)}$ (i.e., $X(t)$ passes Border n to Layer n) with transition probability matrix $P_{-b-}^{(n)}$; or iii) to $\mathcal{S}_b^{(n)}$ (i.e., $X(t)$ enters Border n) with transition probability matrix $P_{-bb}^{(n)}$. Note that $P_{-b+}^{(n)}\mathbf{e} + P_{-b-}^{(n)}\mathbf{e} + P_{-bb}^{(n)}\mathbf{e} = \mathbf{e}$.

4. **Transitions of Border n, i.e, $X(t) = l_n$ and $\phi(t) \in \mathcal{S}_b^{(n)}$, where $\mathcal{S}_b^{(n)}$ is the set of states within Border n, for $n = 1, ..., M+N-1$:**

   - $X(t)$ remains at l_n during the period that $\phi(t)$ is in $\mathcal{S}_b^{(n)}$ with sub-generator $Q_b^{(n)}$, until $\phi(t)$ switches from $\mathcal{S}_b^{(n)}$ to either $\mathcal{S}_-^{(n)}$ with transition rate matrix $Q_{b-}^{(n)}$ or $\mathcal{S}_+^{(n+1)}$ with transition rate matrix $Q_{b+}^{(n)}$. Note that $Q_b^{(n)}\mathbf{e} + Q_{b+}^{(n)}\mathbf{e} + Q_{b-}^{(n)}\mathbf{e} = 0$.

With the above definition, if we define $c_i^{(n)} = 0$ for all n and $i \in \mathcal{S}_b^{(n)}$, then $X(t)$ is controlled by $\phi(t)$ explicitly as

$$X(t) = X(0) + \int_0^t c_{\phi(s)}^{(L(X(s)))} \mathrm{d}s, \quad \text{or} \quad \frac{\mathrm{d}X(t)}{\mathrm{d}t} = c_{\phi(t)}^{(L(X(t)))}, \tag{2.5}$$

where $L(x) = n$ if $l_{n-1} < x \leq l_n$, for $n = 1, ..., N$.

The stationary joint distribution of multi-layer $MMFF$ processes is crucial for obtaining the steady-state queueing quantities in our model and can be expressed by a set of basic quantities in classical $MMFF$ processes as follows.

1. $\Psi^{(n)}$ ($\widehat{\Psi}^{(n)}$) (Latouche and Nguyen (2018); Meini (2013); Ramaswami (1999); Rogers (1994)) is defined

as the transition probability matrix of the state of underlying Markov chain $\phi(t)$ from an epoch that $X(t)$ starts to increase (decrease) from 0 to the next first epoch that $X(t)$ returns to 0. The two matrices can be obtained by solving the following (matrix) quadratic Riccati equations:

$$\begin{aligned} Q_{+-}^{(n)} + Q_{++}^{(n)}\Psi^{(n)} + \Psi^{(n)}Q_{--}^{(n)} + \Psi^{(n)}Q_{-+}^{(n)}\Psi^{(n)} &= 0; \\ Q_{-+}^{(n)} + Q_{--}^{(n)}\widehat{\Psi}^{(n)} + \widehat{\Psi}^{(n)}Q_{++}^{(n)} + \widehat{\Psi}^{(n)}Q_{+-}^{(n)}\widehat{\Psi}^{(n)} &= 0. \end{aligned} \tag{2.6}$$

2. $\mathcal{U}^{(n)}$ ($\widehat{\mathcal{U}}^{(n)}$) (Asmussen (1995)) is defined as the infinitesimal generator of the state of the underlying Markov chain $\phi(t)$ when the fluid level reaches a new lower (higher) point. The two matrices can be derived from $\Psi^{(n)}$ and $\widehat{\Psi}^{(n)}$:

$$\begin{aligned} \mathcal{U}^{(n)} &= Q_{--}^{(n)} + Q_{-+}^{(n)}\Psi^{(n)}; \\ \widehat{\mathcal{U}}^{(n)} &= Q_{++}^{(n)} + Q_{+-}^{(n)}\widehat{\Psi}^{(n)}. \end{aligned} \tag{2.7}$$

3. $\mathcal{K}^{(n)}$ ($\widehat{\mathcal{K}}^{(n)}$) (Ramaswami (1999)) is a matrix associated with the numbers of visits to certain fluid level and underlying state during first passage periods. The two matrices can be derived from $\Psi^{(n)}$ and $\widehat{\Psi}^{(n)}$:

$$\begin{aligned} \mathcal{K}^{(n)} &= Q_{++}^{(n)} + \Psi^{(n)}Q_{-+}^{(n)}; \\ \widehat{\mathcal{K}}^{(n)} &= Q_{--}^{(n)} + \widehat{\Psi}^{(n)}Q_{+-}^{(n)}. \end{aligned} \tag{2.8}$$

To analyze the queueing model of interest, only a special type of multi-layer $MMFF$ processes with $\mathbf{c}^{(n)} = (\mathbf{e}^T, -\mathbf{e}^T)$ and one sticky Border (i.e., fluid can only enter Border $l_M = 0$) is required, i.e., $C_+^{(n)} = I$, $C_-^{(n)} = I$, for all n; and $\mathcal{S}_b^{(n)} = \phi$ (empty set), for all n except $n = M$. Therefore, we present the stationary joint distribution and its computational steps only for the special multi-layer $MMFF$ process in Theorem 1 and Appendix A, respectively.

Let $\pi_j^{(n)}(x)$ be the stationary joint density of level x with underlying state j of the fluid flow process $\{(X(t), \phi(t)), t \geq 0\}$. Define density vectors $\boldsymbol{\pi}_+^{(n)}(x) = \left(\pi_j^{(n)}(x) : \ j \in \mathcal{S}_+^{(n)}\right)$ and $\boldsymbol{\pi}_-^{(n)}(x) = \left(\pi_j^{(n)}(x) : \ j \in \mathcal{S}_-^{(n)}\right)$, for $-\infty < x < \infty$ and $n = 1, 2, ..., M+N$. Let $\mathbf{p}^{(M)}$ be the border probabilities on Border M.

Theorem 1. *(He and Wu (2020)) We assume that i) $\mu^{(M+N)} < 0$ and $\mu^{(1)} > 0$; ii) $\mu^{(n)} \neq 0$ for $n = 2, ..., M+N-1$. Let, for $n = 1, 2, ..., M+N$,*

$$\left(\mathbf{u}_+^{(n)}, \mathbf{u}_-^{(n)}\right) = \left(\boldsymbol{\pi}_+^{(n)}(l_{n-1}), \boldsymbol{\pi}_-^{(n)}(l_n)\right) \begin{pmatrix} I & e^{\mathcal{K}^{(n)}b_n\Psi^{(n)}} \\ e^{\widehat{\mathcal{K}}^{(n)}b_n\widehat{\Psi}^{(n)}} & I \end{pmatrix}^{-1}, \tag{2.9}$$

where $\{\boldsymbol{\pi}_+^{(n)}(l_{n-1}),\ \boldsymbol{\pi}_-^{(n)}(l_n),\ n=1,2,...,M+N\}$ *are given in Appendix A. Then, we have for* $x<l_1$*, the joint density function is*

$$\left(\boldsymbol{\pi}_+^{(1)}(x),\ \boldsymbol{\pi}_-^{(1)}(x)\right)=\left(\mathbf{u}_-^{(1)}e^{\widehat{\mathcal{K}}^{(1)}(l_1-x)}\widehat{\Psi}^{(1)},\quad \mathbf{u}_-^{(1)}e^{\widehat{\mathcal{K}}^{(1)}(l_1-x)}\right). \tag{2.10}$$

For $n=2,...,M+N-1$*, we have, for* $l_{n-1}<x<l_n$*, the joint density function is*

$$\left(\boldsymbol{\pi}_+^{(n)}(x),\ \boldsymbol{\pi}_-^{(n)}(x)\right)=\left(\mathbf{u}_+^{(n)}e^{\mathcal{K}^{(n)}(x-l_{n-1})}+\mathbf{u}_-^{(n)}e^{\widehat{\mathcal{K}}^{(n)}(l_n-x)}\widehat{\Psi}^{(n)},\quad \mathbf{u}_+^{(n)}e^{\mathcal{K}^{(n)}(x-l_{n-1})}\Psi^{(n)}+\mathbf{u}_-^{(n)}e^{\widehat{\mathcal{K}}^{(n)}(l_n-x)}\right). \tag{2.11}$$

For $x>l_{M+N-1}$*, the joint density function is*

$$\left(\boldsymbol{\pi}_+^{(M+N)}(x),\ \boldsymbol{\pi}_-^{(M+N)}(x)\right)=\left(\mathbf{u}_+^{(M+N)}e^{\mathcal{K}^{(M+N)}(x-l_{M+N-1})},\quad \mathbf{u}_+^{(M+N)}e^{\mathcal{K}^{(M+N)}(x-l_{M+N-1})}\Psi^{(M+N)}\right). \tag{2.12}$$

Border probabilities $\mathbf{p}^{(M)}$ *are given in Step 3 in Appendix A.*

Remark: Condition i) is imposed to ensure the existence of the stationary joint distribution. Condition ii) is not essential and can be removed. Yet the algorithm for computing the stationary joint distribution is much more involved without Condition ii) (see Latouche and Nguyen (2018)).

3 The Model

In this section, we define a double-sided queueing system with batch arrivals and abandonment. The system has two types of agents/customers, to be called *buyers* and *sellers*, arriving at the system independently. Each buyer (seller) has a number of buyer (seller) orders, which are expected to be matched by seller (buyer) orders. The number of orders held by a buyer (seller) is called its *batch size*. The batch size of a buyer (seller) may change after some of its orders are matched by seller (buyer) orders. A buyer (seller) order has to be matched with a seller (buyer) order if there is one or more than one seller (buyer) orders in the system. As soon as a buyer order is matched with a seller order, the pair leaves the system immediately, thus the *buyer queue* and the *seller queue* do not co-exist in the system at any time. The matching rule is *first-arrived-first-matched*. Within a batch (i.e., a buyer or seller), we do not specify the matching order since it does not affect the quantities of interest in this paper.

Each buyer (seller) has limited patience and may abandon the system before all its orders are matched. If a buyer (seller) abandons the system, all remaining (unmatched) orders of the buyer (seller) are removed from the system. The abandonment time of a buyer (seller) depends on its batch size and its position in

the buyer (seller) queue. A buyer (seller) in the buyer (seller) queue can be at the head of the queue or in a position behind the head of the queue (i.e., any position other than the head of the queue). Specifically, the abandonment mechanism for buyers is defined as follows:

- If a buyer arrives and finds a buyer queue, the buyer joins the buyer queue and, conditioning on its batch size, its abandonment time is sampled. The abandonment time of the buyer will stay with the buyer until it abandons the queue or it becomes the head of the queue.

- If a buyer arrives and finds an empty system, the buyer forms a buyer queue and its abandonment time is sampled. The abandonment time of the buyer will stay with the buyer until it abandons the queue or its batch size changes due to the arrival of sellers.

- If a buyer arrives and finds a seller queue, its orders are matched by seller orders in the seller queue. If the buyer's batch size is greater than the total batch size of all sellers in the queue, the seller queue disappears and the buyer forms a buyer queue by itself. Its abandonment time is sampled accordingly. For every buyer, matching takes priority over abandonment.

- If a waiting buyer becomes the head of the queue, its abandonment time is re-sampled based on its new position and batch size, conditioning on its elapsed waiting time. We assume that the conditional distribution exists for any possible elapsed waiting time.

- At the head of the queue, if the batch size of a buyer is changed (this may happen more than one time due to the arrivals of sellers), its abandonment time is re-sampled based on its new batch size, conditioning on the elapsed waiting time. Again, we assume that the conditional distribution exists for any possible elapsed waiting time.

The abandonment mechanism for sellers is defined similarly. We note that Example 1 at the end of this section is a running example containing all assumptions and mechanisms described above, while real applications may have simpler abandonment mechanisms (i.e., the abandonment time is independent of batch size and/or position in the queue). We consider the dependency to make our model more generic. Next, the arrival processes and abandonment time distributions of the queueing model are defined explicitly in the following four items:

1. **Buyer's arrival process:** Assume that the maximum batch size of a buyer is K. Buyers arrive to the queueing system according to a continuous time $BMAP$ with matrix representation $(D_0, D_1, ..., D_K)$ of order m_b. The underlying Markov chain of the arrival process $\{I_b(t), t \geq 0\}$ with generator $D =$

$D_0 + D_1 + ... + D_K$ is irreducible and has stationary distribution $\boldsymbol{\theta}_b$. The (average) arrival rate of batch size k buyers is given by $\lambda_k = \boldsymbol{\theta}_b D_k \mathbf{e}$, for $k = 1, ..., K$. Define $\lambda = \sum_{k=1}^{K} k\lambda_k$ as the arrival rate of buyer orders.

2. **Buyer's abandonment times:** Assume that the abandonment time of a buyer of batch size k, for $k = 1, 2, ..., K$, before becoming the head of the buyer queue, is τ_k, which has a discrete distribution: $P\{\tau_k = \tilde{l}_n\} = \eta_{k,n}$, for $n = 0, 1, ..., N$, where $\tilde{l}_0 = 0 < \tilde{l}_1 < ... < \tilde{l}_{N-1} < \tilde{l}_N = \infty$ are the *possible abandonment times*. Assume that the abandonment time of a buyer of batch size k, after becoming the head of the buyer queue, is $\dot{\tau}_k$, which has a discrete distribution: $P\{\dot{\tau}_k = \tilde{l}_n\} = \dot{\eta}_{k,n}$, for $n = 0, 1, ..., N$. Note that the batch size k of a buyer may change after the buyer becomes the head of the buyer queue, due to matching of orders.

3. **Seller's arrival process:** Without loss of generality, we assume the maximum batch size of sellers is also K. Sellers arrive to the queueing system according to a different $BMAP$ with matrix representation $(\hat{D}_0, \hat{D}_1, ..., \hat{D}_K)$ of order m_s. The underlying Markov chain of the arrival process $\{I_s(t), t \geq 0\}$ with generator $\hat{D} = \hat{D}_0 + \hat{D}_1 + ... + \hat{D}_K$ is irreducible and has stationary distribution $\boldsymbol{\theta}_s$. The (average) arrival rate of size k sellers is given by $\mu_k = \boldsymbol{\theta}_s \hat{D}_k \mathbf{e}$, for $k = 1, ..., K$. Define $\mu = \sum_{k=1}^{K} k\mu_k$ as the arrival rate of the seller orders.

4. **Seller's abandonment times:** Assume that the abandonment time of a seller of batch size k, for $k = 1, 2, ..., K$, before reaching the head of the seller queue, is $\hat{\tau}_k$, which has a discrete distribution: $P\{\hat{\tau}_k = \hat{l}_m\} = \hat{\eta}_{k,m}$, for $m = 0, 1, ..., M$, where $\hat{l}_0 = 0 < \hat{l}_1 < ... < \hat{l}_{M-1} < \hat{l}_M = \infty$ are the possible abandonment times. Assume that the abandonment time of a seller of size k, after becoming the head of the queue, is $\dot{\hat{\tau}}_k$, which has a discrete distribution: $P\{\dot{\hat{\tau}}_k = \hat{l}_m\} = \dot{\hat{\eta}}_{k,m}$, for $m = 0, 1, ..., M$.

In the rest of the paper, we make the following assumptions to ensure the stability of the queueing model (i.e., a finite buyer queue and a finite seller queue probabilistically).

$$\sum_{k=1}^{K} k\lambda_k \eta_{k,N} < \mu \quad \text{and} \quad \sum_{k=1}^{K} k\mu_k \hat{\eta}_{k,M} < \lambda. \tag{3.1}$$

It is important to note that the abandonment of a buyer (seller) means that all remaining orders of the buyer (seller) abandon the system together. The model is more complicated if orders of a buyer (seller) can abandon the queue individually since that allows partial abandonment of a buyer (seller). That model is still solvable, but it is much more complicated, and we do not consider that case in this paper.

In order to illustrate the queueing model, we present the following example, in which we will draw sample paths of the age processes to show the dynamics of the model in Section 4 and give the queueing quantities as we derive the results in Section 5.

Example 1. A double-sided queue with maximum batch size $K = 3$ and $M = N = 5$. All the input parameters are presented in Table 1.

Buyer Arrival	Batch Size	$\tilde{l}_1 = 1$	$\tilde{l}_2 = 3$	$\tilde{l}_3 = 5$	$\tilde{l}_4 = 7$	$\tilde{l}_5 = \infty$
$D_0 = \begin{pmatrix} -8, & 2 \\ 3, & -7 \end{pmatrix}$	$D_1 = \begin{pmatrix} 1, & 2 \\ 0, & 2 \end{pmatrix}$	$\eta_{1,1} = 0$ $\dot{\eta}_{1,1} = 0$	$\eta_{1,2} = 0$ $\dot{\eta}_{1,2} = 0$	$\eta_{1,3} = 0$ $\dot{\eta}_{1,3} = 0$	$\eta_{1,4} = 0.9$ $\dot{\eta}_{1,4} = 0.1$	$\eta_{1,5} = 0.1$ $\dot{\eta}_{1,5} = 0.9$
	$D_2 = \begin{pmatrix} 2, & 0 \\ 0, & 1 \end{pmatrix}$	$\eta_{2,1} = 0.4$ $\dot{\eta}_{2,1} = 0.1$	$\eta_{2,2} = 0.3$ $\dot{\eta}_{2,2} = 0.1$	$\eta_{2,3} = 0.2$ $\dot{\eta}_{2,3} = 0.1$	$\eta_{2,4} = 0.1$ $\dot{\eta}_{2,4} = 0.1$	$\eta_{2,5} = 0$ $\dot{\eta}_{2,5} = 0.6$
	$D_3 = \begin{pmatrix} 1, & 0 \\ 0, & 1 \end{pmatrix}$	$\eta_{3,1} = 0.1$ $\dot{\eta}_{3,1} = 0.1$	$\eta_{3,2} = 0.1$ $\dot{\eta}_{3,2} = 0.1$	$\eta_{3,3} = 0.1$ $\dot{\eta}_{3,3} = 0.1$	$\eta_{3,4} = 0.1$ $\dot{\eta}_{3,4} = 0.1$	$\eta_{3,5} = 0.6$ $\dot{\eta}_{3,5} = 0.6$
Seller Arrival	Batch Size	$\hat{l}_1 = 2$	$\hat{l}_2 = 3$	$\hat{l}_3 = 4$	$\hat{l}_4 = 5$	$\hat{l}_5 = \infty$
$\hat{D}_0 = \begin{pmatrix} -5, & 1 \\ 1, & -10 \end{pmatrix}$	$\hat{D}_1 = \begin{pmatrix} 1, & 0 \\ 1, & 6 \end{pmatrix}$	$\hat{\eta}_{1,1} = 0.2$ $\dot{\hat{\eta}}_{1,1} = 0$	$\hat{\eta}_{1,2} = 0.2$ $\dot{\hat{\eta}}_{1,2} = 0$	$\hat{\eta}_{1,3} = 0.2$ $\dot{\hat{\eta}}_{1,3} = 0$	$\hat{\eta}_{1,4} = 0.2$ $\dot{\hat{\eta}}_{1,4} = 0.2$	$\hat{\eta}_{1,5} = 0.2$ $\dot{\hat{\eta}}_{1,5} = 0.8$
	$\hat{D}_2 = \begin{pmatrix} 2, & 0 \\ 0, & 2 \end{pmatrix}$	$\hat{\eta}_{2,1} = 0.9$ $\dot{\hat{\eta}}_{2,1} = 0$	$\hat{\eta}_{2,2} = 0$ $\dot{\hat{\eta}}_{2,2} = 0$	$\hat{\eta}_{2,3} = 0$ $\dot{\hat{\eta}}_{2,3} = 0$	$\hat{\eta}_{2,4} = 0$ $\dot{\hat{\eta}}_{2,4} = 0.1$	$\hat{\eta}_{2,5} = 0.1$ $\dot{\hat{\eta}}_{2,5} = 0.9$
	$\hat{D}_3 = \begin{pmatrix} 1, & 0 \\ 0, & 0 \end{pmatrix}$	$\hat{\eta}_{3,1} = 1$ $\dot{\hat{\eta}}_{3,1} = 0$	$\hat{\eta}_{3,2} = 0$ $\dot{\hat{\eta}}_{3,2} = 0$	$\hat{\eta}_{3,3} = 0$ $\dot{\hat{\eta}}_{3,3} = 0.1$	$\hat{\eta}_{3,4} = 0$ $\dot{\hat{\eta}}_{3,4} = 0.1$	$\hat{\eta}_{3,5} = 0$ $\dot{\hat{\eta}}_{3,5} = 0.8$

Table 1: Parameters of Example 1.

In tihs example, the batch sizes for both buyers and sellers are 3. The first two columns in Table 1 present the matrix-representations of the two BMAPS. The distributions of abandonment times of buyers and sellers are given in columns 3 to 7 in Table 1. For example, the arrival of a buyer of batch size 2 is modeled by D_2 in row 3. If the buy joins a buyer queue (after the head of the queue), its abandonment time is sampled by using the probability distribution $\{\eta_{2,1} = 0.4, \eta_{2,2} = 0.3, \eta_{2,3} = 0.2, \eta_{2.4} = 0.1, \eta_{2,5} = 0\}$. If the buyer is at the head of a buyer queue, its abandonment time is sampled from probability distribution $\{\dot{\eta}_{2,1} = 0.4, \dot{\eta}_{2,2} = 0.3, \dot{\eta}_{2,3} = 0.2, \dot{\eta}_{2.4} = 0.1, \dot{\eta}_{2,5} = 0\}$, even if its batch size has been changed to 1.

4 Age Processes and Multi-layer $MMFF$ Process

In this section, we first introduce the age processes of the buyers and sellers in the double-sided queueing model in Subsection 4.1. Then we convert the age processes into a multi-layer $MMFF$ process for the analysis of the queueing model in Subsection 4.2.

4.1 Age Processes

We define *age* of a buyer (seller) in the system as the time elapsed since the buyer (seller) enters the system. The ages of the buyers and the ages of sellers can never both be positive because the buyer queue and the seller queue can never coexist in the system. Let $a_B(t)$ be the age of the buyer at the head of the buyer queue at time t, if the buyer queue is not empty; otherwise, $a_B(t) = 0$. The age $a_S(t)$ of the seller at the head of the seller queue at time t is defined similarly. If both $a_B(t)$ and $a_S(t)$ are zero, then the system is empty at time t. If we flip $a_S(t)$ of sellers over the horizontal axis (i.e., change $a_S(t)$ to $-a_S(t)$) (see the green lines in Figure 1), we can combine the two age processes $\{a_B(t), t \geq 0\}$ and $\{a_S(t), t \geq 0\}$ into a one-dimensional stochastic process $\{a(t), t \geq 0\}$, to be called *the age process*, as $a(t) = a_B(t)$, if $a_B(t) > 0$; $a(t) = -a_S(t)$, if $a_S(t) > 0$; and $a(t) = 0$, otherwise.

Likewise, we track the remaining batch size of the buyer (seller) at the head of the queue. Let $s_B(t)$ be the remaining batch size of the buyer at the head of the queue at time t, if the buyer queue is not empty; otherwise, $s_B(t) = 0$. Let $s_S(t)$ be the remaining batch size of the seller at the head of the queue at time t, if the seller queue is not empty; otherwise, $s_S(t) = 0$. By flipping the batch size of sellers over the horizontal axis, we can convert the two-dimensional process $\{(s_B(t), s_S(t)), t \geq 0\}$ into a one-dimensional stochastic process $\{s(t), t \geq 0\}$ as $s(t) = s_B(t)$, if $s_B(t) > 0$; $s(t) = -s_S(t)$, if $s_S(t) > 0$; and $s(t) = 0$, otherwise.

Note that in Figure 1, we use blue lines to represent the age of buyers, and green lines to represent the flipped age of sellers. The blue lines together with the green lines show the age process $\{a(t), t \geq 0\}$. We also use single, double, and triple lines to indicate that the batch size is one, two, and three, respectively.

Now, we have a two-dimensional stochastic process $\{(a(t), s(t)), t \geq 0\}$ capturing both the age and remaining batch size. For notational convenience, we define constants $\{l_n, n = 0, 1, ..., M + N\}$ as: $l_0 = -\infty$, $l_n = -\hat{l}_{M-n}$, for $n = 1, 2, ..., M - 1$, $l_M = 0$, $l_{M+n} = \tilde{l}_n$, for $n = 1, 2, ..., N - 1$, and $l_{M+N} = \infty$. The dynamics of $(a(t), s(t))$ can be described by five cases at any time t as follows.

1. If $0 \leq l_{M+n} < a(t) < l_{M+n+1}$, for $n = 0, 1, ..., N - 1$, there are two situations for $\{(a(t), s(t)), t \geq 0\}$.

   (a) If there is no seller arrival at time t, $a(t)$ equals the age of the buyer at the head of the queue and increases linearly at rate 1; and $s(t + 0) = s(t)$.

   (b) If a seller arrives at time t with batch size v, and the total batch size of all buyers currently in the queue is $\hat{s}(t)$ (Note that $\hat{s}(t) \geq s(t)$), we have

      i) if $v < s(t)$, then $a(t)$ equals the age of the buyer at the head of the queue and increases

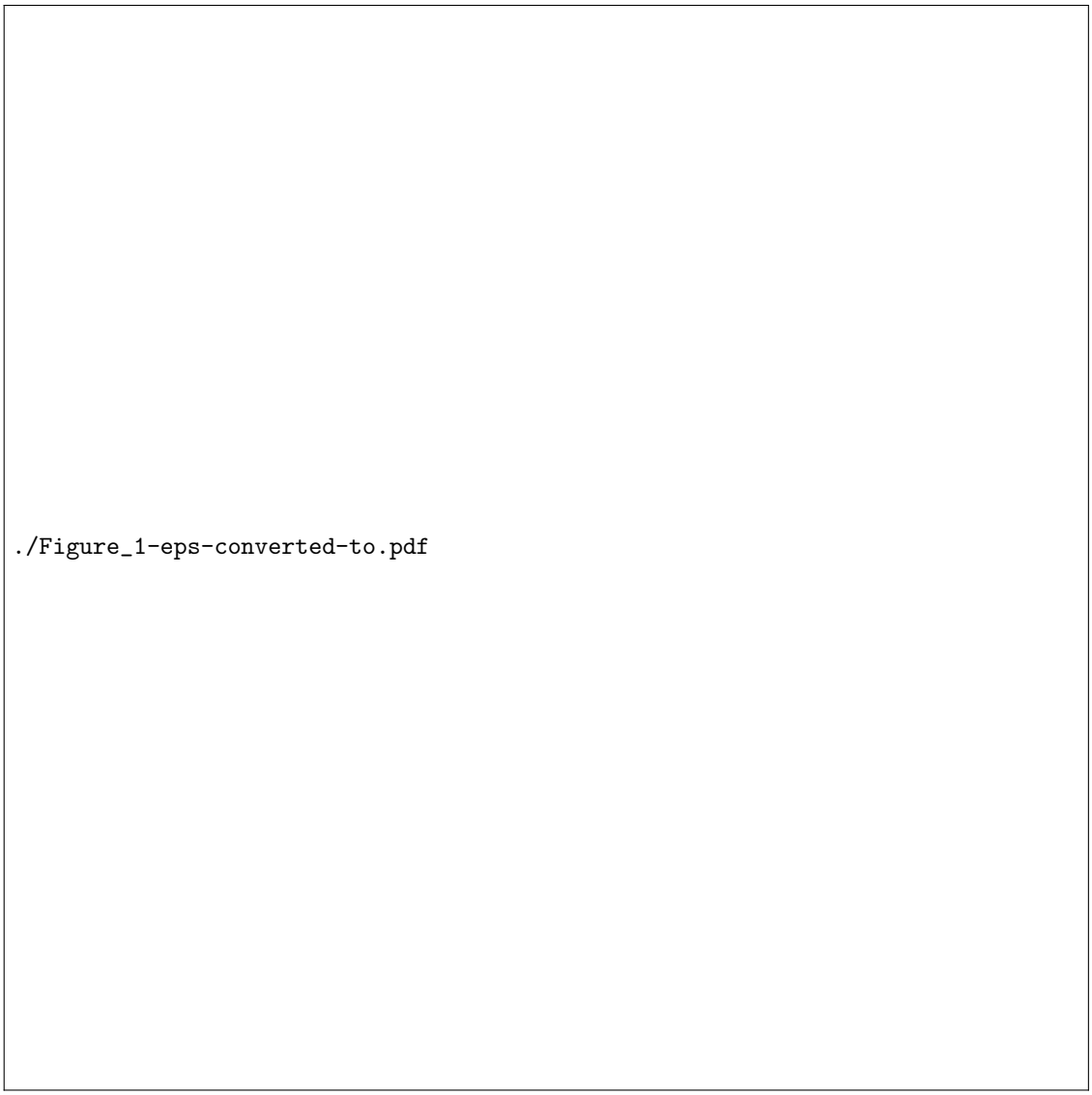

Figure 1: A sample path of the age process of Example 1.

linearly at rate 1; and $s(t+0) = s(t) - v$;

ii) if $s(t) \leq v < \hat{s}(t)$, we find the first waiting buyer (to be called a *tagged* buyer) such that the total batch size $\hat{s}_{tag}(t)$ of the tagged buyer and buyers currently ahead of the tagged buyer is strictly greater than v, then $a(t+0) = a(t) - u$, where u is the interarrival time between the buyer at the head of the queue and the tagged buyer, and $s(t+0) = \hat{s}_{tag}(t) - v$; (Note: The tagged buyer will be the next buyer to be at the head of the queue. Some buyers arrived earlier than the tagged buyer may have abandoned the queue after their arrivals to the queue.)

iii) if $s(t) \leq v = \hat{s}(t)$, then $a(t+0) = 0$ and $s(t+0) = 0$; and

iv) if $\hat{s}(t) < v$, then $a(t+0) = 0$ and starts to decrease at rate -1, and $s(t+0) = \hat{s}(t) - v$.

2. If $a(t) = l_{M+n} > 0$ for $n = 1, 2, ..., N-1$,

   (a) with probability $1 - \dot{\eta}_{k,n}/(\dot{\eta}_{k,n} + ... + \dot{\eta}_{k,N})$, $a(t)$ continues to increase linearly at rate 1 and $s(t+0) = s(t) = k$;

   (b) otherwise, $a(t+0) = \max\{0, l_{M+n} - u\}$, where u is the interarrival time between the departing buyer (due to abandonment) and the buyer who is currently behind it, and if $a(t+0) > 0$, then $s(t+0)$ is the batch size of the buyer who is now at the head of the queue; if $a(t+0) = 0$, then $s(t+0) = 0$. We note that some buyers arrived after the departing buyer may have left the queueing system due to abandonment.

3. This is a symmetric case of Case 1. If $l_{n-1} < a(t) < l_n \leq 0$, for $n = 1, ..., M$, there are two situations for $\{(a(t), s(t)), t \geq 0\}$.

   (a) If there is no buyer arrival at time t, $a(t)$ equals the age of the seller at the head of the queue and decreases linearly at rate -1; and $s(t+0) = s(t)$.

   (b) If a buyer arrives at time t with batch size v, and the total batch size of all sellers currently in the queue is $\hat{s}(t)$ (Note that $\hat{s}(t) \geq -s(t)$), we have

      i) if $v < -s(t)$, then $a(t)$ equals the age of the seller at the head of the queue and increases linearly at rate one; and $s(t+0) = s(t) + v$;

      ii) if $-s(t) \leq v < \hat{s}(t)$, we find the first waiting seller (a.k.a. *tagged* seller) such that the total batch size $\hat{s}_{tag}(t)$ of the tagged seller and sellers currently ahead of the tagged seller is greater than v, then $a(t+0) = a(t) - u$, where u is the interarrival time between the seller at the head of the queue and the tagged seller, and $s(t+0) = v - \hat{s}_{tag}(t)$;

      iii) if $-s(t) \leq v = \hat{s}(t)$, then $a(t+0) = 0$ and $s(t+0) = 0$; and

      iv) if $\hat{s}(t) < v$, then $a(t+0) = 0$ and starts to increase at rate 1, and $s(t+0) = v - \hat{s}(t)$.

4. This is a symmetric case of Case 2. If $a(t) = l_n < 0$, for $n = 1, 2..., M-1$,

   (a) with probability $1 - \dot{\hat{\eta}}_{k,M-n}/(\dot{\hat{\eta}}_{k,M-n} + ... + \dot{\hat{\eta}}_{k,M})$, $a(t)$ continues to decrease linearly at rate -1

and $s(t+0) = s(t) = -k$;

(b) otherwise, $a(t+0) = \min\{0, l_n + u\}$, where u is the interarrival time between the departing seller (due to abandonment) and the seller is currently behind it, and if $a(t+0) < 0$, then $s(t+0)$ is the flipped batch size of the seller who is now at the head of the queue; if $a(t+0) = 0$, then $s(t+0) = 0$.

5. If $a(t) = 0$, we have $s(t) = 0$. That $(a(t), s(t))$ remains to be $(0, 0)$ until the arrival of the next buyer or seller.

In order to analyze the queueing model, we need to track the underlying states of the two $BMAPs$. However, the process becomes very complicated as both $BMAPs$ are evolving at the same time so that we need to track the ages of all buyers (sellers) in the system instead of the age of the buyer (seller) at the head of the queue. Based on the assumption that two $BMAPs$ are independent, we can "freeze" one of the $BMAPs$ during some periods of the age process and then "unfreeze" it during periods of jumps. We use two supplementary variables $\{I_{(b)}(t), I_{(s)}(t)\}$ to represent the underlying states of the two $BMAPs$ $\{I_b(t), I_s(t)\}$ as follows,

- When $a(t) > 0$ is in an increasing period, the buyer arrival process is frozen and the seller arrival process is evolving, so $I_{(b)}(t)$ is fixed at the phase of the buyer arrival process at the beginning of the period and $I_{(s)}(t) = I_s(t)$; When $a(t) < 0$ is in a decreasing period, the buyer arrival process is evolving and the seller arrival process is frozen, so $I_{(s)}(t)$ is fixed at the phase of the seller arrival process at the beginning of the period and $I_{(b)}(t) = I_b(t)$;

- When $a(t) > 0$ is in a down jump period, the buyer arrival process is evolving and the seller arrival process is frozen, so $I_{(b)}(t) = I_b(t)$ and $I_{(s)}(t)$ is fixed at the phase of the seller arrival process at the beginning of the period; When $a(t) < 0$ is in an up jump period, the buyer arrival process is frozen and the seller arrival process is evolving, so $I_{(b)}(t)$ is fixed at the phase of the buyer arrival process at the beginning of the period and $I_{(s)}(t) = I_s(t)$; and

- When $a(t)$ is in a period in which $a(t) = 0$ and $s(t) = 0$, both arrival processes are evolving, so $I_{(b)}(t) = I_b(t)$ and $I_{(s)}(t) = I_s(t)$.

We recycle the name *age process* and call the stochastic process $\{(a(t), s(t), I_{(b)}(t), I_{(s)}(t)), t \geq 0\}$ an age

process with state space

$$\{\{(-\infty,0)\times\{-K,...,-1\}\}\cup\{0\times 0\}\cup\{(0,\infty)\times\{1,...,K\}\}\}\times\{1,...,m_b\}\times\{1,...,m_s\}. \tag{4.1}$$

4.2 Multi-layer $MMFF$ Process

The analysis of the age process is challenging because of the jumps in the process. Thus, we add fictitious time periods and replace those jumps with linear increasing and decreasing periods to construct a multi-layer $MMFF$ process $\{(X(t), s(t), I_{(b)}(t), I_{(s)}(t)),\ t \geq 0\}$ (see Figure 2). Specifically, the original increasing and decreasing periods in the age process are kept in the $MMFF$ process and called *real (time) periods*. However, the up and down jumps in the age process are replaced with *fictitious (time) periods* of linear increase at rate 1 and decrease at rate -1, respectively, in the $MMFF$ process. The length of these fictitious periods equals the heights of the up or down jumps.

- In a real period, $X(t) = a(t)$, and $|s(t)|$ is the remaining batch size of the buyer/seller at the head of the queue. If it is a buyer queue at time t, $X(t)$ increases linearly at rate 1 and $s(t)$ may decreases when a seller arrives and a part of the $s(t)$ orders is matched with the seller's orders. When the orders of the buyer at the head of the queue are all gone, the buyer leaves the system. The process $X(t)$ switches to a fictitious period and starts to decrease. At this point, $s(t)$ is either zero or negative, and $-s(t)$ is the number of unmatched seller orders. The dynamics of $X(t)$ and $s(t)$ for a seller queue at time t is similar.

- In a fictitious period, $X(t)$ is determined by the fluid level at the start of the period plus the product of the changing rate (1 or -1) and the elapsed time from the start of the period, and $|s(t)|$ is the remaining batch size of the newly arrived seller, if it is a buyer queue at time t, or the newly arrived buyer, if it is a seller queue at time t, that will be matched by orders of buyers (sellers) in the queue. When the $|s(t)|$ orders (either buyer's or seller's) are all matched, $X(t)$ will switch to a real period. In a fictitious period, $|s(t)|$ takes values $\{0, 1, ..., K-1\}$ since at least one order of a newly arrived seller/buyer is matched by orders of the buyers/sellers already in queue.

Accordingly, in an increasing period, $I_{(b)}(t)$ is fixed (i.e., frozen) at the phase of the buyer arrival process at the beginning of the increasing period and $I_{(s)}(t)$ is evolving (i.e., $I_{(s)}(t) = I_s(t)$), while in a decreasing period, $I_{(b)}(t)$ is evolving (i.e., $I_{(b)}(t) = I_b(t)$) and $I_{(s)}(t)$ is fixed (i.e., frozen) at the phase of the seller arrival process at the beginning of the decreasing period.

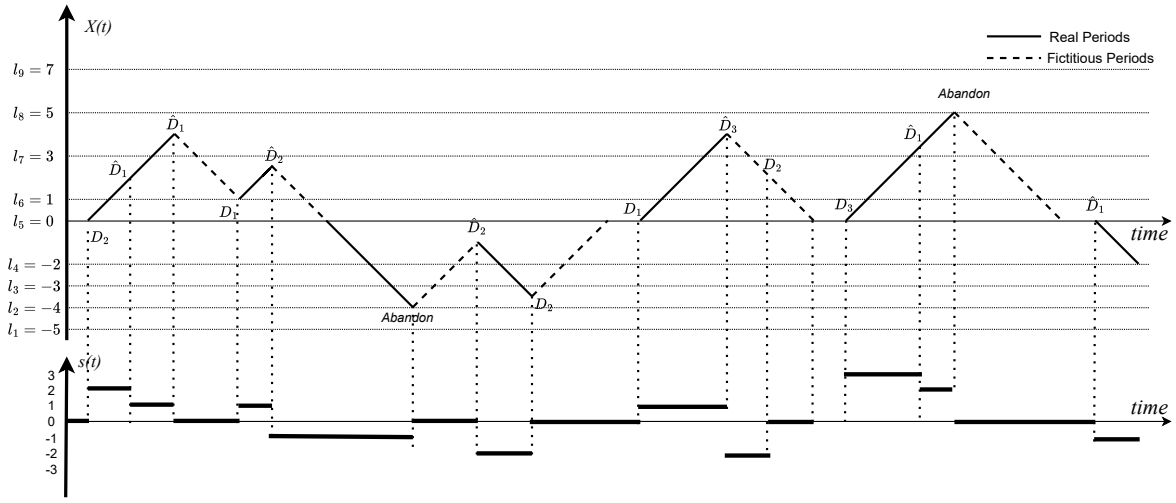

Figure 2: The corresponding $MMFF$ process of the sample path of the age process in Figure 1.

Let $\phi(t) = (s(t), I_{(b)}(t), I_{(s)}(t))$ for $t \geq 0$, the multi-layer $MMFF$ process $\{(X(t), \phi(t)), t \geq 0\}$ is a continuous time Markov process since $\{\phi(t), t \geq 0\}$ is a continuous time Markov chain when $X(t)$ is in each of some predetermined intervals (i.e., layers). We can obtain the stationary joint distribution of the age process by analyzing the corresponding multi-layer $MMFF$ process.

Next, we use the parameters of the queueing model to construct the multi-layer $MMFF$ process $\{(X(t), \phi(t)), t \geq 0\}$ explicitly. The associated state space and transition matrices are specified as follows:

1. There are $M + N$ layers with Borders l_n, for $n = 0, 1, ...M, M + 1, ..., M + N$. Note that $l_0 = -\infty$, $l_M = 0$ and $l_{M+N} = \infty$.

2. The state space of $\phi(t)$ for Layer n, for $n = 1, 2, ..., M + N$, is $\mathcal{S}^{(n)} = \mathcal{S}_+^{(n)} \cup \mathcal{S}_-^{(n)}$, where, for $n = M + 1, M + 2, ..., M + N$,

$$\begin{aligned} \mathcal{S}_+^{(n)} &= \{(k, i, j), k = K, K-1, ..., 1, i = 1, ..., m_b, j = 1, ..., m_s\}; \\ \mathcal{S}_-^{(n)} &= \{(k, i, j), k = 0, -1, ..., -K+1, i = 1, ..., m_b, j = 1, ..., m_s\}, \end{aligned} \tag{4.2}$$

and, for $n = 1, 2, ..., M$,

$$\begin{aligned} \mathcal{S}_+^{(n)} &= \{(k, i, j), k = K-1, K-2, ..., 0, i = 1, ..., m_b, j = 1, ..., m_s\}; \\ \mathcal{S}_-^{(n)} &= \{(k, i, j), k = -1, -2, ..., -K, i = 1, ..., m_b, j = 1, ..., m_s\}. \end{aligned} \tag{4.3}$$

Note that, for all layers, $\mathcal{S}_0^{(n)} = \emptyset$. In any state in $\mathcal{S}_+^{(n)}$, $X(t)$ is increasing. In any state in $\mathcal{S}_-^{(n)}$, $X(t)$ is decreasing. According to $\mathcal{S}_+^{(n)}$ and $\mathcal{S}_-^{(n)}$, the transition rate matrix $Q^{(n)}$ of the underlying Markov chain is partitioned into four blocks $Q_{++}^{(n)}$, $Q_{+-}^{(n)}$, $Q_{-+}^{(n)}$, and $Q_{--}^{(n)}$ (see equation (2.2)). Putting the four blocks together, which makes it easier to see the change of $s(t)$, we obtain, for $n = M+1, M+2, ..., M+N$,

$$Q^{(n)} = \left(\begin{array}{ccccc|ccccc} I\otimes\hat{D}_0 & I\otimes\hat{D}_1 & ... & I\otimes\hat{D}_{K-1} & I\otimes\hat{D}_K & & & & \\ & I\otimes\hat{D}_0 & ... & I\otimes\hat{D}_{K-2} & I\otimes\hat{D}_{K-1} & I\otimes\hat{D}_K & & & \\ & & \ddots & \vdots & \vdots & \vdots & \ddots & & \\ & & & I\otimes\hat{D}_0 & I\otimes\hat{D}_1 & I\otimes\hat{D}_2 & ... & I\otimes\hat{D}_K \\ \hline \mathcal{D}_{K,n}\otimes I & \mathcal{D}_{K-1,n}\otimes I & ... & \mathcal{D}_{1,n}\otimes I & \bar{\mathcal{D}}_n\otimes I & & & & \\ & \mathcal{D}_{K,n}\otimes I & ... & \mathcal{D}_{2,n}\otimes I & \mathcal{D}_{1,n}\otimes I & \bar{\mathcal{D}}_n\otimes I & & & \\ & & \ddots & \vdots & \vdots & \vdots & \ddots & & \\ & & & \mathcal{D}_{K,n}\otimes I & \mathcal{D}_{K-1,n}\otimes I & \mathcal{D}_{K-2,n}\otimes I & ... & \bar{\mathcal{D}}_n\otimes I \end{array}\right), \tag{4.4}$$

where $\mathcal{D}_{k,n} = D_k(\eta_{k,n-M} + ... + \eta_{k,N})$, $\bar{\mathcal{D}}_n = D - \sum_{k=1}^K D_k(\eta_{k,n-M} + ... + \eta_{k,N})$, and $\otimes$ is for Kronecker product for matrices defined as, for matrices $A = (a_{i,j})$ and B, $A \otimes B = (a_{i,j}B)$; and, for $n = 1, 2, ..., M$,

$$Q^{(n)} = \left(\begin{array}{ccccc|ccccc} I\otimes\bar{\hat{\mathcal{D}}}_n & I\otimes\hat{\mathcal{D}}_{1,n} & ... & I\otimes\hat{\mathcal{D}}_{K-1,n} & I\otimes\hat{\mathcal{D}}_{K,n} & & & & \\ & I\otimes\bar{\hat{\mathcal{D}}}_n & ... & I\otimes\hat{\mathcal{D}}_{K-2,n} & I\otimes\hat{\mathcal{D}}_{K-1,n} & I\otimes\hat{\mathcal{D}}_{K,n} & & & \\ & & \ddots & \vdots & \vdots & \vdots & \ddots & & \\ & & & I\otimes\bar{\hat{\mathcal{D}}}_n & I\otimes\hat{\mathcal{D}}_{1,n} & I\otimes\hat{\mathcal{D}}_{2,n} & ... & I\otimes\hat{\mathcal{D}}_{K,n} \\ \hline D_K\otimes I & D_{K-1}\otimes I & ... & D_1\otimes I & D_0\otimes I & & & & \\ & D_K\otimes I & ... & D_2\otimes I & D_1\otimes I & D_0\otimes I & & & \\ & & \ddots & \vdots & \vdots & \vdots & \ddots & & \\ & & & D_K\otimes I & D_{K-1}\otimes I & D_{K-2}\otimes I & ... & D_0\otimes I \end{array}\right), \tag{4.5}$$

where $\hat{\mathcal{D}}_{k,n} = \hat{D}_k(\hat{\eta}_{k,M-n+1} + ... + \hat{\eta}_{k,M})$, $\bar{\hat{\mathcal{D}}}_n = \hat{D} - \sum_{k=1}^K \hat{D}_k(\hat{\eta}_{k,M-n+1} + ... + \hat{\eta}_{k,M})$.

3. Within Border M (i.e., $l_M = 0$), the underlying Markov chain has states $\{1, ..., m_b\} \times \{1, ..., m_s\}$, and its transition rate matrices are

$$\begin{aligned} Q_{bb}^{(M)} &= \left(D_0 + \sum_{k=1}^K \dot{\eta}_{k,0} D_k\right) \otimes I + I \otimes \left(\hat{D}_0 + \sum_{k=1}^K \dot{\hat{\eta}}_{k,0} \hat{D}_k\right); \\ Q_{b+}^{(M)} &= ((1-\dot{\eta}_{K,0})D_K \otimes I, \ \ldots, \ (1-\dot{\eta}_{1,0})D_1 \otimes I); \\ Q_{b-}^{(M)} &= \left(I \otimes (1-\dot{\hat{\eta}}_{1,0})\hat{D}_1, \ \ldots, \ I \otimes (1-\dot{\hat{\eta}}_{K,0})\hat{D}_K\right). \end{aligned} \tag{4.6}$$

4. The transition probabilities of approaching Border M are given by, in matrix form, $P^{(M)}_{-b+}=0$; $P^{(M)}_{+b-}=0$;

$$P^{(M)}_{-bb}=\begin{pmatrix} I \\ \dot{\hat{\eta}}_{1,0}I \\ \dot{\hat{\eta}}_{2,0}I \\ \vdots \\ \dot{\hat{\eta}}_{K-1,0}I \end{pmatrix};\quad P^{(M)}_{-b-}=\begin{pmatrix} 0 & & & & \\ (1-\dot{\hat{\eta}}_{1,0})I & 0 & & & \\ & (1-\dot{\hat{\eta}}_{2,0})I & 0 & & \\ & & \ddots & \ddots & \\ & & & (1-\dot{\hat{\eta}}_{K-1,0})I & 0 \end{pmatrix};$$
$$P^{(M)}_{+bb}=\begin{pmatrix} \dot{\eta}_{K-1,0}I \\ \dot{\eta}_{K-2,0}I \\ \vdots \\ \dot{\eta}_{1,0}I \\ I \end{pmatrix};\quad P^{(M)}_{+b+}=\begin{pmatrix} 0 & (1-\dot{\eta}_{K-1,0})I & & & \\ & 0 & (1-\dot{\eta}_{K-2,0})I & & \\ & & 0 & \ddots & \\ & & & \ddots & (1-\dot{\eta}_{1,0})I \\ & & & & 0 \end{pmatrix}. \tag{4.7}$$

Note that the batch size of buyer or seller does not change when crossing Border M if the buy or order has a positive abandonment time, but the state spaces of $\phi(t)$ are different between layers above and below Border M, so that we need to shift the underlying states using the upper shift matrix $P^{(M)}_{+b+}$ for up-crossing and lower shift matrix $P^{(M)}_{-b-}$ for down-crossing. Also note that there is no reflection for Border M since that $X(t)$ approaching zero means that there is no buyer order, if $X(t)>0$, or no seller order, if $X(t)<0$.

5. All borders, except for Border M, have no state. The probabilities of approaching Border $n+M$, for $1\le n\le N-1$, are $P^{(n+M)}_{-b+}=0$, $P^{(n+M)}_{-bb}=0$, $P^{(n+M)}_{-b-}=I$, $P^{(n+M)}_{+bb}=0$,

$$P^{(n+M)}_{+b-}=\begin{pmatrix} \frac{\dot{\eta}_{K,n}}{\dot{\eta}_{K,n}+\ldots+\dot{\eta}_{K,N}}I & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\dot{\eta}_{1,n}}{\dot{\eta}_{1,n}+\ldots+\dot{\eta}_{1,N}}I & 0 & \ldots & 0 \end{pmatrix};$$
$$P^{(n+M)}_{+b+}=\begin{pmatrix} \frac{\dot{\eta}_{K,n+1}+\ldots+\dot{\eta}_{K,N}}{\dot{\eta}_{K,n}+\dot{\eta}_{K,n+1}+\ldots+\dot{\eta}_{K,N}}I & 0 & \ldots & 0 \\ 0 & \frac{\dot{\eta}_{K-1,n+1}+\ldots+\dot{\eta}_{K-1,N}}{\dot{\eta}_{K-1,n}+\dot{\eta}_{K-1,n+1}+\ldots+\dot{\eta}_{K-1,N}}I & \ldots & 0 \\ \ddots & \ddots & \ddots & \ddots \\ 0 & \ldots & 0 & \frac{\dot{\eta}_{1,n+1}+\ldots+\dot{\eta}_{1,N}}{\dot{\eta}_{1,n}+\dot{\eta}_{1,n+1}+\ldots+\dot{\eta}_{1,N}}I \end{pmatrix}. \tag{4.8}$$

The probabilities of approaching Border n, for $1\le n\le M-1$, are $P^{(n)}_{+b-}=0$, $P^{(n)}_{+bb}=0$, $P^{(n)}_{+b+}=I$,

$P_{-bb}^{(n)} = 0,$

$$P_{-b-}^{(n)} = \begin{pmatrix} \frac{\dot{\hat{\eta}}_{1,M-n+1}+\ldots+\dot{\hat{\eta}}_{1,M}}{\dot{\hat{\eta}}_{1,M-n}+\ldots+\dot{\hat{\eta}}_{1,M}} I & 0 & \ldots & 0 \\ 0 & \frac{\dot{\hat{\eta}}_{2,M-n+1}+\ldots+\dot{\hat{\eta}}_{2,M-n}}{\dot{\hat{\eta}}_{2,M-n}+\ldots+\dot{\hat{\eta}}_{2,M}} I & \ldots & 0 \\ \ddots & \ddots & \ddots & \ddots \\ 0 & \ldots & 0 & \frac{\dot{\hat{\eta}}_{K,M-n+1}+\ldots+\dot{\hat{\eta}}_{K,M}}{\dot{\hat{\eta}}_{K,M-n}+\ldots+\dot{\hat{\eta}}_{K,M}} I \end{pmatrix}; \tag{4.9}$$

$$P_{-b+}^{(n)} = \begin{pmatrix} 0 & \ldots & 0 & \frac{\dot{\hat{\eta}}_{1,M-n}}{\dot{\hat{\eta}}_{1,M-n}+\ldots+\dot{\hat{\eta}}_{1,M}} I \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \ldots & 0 & \frac{\dot{\hat{\eta}}_{K,M-n}}{\dot{\hat{\eta}}_{K,M-n}+\ldots+\dot{\hat{\eta}}_{K,M}} I \end{pmatrix}.$$

Under the conditions given in equation (3.1), the multi-layer $MMFF$ process can reach steady-state. We can find the stationary joint distribution of the process by Theorem 1 in Section 2 and the computational steps in Appendix A.

Remark: In the above construction, the abandonment time distributions of a buyer (seller) are used indirectly. Instead of sampling abandonment times at arrivals or other state changing epochs, abandonment decisions, to continue to wait or to abandon the queue, are made at all possible abandonment epochs, by using conditional distributions of the abandonment times. The two models are equivalent probabilistically since the distributions of the actual abandonment times for the latter model are the same as that of the original model. Further, for notational convenience, we assume that $\max\{\tilde{l}_n : P\{\dot{\tau}_k = \tilde{l}_n\} \neq 0\} \geq \max\{\tilde{l}_n : P\{\dot{\tau}_{k+1} = \tilde{l}_n\} \neq 0\} \geq \max\{\tilde{l}_n : P\{\tau_{k+1} = \tilde{l}_n\} \neq 0\}$, and $\max\{\hat{l}_n : P\{\dot{\hat{\tau}}_k = \hat{l}_n\} \neq 0\} \geq \max\{\hat{l}_n : P\{\dot{\hat{\tau}}_{k+1} = \hat{l}_n\} \neq 0\} \geq \max\{\hat{l}_n : P\{\hat{\tau}_{k+1} = \hat{l}_n\} \neq 0\}$ to ensure the existence of conditional distributions of abandonment times for re-sampling. This assumption is intuitive and realistic as a buyer/seller becomes more patient when its order size becomes smaller and when it becomes the head of the queue. when the conditions are not satisfied, the multi-layer MMFF process for the age process can still be constructed. Yet the construction process becomes even more tedious.

5 Queueing Quantities

In this section, we use the stationary joint distribution given in Theorem 1 and some basic quantities of the $MMFF$ processes to find queueing quantities for the double-sided queue. As the two sides of the model are symmetric, we mainly present the results related to the buyer side. The results for the seller side can be obtained similarly.

5.1 Stationary Joint Distributions of Age Processes

Stationary joint distributions of age processes are the starting point for finding other queueing quantities. Let $\mathbf{f}(x)$ be the stationary joint density function of the age process $\{(a(t), s(t), I_b(t), I_s(t)), t \geq 0\}$, which is a row vector of size Km_bm_s. Let $\mathbf{f}^{(n)}(x) = \mathbf{f}(x)$, if $l_{n-1} < x < l_n$, for $n = 1, 2, ..., M+N$. By censoring out the fictitious periods in the $MMFF$ process, we obtain

Theorem 2. *(Wu and He (2021)) Under the assumptions given in equation (3.1), the age process* $\{(a(t), s(t), I_b(t), I_s(t)), t \geq 0\}$ *has a stationary joint distribution, and its density function can be expressed as*

$$\begin{aligned}
&P\{a(t) = 0\} = \hat{\mathbf{p}}^{(M)}\mathbf{e};\\
&\mathbf{f}^{(n)}(x) = \mathbf{v}_+^{(n)} e^{\mathcal{K}^{(n)}(x-l_{n-1})}\Psi^{(n)} + \mathbf{v}_-^{(n)} e^{\widehat{\mathcal{K}}^{(n)}(l_n-x)}, \text{ for } l_{n-1} < x \leq l_n, \ n = 1, ..., M;\\
&\mathbf{f}^{(n)}(x) = \mathbf{v}_+^{(n)} e^{\mathcal{K}^{(n)}(x-l_{n-1})} + \mathbf{v}_-^{(n)} e^{\widehat{\mathcal{K}}^{(n)}(l_n-x)}\widehat{\Psi}^{(n)}, \text{for } l_{n-1} \leq x < l_n, \ n = M+1, ..., M+N,
\end{aligned} \tag{5.1}$$

where $\hat{\mathbf{p}}^{(M)} = \mathbf{p}^{(M)}/\hat{c}_{norm}$, $\mathbf{v}_+^{(n)} = \mathbf{u}_+^{(n)}/\hat{c}_{norm}$, $\mathbf{v}_-^{(n)} = \mathbf{u}_-^{(n)}/\hat{c}_{norm}$ *and* $\mathbf{v}_+^{(1)} = 0$, $\mathbf{v}_-^{(M+N)} = 0$, *and* $\mathbf{p}^{(M)}$, $\mathbf{u}_+^{(n)}$, *and* $\mathbf{u}_-^{(n)}$ *are given in Theorem 1. The normalization factor is*

$$\hat{c}_{norm} = \mathbf{p}^{(M)}\mathbf{e} + \sum_{n=1}^{M}\left(\mathbf{u}_+^{(n)}\mathcal{L}_{l_{n-1},l_n}^{\mathcal{K}^{(n)}}\Psi^{(n)} + \mathbf{u}_-^{(n)}\widetilde{\mathcal{L}}_{l_{n-1},l_n}^{\widehat{\mathcal{K}}^{(n)}}\right)\mathbf{e} + \sum_{n=M+1}^{M+N}\left(\mathbf{u}_+^{(n)}\mathcal{L}_{l_{n-1},l_n}^{\mathcal{K}^{(n)}} + \mathbf{u}_-^{(n)}\widetilde{\mathcal{L}}_{l_{n-1},l_n}^{\widehat{\mathcal{K}}^{(n)}}\widehat{\Psi}^{(n)}\right)\mathbf{e}, \tag{5.2}$$

where $\mathcal{L}_{l_{n-1},l_n}^{\mathcal{K}^{(n)}}$ *and* $\widetilde{\mathcal{L}}_{l_{n-1},l_n}^{\widehat{\mathcal{K}}^{(n)}}$ *are defined in equation (A.7).*

Proof. The results are obtained by censoring out the fictitious periods (i.e., $\{(x, j) : l_{n-1} < x < l_n, j \in \mathcal{S}_-^{(n)}, n = M+1, M+2, ..., M+N\} \cup \{(x, j) : l_{n-1} < x < l_n, j \in \mathcal{S}_+^{(n)}, n = 1, 2, ...M\}$) in the multi-layer $MMFF$ process. Theorem 1 leads to the desired results. □

Remark: For notational convenience, the steady-state quantities are sometimes denoted by notations with the time variable t, (e.g., $a(t)$ denotes the age in steady-state).

Next, the joint distribution related to the buyer age can be obtained immediately. Let $\mathbf{f}_B^{(n)}(x)$, for $l_{n-1} < x < l_n$ and $n = M+1, ..., M+N$ be the stationary joint density functions of the age process of buyers.

Corollary 1. *The stationary joint distribution for the age process of buyers is*

$$\begin{aligned} P\{a_B(t)=0\} &= P\{a(t)\le 0\} = \hat{\mathbf{p}}^{(M)}\mathbf{e} + \sum_{n=1}^{M}\left(\mathbf{v}_+^{(n)}\mathcal{L}_{l_{n-1},l_n}^{\mathcal{K}^{(n)}}\Psi^{(n)} + \mathbf{v}_-^{(n)}\widetilde{\mathcal{L}}_{l_{n-1},l_n}^{\widehat{\mathcal{K}}^{(n)}}\right)\mathbf{e}; \\ \mathbf{f}_B^{(n)}(x) &= \mathbf{v}_+^{(n)}e^{\mathcal{K}^{(n)}(x-l_{n-1})} + \mathbf{v}_-^{(n)}e^{\widehat{\mathcal{K}}^{(n)}(l_n-x)}\widehat{\Psi}^{(n)}, \\ &\qquad \textit{for } l_{n-1}\le x<l_n,\ n=M+1,...,M+N. \end{aligned} \tag{5.3}$$

Similarly, queueing quantities for the remaining batch size of the buyer at the head of the queue can also be obtained by fixing the underlying states $s(t)$ at a specific value k. Let $\mathbf{f}_B^{(n)}(k,x)$ be the joint density function of $(a(t), s(t)=k, I_{(b)}(t), I_{(s)}(t))$, for $k=1,2,...,K$, $l_{n-1}<x<l_n$, and $n=M+1,...,M+N$, we have

$$\mathbf{f}_B^{(n)}(k,x) = \mathbf{f}_B^{(n)}(x)(\mathbf{e}(k)\otimes I), \tag{5.4}$$

where $\mathbf{e}(k)$ is a column logical vector of size K with only the k-th element being 1.

Let $p_B(k)$ be the probability that k orders remained for the buyer at the head of the queue at an arbitrary time. We have, for $k=1,...,K$,

$$p_B(k) = \sum_{n=M+1}^{M+N}\left(\mathbf{v}_+^{(n)}\mathcal{L}_{l_{n-1},l_n}^{\mathcal{K}^{(n)}} + \mathbf{v}_-^{(n)}\widetilde{\mathcal{L}}_{l_{n-1},l_n}^{\widehat{\mathcal{K}}^{(n)}}\widehat{\Psi}^{(n)}\right)(\mathbf{e}(k)\otimes\mathbf{e}). \tag{5.5}$$

Let p_B be the probability that there is a buyer queue (i.e., $a(t)>0$). We have

$$p_B = p_B(1) + p_B(2) + ... + p_B(K). \tag{5.6}$$

Using the joint density functions of the age processes, in the following sections, we derive a number of queueing quantities for buyer orders (Subsection 5.2) and for buyers (Subsection 5.3). To distinguish the two sets of quantities, we use notations with superscript "o" for order level queueing quantities in Subsection 5.2.

5.2 Order Level Queueing Quantities

5.2.1 Matching rates and fill rates of orders

Let ω^o be the number of matched orders per unit time, to be called *the matching rate* of the orders. Note that the matching rates of buyer orders and seller orders are the same.

Proposition 1. *The matching rate of the orders is*

$$\begin{aligned}\omega^o &= \frac{1}{\hat{c}_{norm}}\sum_{n=1}^{M}\left(\left(\int_{l_{n-1}}^{l_n}\boldsymbol{\pi}_+^{(n)}(x)\mathrm{d}x\right)\boldsymbol{\delta}(-,+,n)+\left(\int_{l_{n-1}}^{l_n}\boldsymbol{\pi}_-^{(n)}(x)\mathrm{d}x\right)\boldsymbol{\delta}(-,-)\right)\\ &\quad+\frac{1}{\hat{c}_{norm}}\sum_{n=M+1}^{M+N}\left(\left(\int_{l_{n-1}}^{l_n}\boldsymbol{\pi}_+^{(n)}(x)\mathrm{d}x\right)\boldsymbol{\delta}(+,+)+\left(\int_{l_{n-1}}^{l_n}\boldsymbol{\pi}_-^{(n)}(x)\mathrm{d}x\right)\boldsymbol{\delta}(+,-,n)\right),\end{aligned} \tag{5.7}$$

where

$$\begin{aligned}\textstyle\int_{l_{n-1}}^{l_n}\boldsymbol{\pi}_+^{(n)}(x)\mathrm{d}x &= \mathbf{u}_+^{(n)}\mathcal{L}_{l_{n-1},l_n}^{\mathcal{K}^{(n)}}+\mathbf{u}_-^{(n)}\widetilde{\mathcal{L}}_{l_{n-1},l_n}^{\widehat{\mathcal{K}}^{(n)}}\widehat{\Psi}^{(n)};\\ \textstyle\int_{l_{n-1}}^{l_n}\boldsymbol{\pi}_-^{(n)}(x)\mathrm{d}x &= \mathbf{u}_+^{(n)}\mathcal{L}_{l_{n-1},l_n}^{\mathcal{K}^{(n)}}\Psi^{(n)}+\mathbf{u}_-^{(n)}\widetilde{\mathcal{L}}_{l_{n-1},l_n}^{\widehat{\mathcal{K}}^{(n)}},\end{aligned} \tag{5.8}$$

and

$$\begin{aligned}\boldsymbol{\delta}(-,-) &= \left(\left(A\otimes\mathbf{e}\mathbf{e}^T\right)\odot\left(\mathbf{e}\otimes(D_1,D_2,...,D_K)\otimes I\right)\right)\mathbf{e};\\ \boldsymbol{\delta}(+,-,n) &= \left(\left(\hat{A}\otimes\mathbf{e}\mathbf{e}^T\right)\odot\left(\mathbf{e}\otimes(\mathcal{D}_{1,n},\mathcal{D}_{2,n},...,\mathcal{D}_{K,n})\otimes I\right)\right)\mathbf{e};\\ \boldsymbol{\delta}(-,+,n) &= \left(\left(\tilde{\hat{A}}\otimes\mathbf{e}\mathbf{e}^T\right)\odot\left(\mathbf{e}\otimes(I\otimes\hat{\mathcal{D}}_{1,n},I\otimes\hat{\mathcal{D}}_{2,n},...,I\otimes\hat{\mathcal{D}}_{K,n})\right)\right)\mathbf{e};\\ \boldsymbol{\delta}(+,+) &= \left(\left(\tilde{A}\otimes\mathbf{e}\mathbf{e}^T\right)\odot\left(\mathbf{e}\otimes(I\otimes\hat{D}_1,I\otimes\hat{D}_2,...,I\otimes\hat{D}_K)\right)\right)\mathbf{e},\end{aligned} \tag{5.9}$$

where $\odot$ is the Hadamard product, $A=(a_{i,j})$ is a square matrix of order K with element $a_{i,j}=min(i,j)$ and $\tilde{A}=(\tilde{a}_{i,j})$ is an upside down flipped A (i.e., $\tilde{a}_{i,j}=a_{K-i+1,j}$), $\hat{A}=(\hat{a}_{i,j})$ is a square matrix of order K with element $\hat{a}_{i,j}=min(i-1,j)$ and $\tilde{\hat{A}}$ is an upside down flipped $\hat{A}$.

(Note: For notational simplicity, in the rest of the paper we shall use the integrals on the left hand sides, instead of the explicit expressions on the right hand sides, in equation (5.8).)

Proof. At an arbitrary time, the matching of orders can take place only if a buyer or seller arrives. Suppose that the state of the $MMFF$ process is (x,s,i_b,i_s) at an arbitrary time. We consider four cases: i) $x>0$ and $s>0$; ii) $x>0$ and $s<0$; iii) $x<0$ and $s<0$; and iv) $x<0$ and $s>0$.

i) If $x>0$ and $s>0$, there are buyers in queue and matching will take place when the next seller arrives. Given the phase i_s of the seller arrival process, the arrival rate of a seller of batch size k is given by $(\hat{D}_k\mathbf{e})_{i_s}$. Consequently, the conditional matching rate for state (x,s,i_b,i_s) is

$$\delta(x,s,i_b,i_s)=\sum_{k=1}^{K}\min\{s,k\}(\hat{D}_k\mathbf{e})_{i_s}. \tag{5.10}$$

Let $\boldsymbol{\delta}(+,+)$ be the vector with elements $\delta(x,s,i_b,i_s)$ defined in equation (5.10). Recall that the stationary distribution of the $MMFF$ process is given by $\left(\boldsymbol{\pi}_+^{(n)}(x),\boldsymbol{\pi}_-^{(n)}(x)\right)$. Then the matching rate for case i) at $X(t)=x$ is given by $\boldsymbol{\pi}_+^{(n)}(x)\boldsymbol{\delta}(+,+)$. Summing up over x from l_{n-1} to l_n for $n=M+1,M+2,...,M+N$, the matching rate for case i) is obtained. With the explicit expressions of $\int_{l_{n-1}}^{l_n}\boldsymbol{\pi}_+^{(n)}(x)\mathrm{d}x$ and $\boldsymbol{\delta}(+,+)$, the

matching rate for this case can be written as the product of the first line in equation (5.8) and the last line in equation (5.9), for $n = M+1, M+2, ..., M+N$.

ii) If $x > 0$ and $s < 0$, this is a fictitious time period. There is possibly a seller waiting to be matched by the next arriving buyer. Different from case i), in the real queueing system, the buyer has arrived earlier and survived up to the current time epoch, i.e., the buyer has an age of x. The probability for the buyer to have an age x is $\eta_{k,n-M} + ... + \eta_{k,N}$, which depends on the batch size of the buyer too. For the case with batch size k of the buyer, $l_{n-1} < x < l_n$, and the phase i_b of the buyer arrival process, the conditional arrival rate is given by $(\mathcal{D}_{k,n}\mathbf{e})_{i_b}$. Consequently, the conditional matching rate for state (x, s, i_b, i_s) is, for $l_{n-1} < x < l_n$,

$$\delta(x, n, s, i_b, i_s) = \sum_{k=1}^{K} \min\{-s, k\}(\mathcal{D}_{k,n}\mathbf{e})_{i_b}. \tag{5.11}$$

Let $\boldsymbol{\delta}(+, -, n)$ be the vector with elements $\delta(x, n, s, i_b, i_s)$. Then the matching rate for case ii) at $X(t) = x$ is given by $\boldsymbol{\pi}_{-}^{(n)}(x)\boldsymbol{\delta}(+, -, n)$. Summing up over x from l_{n-1} to l_n, for $n = M+1, M+2, ..., M+N$, the matching rate for case ii) is obtained. With the explicit expressions of $\int_{l_{n-1}}^{l_n} \boldsymbol{\pi}_{-}^{(n)}(x)\mathrm{d}x$ and $\boldsymbol{\delta}(+, -, n)$, the matching rate for this case can be written as the product of the second line in equation (5.8) and the second line in equation (5.9), for $n = M+1, M+2, ..., M+N$.

Cases iii) and iv) can be obtained similarly.

iii) If $x < 0$ and $s < 0$, there are sellers in the queue and matching will take place when the next buyer arrives. Given the phase i_b of the buyer arrival process, the arrival rate of a buyer of batch size k is given by $(D_k\mathbf{e})_{i_b}$. Consequently, the conditional matching rate for state (x, s, i_b, i_s) is

$$\delta(x, s, i_b, i_s) = \sum_{k=1}^{K} \min\{-s, k\}(D_k\mathbf{e})_{i_b}. \tag{5.12}$$

Let $\boldsymbol{\delta}(-, -)$ be the vector with elements $\delta(x, s, i_b, i_s)$ for this case. With the explicit expressions of $\int_{l_{n-1}}^{l_n} \boldsymbol{\pi}_{-}^{(n)}(x)\mathrm{d}x$ and $\boldsymbol{\delta}(-, -)$, the matching rate for this case can be written as the product of the second line in equation (5.8) and the first line in equation (5.9), for $n = 1, 2, ..., N$.

iv) If $x < 0$ and $s > 0$, there is possibly a buyer to be matched by arriving sellers. Different from case iii), in the real queueing system, those sellers have arrived earlier and survived up to the current time epoch, i.e., the seller has an age of x. The probability for that depends on the batch size of the seller. For the case with batch size k of the seller, $l_{n-1} < x < l_n$, and the phase i_s of the seller arrival process, the arrival rate

is given by $(\hat{\mathcal{D}}_{k,n}\mathbf{e})_{i_s}$. Consequently, the matching rate for state (x, s, i_b, i_s) is, for $l_{n-1} < x < l_n$,

$$\delta(x, n, s, i_b, i_s) = \sum_{k=1}^{K} \min\{s, k\}(\hat{\mathcal{D}}_{k,n}\mathbf{e})_{i_s}. \tag{5.13}$$

Let $\boldsymbol{\delta}(-, +, n)$ be the vector with elements $\delta(x, n, s, i_b, i_s)$. With the explicit expressions of $\int_{l_{n-1}}^{l_n} \boldsymbol{\pi}_+^{(n)}(x)\mathrm{d}x$ and $\boldsymbol{\delta}(-, +, n)$, the matching rate for this case can be written as the product of the first line in equation (5.8) and the third line in equation (5.9), for $n = 1, 2, ..., N$.

Combining the four cases and normalizing the coefficient vectors, the matching rate is obtained. □

Let $p^o_{B,F}$ be the fill rate of buyer orders. Note that fill rate is a commonly used term in inventory and supply chain systems and is defined as the fraction of orders being matched. Recall that λ is the arrival rate of buyer orders.

Corollary 2. *We have*

$$p^o_{B,F} = \frac{\omega^o}{\lambda}. \tag{5.14}$$

The loss probability of buyer orders is $p^o_{B,L} = 1 - p^o_{B,F}$.

Loss probability $p^o_{B,L}$ can be decomposed into two parts based on the location of the buyers in the queue: i) loss probability $p^o_{BL,1}$ of buyer orders at the head of the queue; and ii) loss probability $p^o_{BL,>1}$ of buyer orders before reaching the head of the queue. Then we have

Proposition 2.

$$p^o_{BL,1} = \frac{1}{\lambda} \sum_{n=M+1}^{M+N-1} \mathbf{f}^{(n)}(l_n) \left(\left(\sum_{k=1}^{K} \frac{k\dot{\eta}_{k,n-M}}{\sum_{m=n}^{M+N} \dot{\eta}_{k,m-M}} \mathbf{e}(K-k+1) \right) \otimes \mathbf{e} \right) + \frac{1}{\lambda} \hat{\mathbf{p}}^{(M)} \left(\left(\sum_{k=1}^{K} k\dot{\eta}_{k,0} D_k \mathbf{e} \right) \otimes \mathbf{e} \right). \tag{5.15}$$

and $p^o_{BL,>1} = p^o_{B,L} - p^o_{BL,1}$, *and* $\mathbf{e}(K-k+1)$ *is a column logical vector with only the* $(K-k+1)$*-th element being one.*

Proof. The abandonment of the buyers at the head of the queue can only happen when $X(t) = 0$ or $X(t) > 0$ and fluid increases, therefore, we can use the age process to find this probability.

For $X(t) = 0$, in the corresponding age process, the probability of $a(t) = 0$ is $\hat{\mathbf{p}}^{(M)}$. Given state (i_b, i_s), the

abandonment rate equals the total arrival rates of all orders with buyers with 0 patience time $\left(\left(\sum_{k=1}^{K} k\dot{\eta}_{k,0} D_k \mathbf{e}\right) \otimes \mathbf{e}\right)$. Divided by λ leads to the second item of the right-hand side of equation (5.15) (i.e., $\frac{1}{\lambda}\hat{\mathbf{p}}^{(M)} \left(\left(\sum_{k=1}^{K} k\dot{\eta}_{k,0} D_k \mathbf{e}\right) \otimes \mathbf{e}\right)$).

For $X(t) > 0$ and fluid increases, in the corresponding age process, the density function for $a(t) = l_n$ is $\mathbf{f}^{(n)}(l_n) = \mathbf{v}_+^{(n)} e^{\mathcal{K}^{(n)}(l_n - l_{n-1})} + \mathbf{v}_-^{(n)} \widehat{\Psi}^{(n)}$. Given $a(t) = l_n$, $s(t) = k$, and state (i_b, i_s), the probability that the buyer at the head of the queue abandons the system at l_n is $\frac{\dot{\eta}_{k,n-M}}{\sum_{m=n}^{M+N} \dot{\eta}_{k,m-M}}$. Multiplying the size k with the abandonment probability of the buyer at the head of the queue with batch size k, and then combining with the joint density function, we get the abandonment rate of buyer orders when $a(t) = l_n$. Considering all possible abandonment points from $M+1$ to $M+N-1$, the total abandonment rate of buyer orders divided by λ leads to the first item on the right-hand side of equation (5.15). □

5.2.2 Sojourn times of orders

In this subsection, we present the sojourn times of buyer orders. We consider four types of sojourn times, namely, the sojourn times of i) filled orders; ii) lost orders; iii) lost orders at the head of the queue; and iv) lost orders before reaching the head of the queue.

Proposition 3. *The distribution of sojourn time $W^o_{B,F}$ of filled buyer orders is*

$$\begin{aligned} P\{W^o_{B,F} = 0\} &= \frac{1}{\lambda p_{B,F} \hat{c}_{norm}} \sum_{n=1}^{M} \left(\int_{l_{n-1}}^{l_n} \boldsymbol{\pi}_-^{(n)}(x) \mathrm{d}x \boldsymbol{\delta}(-,-) + \int_{l_{n-1}}^{l_n} \boldsymbol{\pi}_+^{(n)}(x) \mathrm{d}x \boldsymbol{\delta}(-,+,n) \right); \\ \frac{\mathrm{d}P\{W^o_{B,F} < x\}}{\mathrm{d}x} &= \frac{1}{\lambda p_{B,F} \hat{c}_{norm}} \left(\boldsymbol{\pi}_+^{(n)}(x) \boldsymbol{\delta}(+,+) + \boldsymbol{\pi}_-^{(n)}(x) \boldsymbol{\delta}(+,-,n) \right), \\ & \qquad \text{for } l_{n-1} \le x < l_n,\ n = M+1, ..., M+N, \end{aligned} \tag{5.16}$$

where the integrals are given in equation (5.8). The distribution of sojourn time $W^o_{BL,1}$ of lost orders at the head of the queue is given by

$$P\{W^o_{BL,1} = 0\} = \frac{1}{\lambda p^o_{BL,1}} \hat{\mathbf{p}}^{(M)} \left(\left(\sum_{k=1}^{K} k\dot{\eta}_{k,0} D_k \mathbf{e} \right) \otimes \mathbf{e} \right), \tag{5.17}$$

and, for $n = M+1, M+2, ..., M+N-1$,

$$P\{W^o_{BL,1} = l_n\} = \frac{\left(\mathbf{v}_+^{(n)} e^{\mathcal{K}^{(n)}(l_n - l_{n-1})} + \mathbf{v}_-^{(n)} \widehat{\Psi}^{(n)} \right)}{p^o_{BL,1} \lambda} \left(\left(\sum_{k=1}^{K} \frac{k\dot{\eta}_{k,n-M}}{\dot{\eta}_{k,n-M} + ... + \dot{\eta}_{k,N}} \mathbf{e}(K-k+1) \right) \otimes \mathbf{e} \right). \tag{5.18}$$

The distribution of sojourn time $W^o_{BL,>1}$ of lost orders before reaching the head of the queue, we have, for

$n = M, M+1, ..., M+N-1$,

$$P\{W^o_{BL,>1} = l_n\} = \frac{1}{\lambda p^o_{BL,>1}\hat{c}_{norm}} \left(\sum_{m=n+1}^{M+N} \int_{l_{m-1}}^{l_m} \boldsymbol{\pi}_-^{(m)}(x)\mathrm{d}x \left(\mathbf{e} \otimes \left(\sum_{k=1}^{K} k\eta_{k,n-M} D_k \mathbf{e} \right) \otimes \mathbf{e} \right) \right). \quad (5.19)$$

Proof. Since matching can happen during fictitious time periods in the multi-layer $MMFF$ processes, we need to use the stationary joint distribution of the multi-layer $MMFF$ process to find the sojourn time distribution of filled buyer orders.

For $W^o_{B,F} = 0$, matching can only happen when there is a seller queue and a buyer arrives, the ratio of the probability that buyer orders get filled without waiting (i.e., cases iii) and iv) in the proof of Proposition 1) and $p^o_{B,F}$ leads to the expression for $P\{W^o_{B,F} = 0\}$.

For $W^o_{B,F} > 0$, we condition on the fluid level $X(t)$. Similar to cases i) and ii) in the proof of Proposition 1, we change the total probability that $X(t) > 0$ to the density function of $X(t)$ and then divided by $\hat{c}_{norm}$ for normalization, which leads to the desired result.

For the sojourn time distribution of lost orders when the buyer is at the head of the queue, we can use the age process similar to Proposition 2. The distribution of $W^o_{BL,1}$ can be obtained from the proof of Proposition 2 easily.

For the distribution of $W^o_{BL,>1}$, we need to use the stationary joint distribution of the multi-layer $MMFF$ process $\{(X(t), \phi(t), t \geq 0\}$. When $X(t)$ is decreasing and there is a buyer arrival, which may take place when $X(t) \geq 0$, the arriving buyer will abandon the queue in the future with probability $\eta_{k,1} + ... + \eta_{k,n-1}$ if $l_{n-1} < x < l_n$ and the batch size of the buyer is k. Therefore, we have the abandonment rate at l_n of orders before reaching the head of the queue as $\sum_{m=n+1}^{M+N} \int_{l_{n-1}}^{l_n} \boldsymbol{\pi}_-^{(n)}(x)\mathrm{d}x \left(\mathbf{e} \otimes \left(\sum_{k=1}^{K} k\eta_{k,n-M} D_k \mathbf{e} \right) \otimes \mathbf{e} \right)$. Since the buyer arrival process evolves only when $X(t)$ is decreasing, we censor out the real periods of time when $X(t) \neq 0$ and get the same normalization factor $\hat{c}_{norm}$ as the age process. Last, divided by the total abandonment rate of orders before reaching the head of the queue (i.e., $p_{BL,>1}\lambda$), we get the desired results. □

The mean sojourn time for filled buyer orders $\mathbb{E}[W^o_{B,F}]$ can be obtained by:

$$\begin{aligned} \mathbb{E}[W^o_{B,F}] &= \frac{1}{\lambda p_{B,F}} \sum_{n=M+1}^{M+N} \left(\mathbf{v}_+^{(n)} \mathcal{M}^{\mathcal{K}^{(n)}}_{l_{n-1},l_n} + \mathbf{v}_-^{(n)} \widetilde{\mathcal{M}}^{\widehat{\mathcal{K}}^{(n)}}_{l_{n-1},l_n} \widehat{\Psi}^{(n)} \right) \boldsymbol{\delta}(+,+) \\ &\quad + \frac{1}{\lambda p_{B,F}} \sum_{n=M+1}^{M+N} \left(\mathbf{v}_+^{(n)} \mathcal{M}^{\mathcal{K}^{(n)}}_{l_{n-1},l_n} \Psi^{(n)} + \mathbf{v}_-^{(n)} \widetilde{\mathcal{M}}^{\widehat{\mathcal{K}}^{(n)}}_{l_{n-1},l_n} \right) \boldsymbol{\delta}(+,-,n), \end{aligned} \quad (5.20)$$

where $\mathcal{M}_{l_{n-1},l_n}^{\mathcal{K}^{(n)}}$ and $\widetilde{\mathcal{M}}_{l_{n-1},l_n}^{\widehat{\mathcal{K}}^{(n)}}$ can be found in equation (B.2) in Appendix B. The mean sojourn time of an arbitrary buyer order can be obtained by

$$\mathbb{E}[W_B^o] = p_{B,F}^o \mathbb{E}[W_{B,F}^o] + p_{BL,1}^o \mathbb{E}[W_{BL,1}^o] + p_{BL,>1}^o \mathbb{E}[W_{BL,>1}^o]. \tag{5.21}$$

5.2.3 Queue length of orders

Let $q_B^o(t)$ be the queue length of buyer orders at an arbitrary time t. The mean queue length (i.e., total batch size) of buyer orders can be obtained by Little's law as $\mathbb{E}[q_B^o(t)] = \lambda \mathbb{E}[W_B^o]$. We provide a way to find the probability generating function of the queue length distribution of the buyer orders in this subsection.

If $a(t) = x, s(t) = k, I_{(b)}(t) = m_b$ at an arbitrary time t, the queue length consists of the remaining batch size of the buyer at the head of the queue and the total batch size of the buyers arrived after that buyer (i.e., in the period $(t-x,t)$) who have not abandoned the queue yet. If we only consider buyers who are still in the queue, the arrival process of buyer orders in $(t-l_n, t-l_{n-1})$ can be expressed as a time inhomogeneous $BMAP$ with

$$\Delta(s) := (D_0(s), D_1(s), ..., D_K(s)), \tag{5.22}$$

where $D_0(s) = D_0 + \sum_{k=1}^{K}(1-\xi_{k,n})D_k$ and $D_k(s) = \xi_{k,n} D_k$ for $k = 1, ..., K$ when $s \in (t-l_n, t-l_{n-1})$ and $\xi_{k,n} = \sum_{i=n-M}^{N} \eta_{k,i}$, for $n = M, M+1, ..., M+N$.

In general, for buyers arrived in $(t-l_n, t-l_{n-1})$, the probability generating function for the batch size is given by

$$P^*(n, z, l_n - l_{n-1}) = \exp\left\{ \left(D_0(s) + \sum_{k=1}^{K} z^k D_k(s) \right) (l_n - l_{n-1}) \right\}. \tag{5.23}$$

The probability generating function of $q_B^o(t)$ can be derived based on the joint distribution of the age process. Recall that $P\{a(t) \le 0\} = \hat{\mathbf{p}}^{(M)}\mathbf{e} + \sum_{n=1}^{M} \left(\mathbf{v}_+^{(n)} \mathcal{L}_{l_{n-1},l_n}^{\mathcal{K}^{(n)}} \Psi^{(n)} + \mathbf{v}_-^{(n)} \widetilde{\mathcal{L}}_{l_{n-1},l_n}^{\widehat{\mathcal{K}}^{(n)}} \right) \mathbf{e}$. Conditioning on $a(t)$ at an arbitrary time t, we have

$$\mathbb{E}\left[z^{q_B^o(t)} \right] = P\{a(t) \le 0\} + \sum_{n=M+1}^{M+N} \int_{l_{n-1}}^{l_n} \mathbf{f}_B^{(n)}(x) \left(I(z^K) \otimes \left(P^*(n, z, x - l_{n-1}) \prod_{m=n-1}^{M+1} P^*(m, z, b_m) \right) \otimes I \right) \mathrm{d}x \mathbf{e}, \tag{5.24}$$

where $I(z^K) = diag(z^K, z^{K-1}, ..., z^1)$ and $b_m = l_m - l_{m-1}$, for $m = M+1, M+2, ..., M+N$.

By Theorem 2.4.2 in He (2014), we have

$$\left.\frac{\partial P^*(n,z,x)\mathbf{e}}{\partial z}\right|_{z=1} = \left(\sum_{k=1}^{K} k\xi_{k,n}\lambda_k\right) x\mathbf{e} + (e^{Dx}-I)(D-\mathbf{e}\boldsymbol{\theta}_b)^{-1}\left(\sum_{k=1}^{K} k\xi_{k,n}D_k\right)\mathbf{e}. \tag{5.25}$$

Consequently, we obtain the following result.

Proposition 4. *The mean queue length of remaining buyer orders is given by*

$$\begin{aligned}
\mathbb{E}\left[q_B^o(t)\right] = & \sum_{n=M+1}^{M+N}\left(\mathbf{v}_+^{(n)}\mathcal{L}_{l_{n-1},l_n}^{\mathcal{K}^{(n)}} + \mathbf{v}_-^{(n)}\widetilde{\mathcal{L}}_{l_{n-1},l_n}^{\widehat{\mathcal{K}}^{(n)}}\widehat{\Psi}^{(n)}\right)(I(K)\otimes I\otimes I)\mathbf{e} \\
& + \sum_{n=M+1}^{M+N}\left(\mathbf{v}_+^{(n)}\mathcal{L}_{l_{n-1},l_n}^{(\mathcal{K}^{(n)},D)} + \mathbf{v}_-^{(n)}\widetilde{\mathcal{L}}_{l_{n-1},l_n}^{(\widehat{\mathcal{K}}^{(n)},D)}\right) \\
& \quad \times\left(\sum_{m=M+1}^{n-1}\left(I\otimes e^{D(l_{n-1}-l_m)}\left(\left(\sum_{k=1}^{K} k\xi_{k,m}\lambda_k\right)b_m I + (e^{Db_m}-I)(D-\mathbf{e}\boldsymbol{\theta}_b)^{-1}\left(\sum_{k=1}^{K} k\xi_{k,m}D_k\right)\right)\otimes I\right)\right)\mathbf{e} \\
& + \sum_{n=M+1}^{M+N}\sum_{k=1}^{K} k\xi_{k,n}\lambda_k\left(\mathbf{v}_+^{(n)}\left(\mathcal{M}_{l_{n-1},l_n}^{\mathcal{K}^{(n)}} - l_{n-1}\mathcal{L}_{l_{n-1},l_n}^{\mathcal{K}^{(n)}}\right) + \mathbf{v}_-^{(n)}\left(\widetilde{\mathcal{M}}_{l_{n-1},l_n}^{\widehat{\mathcal{K}}^{(n)}} - l_{n-1}\widetilde{\mathcal{L}}_{l_{n-1},l_n}^{\widehat{\mathcal{K}}^{(n)}}\right)\widehat{\Psi}^{(n)}\right)\mathbf{e} \\
& + \sum_{n=M+1}^{M+N}\left(\mathbf{v}_+^{(n)}\left(\mathcal{L}_{l_{n-1},l_n}^{(\mathcal{K}^{(n)},D)} - \mathcal{L}_{l_{n-1},l_n}^{\mathcal{K}^{(n)}}\right) + \mathbf{v}_-^{(n)}\left(\widetilde{\mathcal{L}}_{l_{n-1},l_n}^{(\widehat{\mathcal{K}}^{(n)},D)} - \widetilde{\mathcal{L}}_{l_{n-1},l_n}^{\widehat{\mathcal{K}}^{(n)}}\widehat{\Psi}^{(n)}\right)\right) \\
& \quad \times\left(I\otimes\left((D-\mathbf{e}\boldsymbol{\theta}_b)^{-1}\left(\sum_{k=1}^{K} k\xi_{k,n}D_k\right)\right)\otimes I\right)\mathbf{e},
\end{aligned} \tag{5.26}$$

where $I(K) = diag(K, K-1, ..., 1)$, *matrices* $\mathcal{L}_{l_{n-1},l_n}^{(\mathcal{K}^{(n)},D)}$,$\widetilde{\mathcal{L}}_{l_{n-1},l_n}^{(\widehat{\mathcal{K}}^{(n)},D)}$,$\mathcal{L}_{l_{n-1},l_n}^{\mathcal{K}^{(n)}}$, $\widetilde{\mathcal{L}}_{l_{n-1},l_n}^{\widehat{\mathcal{K}}^{(n)}}$ $\mathcal{M}_{l_{n-1},l_n}^{\mathcal{K}^{(n)}}$ *and* $\widetilde{\mathcal{M}}_{l_{n-1},l_n}^{\widehat{\mathcal{K}}^{(n)}}$ *can be found in Appendix B.*

Proof. It is readily seen that

$$\begin{aligned}
\mathbb{E}[q_B^o(t)] = & \left.\frac{\partial\mathbb{E}[z^{q_B^o(t)}]}{\partial z}\right|_{z=1} = \sum_{n=M+1}^{M+N}\left(\mathbf{v}_+^{(n)}\mathcal{L}_{l_{n-1},l_n}^{\mathcal{K}^{(n)}} + \mathbf{v}_-^{(n)}\widetilde{\mathcal{L}}_{l_{n-1},l_n}^{\widehat{\mathcal{K}}^{(n)}}\widehat{\Psi}^{(n)}\right)(I(K)\otimes I\otimes I)\mathbf{e} \\
& + \sum_{n=M+1}^{M+N}\sum_{m=M+1}^{n-1}\int_{l_{n-1}}^{l_n}\mathbf{f}_B^{(n)}(x)\left(I\otimes e^{D(x-l_m)}\otimes I\right)\mathrm{d}x \\
& \quad \times\left(I\otimes\left(\left(\sum_{k=1}^{K} k\xi_{k,m}\lambda_k\right)b_m I + (e^{Db_m}-I)(D-\mathbf{e}\boldsymbol{\theta}_b)^{-1}\left(\sum_{k=1}^{K} k\xi_{k,m}D_k\right)\right)\otimes I\right)\mathbf{e} \\
& + \sum_{n=M+1}^{M+N}\int_{l_{n-1}}^{l_n}\mathbf{f}_B^{(n)}(x)\left(I\otimes\left(\left(\sum_{k=1}^{K} k\xi_{k,n}\lambda_k\right)(x-l_{n-1})I\right.\right. \\
& \qquad \left.\left. + (e^{D(x-l_{n-1})}-I)(D-\mathbf{e}\boldsymbol{\theta}_b)^{-1}\left(\sum_{k=1}^{K} k\xi_{k,n}D_k\right)\right)\otimes I\right)\mathbf{e}\mathrm{d}x.
\end{aligned} \tag{5.27}$$

The desired result is obtained by replacing the integrals with matrices. □

5.3 Buyer-Seller Level Queueing Quantities

In this subsection, we find queueing quantities associated with buyers and sellers, instead of orders. The idea for finding those quantities is similar to that for orders. Consequently, some details are omitted.

5.3.1 Matching rates and matching probability of buyers-sellers

Let ω_B be the number of fully filled buyers per unit time, to be called *the matching rate* of the buyers. Similarly, let ω_S be the matching rate of the sellers. Note that ω_B may not be equal to ω_S because of partial matching.

Proposition 5. *The matching rate of the buyers of the system is*

$$\begin{aligned}\omega_B &= \frac{1}{\hat{c}_{norm}}\sum_{n=1}^{M}\left(\left(\int_{l_{n-1}}^{l_n}\boldsymbol{\pi}_+^{(n)}(x)\mathrm{d}x\right)\hat{\boldsymbol{\delta}}(-,+,n)+\left(\int_{l_{n-1}}^{l_n}\boldsymbol{\pi}_-^{(n)}(x)\mathrm{d}x\right)\hat{\boldsymbol{\delta}}(-,-)\right)\\ &\quad+\frac{1}{\hat{c}_{norm}}\sum_{n=M+1}^{M+N}\left(\left(\int_{l_{n-1}}^{l_n}\boldsymbol{\pi}_+^{(n)}(x)\mathrm{d}x\right)\hat{\boldsymbol{\delta}}(+,+)+\left(\int_{l_{n-1}}^{l_n}\boldsymbol{\pi}_-^{(n)}(x)\mathrm{d}x\right)\hat{\boldsymbol{\delta}}(+,-,n)\right),\end{aligned} \tag{5.28}$$

where

$$\begin{aligned}\hat{\boldsymbol{\delta}}(-,-) &= \left(\left(E\otimes\mathbf{e}\mathbf{e}^T\right)\odot\left(\mathbf{e}\otimes(D_1,D_2,...,D_K)\otimes I\right)\right)\mathbf{e};\\ \hat{\boldsymbol{\delta}}(+,-,n) &= \left(\left(\hat{E}\otimes\mathbf{e}\mathbf{e}^T\right)\odot\left(\mathbf{e}\otimes(\mathcal{D}_{1,n},\mathcal{D}_{2,n},...,\mathcal{D}_{K,n})\otimes I\right)\right)\mathbf{e};\\ \hat{\boldsymbol{\delta}}(-,+,n) &= \left(\left(F\otimes\mathbf{e}\mathbf{e}^T\right)\odot\left(\mathbf{e}\otimes(I\otimes\hat{\mathcal{D}}_{1,n},I\otimes\hat{\mathcal{D}}_{2,n},...,I\otimes\hat{\mathcal{D}}_{K,n})\right)\right)\mathbf{e};\\ \hat{\boldsymbol{\delta}}(+,+) &= \left(\left(\hat{F}\otimes\mathbf{e}\mathbf{e}^T\right)\odot\left(\mathbf{e}\otimes(I\otimes\hat{D}_1,I\otimes\hat{D}_2,...,I\otimes\hat{D}_K)\right)\right)\mathbf{e},\end{aligned} \tag{5.29}$$

where matrices E, $\hat{E}$, F and $\hat{F}$ are logical (i.e., elements can be 0 or 1) square matrices of order K: The (i,j)-th element of E is 1 if and only if $i\geq j$; The (i,j)-th element of $\hat{E}$ is 1 if and only if $i-1\geq j$; The (i,j)-th element of F is 1 if and only if $i\geq K-j$ and $i\neq K$; and the (i,j)-th element of $\hat{F}$ is 1 if and only if $i\geq K-j+1$.

Proof. The result is obtained from Proposition 1 by replacing matrices A, $\hat{A}$, $\tilde{\hat{A}}$ and $\tilde{A}$ with logical matrices E, $\hat{E}$, F and $\hat{F}$, respectively. In Proposition 1, we count how many orders are being matched. In this proposition, we use logical matrices to indicate if a buyer is fully filled when matching happens. □

Let $p_{B,F}$ be the fully filled probability of buyers, to be called the *matching probability* of the buyers.

Corollary 3. *The matching probability of the buyers can be obtained by*

$$p_{B,F}=\frac{\omega_B}{\sum_{k=1}^{K}\lambda_k}. \tag{5.30}$$

The abandonment probability of the buyers is $p_{B,L} = 1 - p_{B,F}$.

Remark: $\sum_{k=1}^{K} \lambda_k$ is the arrival rate of buyers, which is different from the arrival rate of buyer orders λ.

Again, we decompose $p_{B,L}$ into two parts based on the location in the queue: i) abandonment probability $p_{BL,1}$ of buyers at the head of the queue; and ii) abandonment probability $p_{BL,>1}$ of buyers before reaching the head of the queue. Then we have the following result.

Proposition 6. *The abandonment probability of buyers at the head of the queue is*

$$\begin{aligned} p_{BL,1} &= \frac{1}{\sum_{k=1}^{K} \lambda_k} \sum_{n=M+1}^{M+N-1} \left(\mathbf{v}_+^{(n)} e^{\mathcal{K}^{(n)}(l_n - l_{n-1})} + \mathbf{v}_-^{(n)} \widehat{\Psi}^{(n)} \right) \left(\left(\sum_{k=1}^{K} \frac{\dot{\eta}_{k,n-M}}{\sum_{m=n}^{M+N} \dot{\eta}_{k,m-M}} \mathbf{e}(K-k+1) \right) \otimes \mathbf{e} \right) \\ &\quad + \frac{1}{\sum_{k=1}^{K} \lambda_k} \hat{\mathbf{p}}^{(M)} \left(\left(\sum_{k=1}^{K} \dot{\eta}_{k,0} D_k \mathbf{e} \right) \otimes \mathbf{e} \right), \end{aligned} \tag{5.31}$$

and the abandonment probability of buyers before reaching the head of the queue is $p_{BL,>1} = p_{B,L} - p_{BL,1}$.

Proof. Similar to Proposition 2. □

5.3.2 Sojourn times of buyers-sellers

In this subsection, we present the sojourn times of buyers.

Proposition 7. *The distribution of sojourn time* $W_{B,F}$ *of fully filled buyers is*

$$\begin{aligned} P\{W_{B,F} = 0\} &= \frac{1}{\hat{c}_{norm} p_{B,F} \sum_{k=1}^{K} \lambda_k} \sum_{n=1}^{M} \left(\int_{l_{n-1}}^{l_n} \boldsymbol{\pi}_-^{(n)}(x) \mathrm{d}x \hat{\boldsymbol{\delta}}(-,-) + \int_{l_{n-1}}^{l_n} \boldsymbol{\pi}_+^{(n)}(x) \mathrm{d}x \hat{\boldsymbol{\delta}}(-,+,n) \right); \\ \frac{\mathrm{d}P\{W_{B,F} < x\}}{\mathrm{d}x} &= \frac{1}{\hat{c}_{norm} p_{B,F} \sum_{k=1}^{K} \lambda_k} \left(\boldsymbol{\pi}_+^{(n)}(x) \hat{\boldsymbol{\delta}}(+,-,n) + \boldsymbol{\pi}_-^{(n)}(x) \hat{\boldsymbol{\delta}}(+,+) \right), \\ &\qquad \text{for } l_{n-1} \le x < l_n,\ n = M+1, ..., M+N. \end{aligned} \tag{5.32}$$

The distribution of sojourn time $W_{BL,1}$ of lost buyers at the head of the queue is given by

$$P\{W_{BL,1}=0\}=\frac{1}{p_{BL,1}\sum_{k=1}^{K}\lambda_k}\hat{\mathbf{p}}^{(M)}\left(\left(\sum_{k=1}^{K}\dot{\eta}_{k,0}D_k\mathbf{e}\right)\otimes\mathbf{e}\right), \tag{5.33}$$

and, for $n=M+1,M+2,...,M+N-1$,

$$P\{W_{BL,1}=l_n\}=\frac{\left(\mathbf{v}_+^{(n)}e^{\mathcal{K}^{(n)}(l_n-l_{n-1})}+\mathbf{v}_-^{(n)}\widehat{\Psi}^{(n)}\right)}{p_{BL,1}\sum_{k=1}^{K}\lambda_k}\left(\left(\sum_{k=1}^{K}\frac{\dot{\eta}_{k,n-M}}{\dot{\eta}_{k,n-M}+...+\dot{\eta}_{k,N}}\mathbf{e}(K-k+1)\right)\otimes\mathbf{e}\right). \tag{5.34}$$

The distribution of sojourn time $W_{BL,>1}$ of lost buyers before reaching the head of the queue, we have, for $n=M,M+1,...,M+N-1$,

$$P\{W_{BL,>1}=l_n\}=\frac{1}{p_{BL,>1}\sum_{k=1}^{K}\lambda_k}\sum_{m=n+1}^{M+N}\left(\mathbf{v}_+^{(m)}\mathcal{L}^{\mathcal{K}^{(m)}}_{l_{m-1},l_m}\Psi^{(m)}+\mathbf{v}_-^{(m)}\widetilde{\mathcal{L}}^{\widehat{\mathcal{K}}^{(m)}}_{l_{m-1},l_m}\right)\left(\mathbf{e}\otimes\left(\sum_{k=1}^{K}\eta_{k,n-M}D_k\mathbf{e}\right)\otimes\mathbf{e}\right). \tag{5.35}$$

Proof. Similar to Proposition 3. □

The mean sojourn time for fully filled buyers $\mathbb{E}[W_{B,F}]$ can be calculated by:

$$\begin{aligned}\mathbb{E}[W_{B,F}] &= \frac{1}{p_{B,F}\sum_{k=1}^{K}\lambda_k}\sum_{n=M+1}^{M+N}\left(\mathbf{v}_+^{(n)}\mathcal{M}^{\mathcal{K}^{(n)}}_{l_{n-1},l_n}+\mathbf{v}_-^{(n)}\widetilde{\mathcal{M}}^{\widehat{\mathcal{K}}^{(n)}}_{l_{n-1},l_n}\widehat{\Psi}^{(n)}\right)\hat{\boldsymbol{\delta}}(+,+)\\ &\quad+\frac{1}{p_{B,F}\sum_{k=1}^{K}\lambda_k}\sum_{n=M+1}^{M+N}\left(\mathbf{v}_+^{(n)}\mathcal{M}^{\mathcal{K}^{(n)}}_{l_{n-1},l_n}\Psi^{(n)}+\mathbf{v}_-^{(n)}\widetilde{\mathcal{M}}^{\widehat{\mathcal{K}}^{(n)}}_{l_{n-1},l_n}\right)\hat{\boldsymbol{\delta}}(+,-,n).\end{aligned} \tag{5.36}$$

The mean sojourn time of an arbitrary buyer can be found by

$$\mathbb{E}[W_B]=p_{B,F}\mathbb{E}[W_{B,F}]+p_{BL,1}\mathbb{E}[W_{BL,1}]+p_{BL,>1}\mathbb{E}[W_{BL,>1}]. \tag{5.37}$$

5.3.3 Queue lengths of buyers-sellers

Proposition 8. *The mean queue length of buyers is given by*

$$\begin{aligned}\mathbb{E}[q_B(t)] = &\sum_{n=M+1}^{M+N}\left(\mathbf{v}_+^{(n)}\mathcal{L}_{l_{n-1},l_n}^{\mathcal{K}^{(n)}} + \mathbf{v}_-^{(n)}\widetilde{\mathcal{L}}_{l_{n-1},l_n}^{\widehat{\mathcal{K}}^{(n)}}\widehat{\Psi}^{(n)}\right)\mathbf{e} + \sum_{n=M+1}^{M+N}\left(\mathbf{v}_+^{(n)}\mathcal{L}_{l_{n-1},l_n}^{(\mathcal{K}^{(n)},D)} + \mathbf{v}_-^{(n)}\widetilde{\mathcal{L}}_{l_{n-1},l_n}^{(\widehat{\mathcal{K}}^{(n)},D)}\right)\\
&\times\left(\sum_{m=M+1}^{n-1}\left(I\otimes e^{D(l_{n-1}-l_m)}\left(\left(\sum_{k=1}^{K}\xi_{k,m}\lambda_k\right)b_m I + (e^{Db_m}-I)(D-\mathbf{e}\boldsymbol{\theta}_b)^{-1}\left(\sum_{k=1}^{K}\xi_{k,m}D_k\right)\right)\otimes I\right)\right)\mathbf{e}\\
&+\sum_{n=M+1}^{M+N}\sum_{k=1}^{K}\xi_{k,n}\lambda_k\left(\mathbf{v}_+^{(n)}\left(\mathcal{M}_{l_{n-1},l_n}^{\mathcal{K}^{(n)}} - l_{n-1}\mathcal{L}_{l_{n-1},l_n}^{\mathcal{K}^{(n)}}\right) + \mathbf{v}_-^{(n)}\left(\widetilde{\mathcal{M}}_{l_{n-1},l_n}^{\widehat{\mathcal{K}}^{(n)}} - l_{n-1}\widetilde{\mathcal{L}}_{l_{n-1},l_n}^{\widehat{\mathcal{K}}^{(n)}}\right)\widehat{\Psi}^{(n)}\right)\mathbf{e}\\
&+\sum_{n=M+1}^{M+N}\left(\mathbf{v}_+^{(n)}\left(\mathcal{L}_{l_{n-1},l_n}^{(\mathcal{K}^{(n)},D)} - \mathcal{L}_{l_{n-1},l_n}^{\mathcal{K}^{(n)}}\right) + \mathbf{v}_-^{(n)}\left(\widetilde{\mathcal{L}}_{l_{n-1},l_n}^{(\widehat{\mathcal{K}}^{(n)},D)} - \widetilde{\mathcal{L}}_{l_{n-1},l_n}^{\widehat{\mathcal{K}}^{(n)}}\widehat{\Psi}^{(n)}\right)\right)\\
&\times\left(I\otimes\left((D-\mathbf{e}\boldsymbol{\theta}_b)^{-1}\left(\sum_{k=1}^{K}\xi_{k,n}D_k\right)\right)\otimes I\right)\mathbf{e}.\end{aligned} \tag{5.38}$$

5.4 Numerical Example

Example 1 (continued) We use the parameters provided in Example 1 and all the results in this section to demonstrate the efficiency and effectiveness of the approach and the algorithm.

The probability that there is a buyer queue is $p_B = 0.2684$ and the probability that there is a seller queue is $p_S = 0.7095$. As demonstrated in Table 2 and Figure 3, we can find the queueing performance for both order level and buyer-seller level. For the notation of the queueing quantities related to sellers, we replace the subscript "B" of queueing quantities for buyers with subscript "S" for sellers.

Buyer order	ω^o	$p^o_{B,F}$	$p^o_{BL,1}$	$p^o_{BL,>1}$	$E[W^o_{B,F}]$	$E[W^o_{B,L}]$	$E[W^o_B]$	$E[q^o_B(t)]$	$P\{W^o_{B,F}=0\}$
	8.1017	0.9778	0.0006	0.0216	0.2893	1.1253	0.3079	2.5510	0.7131
Buyer	ω_B	$p_{B,F}$	$p_{BL,1}$	$p_{BL,>1}$	$E[W_{B,F}]$	$E[W_{B,L}]$	$E[W_B]$	$E[q_B(t)]$	$P\{W_{B,F}=0\}$
	4.7720	0.9825	0.0004	0.0172	0.3030	1.1227	0.3174	1.5417	0.7056
Seller order	ω^o	$p^o_{S,F}$	$p^o_{SL,1}$	$p^o_{SL,>1}$	$E[W^o_{S,F}]$	$E[W^o_{S,L}]$	$E[W^o_S]$	$E[q^o_S(t)]$	$P\{W^o_{S,F}=0\}$
	8.1017	0.9002	0.0000	0.0998	0.9456	2.0069	1.0515	9.4635	0.2869
Seller	ω_S	$p_{S,F}$	$p_{SL,1}$	$p_{SL,>1}$	$E[W_{S,F}]$	$E[W_{S,L}]$	$E[W_S]$	$E[q_S(t)]$	$P\{W_{S,F}=0\}$
	5.2139	0.9201	0.0000	0.0799	0.9964	2.0137	1.0777	6.1067	0.2759

Table 2: Queueing quantities for Example 1.

By the last column of Table 2 and Figure 3, we can see that the sojourn times for buyer-seller level are stochastically larger than the corresponding sojourn times for order level (i.e., $W^o_{BF} \prec W_{BF}$ and $W^o_{SF} \prec W_{SF}$), due to the partially matching.

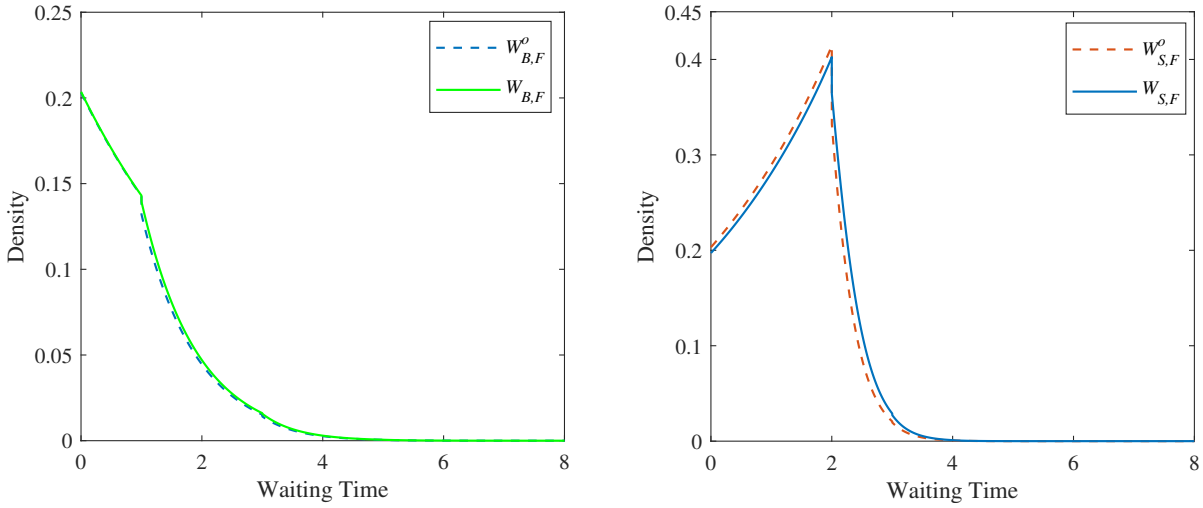

Figure 3: The stationary density functions of W^o_{BF},W_{BF},W^o_{SF} and W_{SF} for Example 1.

6 Application of the Model to Vaccine Inventory Systems

In this section, we illustrate the potential applications of the proposed queueing model by analyzing the performance of vaccination services and inventory systems. For those applications, as illustrated in Figure 4, we consider that vaccines arrive stochastically in batches (e.g., of several multi-dose vials or in the form of individual doses in the same container) to a vaccination center/clinic, and are expected to be administrated to patients randomly arriving to the clinic. We assume that the randomness in the arrival processes of patients and vaccine inventory replenishments follow $BMAP$. Patients may abandon the clinic/center before receiving vaccination service (e.g., if a patient group arrives when the vaccine inventory is depleted, they may wait for or leave before a replenishment); while vaccines may expire after their safe-use time (i.e., open or closed vaccine expiration). This hypothetical case of a vaccine clinic complies with our definition of the double-sided queue with $BMAP$.

We consider two cases in this section: the first one is a hypothetical setting, which may be close to reality for particular vaccination practices in outreach programs; the second one is based on real data from birth dose vaccines in the Gambia, West Africa (refer to Miyahara et al. (2016)).

The hypothetical case may represent, to some extent, the vaccination practices in particular clinical settings including childhood vaccination through outreach programs in developing countries and recent COVID-19 vaccination practices. Several papers have illustrated that the stochastic nature of patient arrivals to the vaccination clinic can be represented with known stochastic distributions (Yang et al. (2014); Azadi et al. (2019)). In both cases, patient arrivals may be in a batch form as multiple members of the same family or sub-

community may seek vaccination or abandon the queue simultaneously. In addition, for COVID-19, patients with a high weight-to-height ratio may need multiple doses of vaccines (Ledford (2020)). Similarly, in both cases, vaccine arrivals refer to randomly timed replenishments in batches, while abandonment refers to vaccine expiration. Random vaccine replenishments could be due to random events (e.g., extreme weather conditions, safety hazards, etc.) causing delays in accessibility to outreach regions during outreach programs, while they may be caused by uncertainties in early vaccine supply and inefficiencies in cold-chain infrastructure/network during COVID-19 vaccination. Arriving vaccines in each batch expire together based on the time from the arrival in these two examples because a) in outreach programs, all vaccines at hand are disposed of at the end of the day for safety; and b) if ultra-freezers are not used, arriving batches of Pfizer COVID-19 vaccines expire in 34-35 days (if stored in special cold-containers) or after 4-5 days (if stored in regular refrigeration units) (Pfizer (2020)). Therefore, the proposed model may represent (to some extent) the vaccination performances in these two cases.

6.1 Hypothetical Example

For the hypothetical example, we considered the case where vaccine replenishments of y doses arrive at a vaccination clinic within exponential interarrival times with mean $1/\hat{\lambda}$ unit of time. Therefore, the supply of vaccines follows a Poisson process with parameter $\hat{\lambda}$. However, each vaccine dose may lose its effectiveness, or its vial may be broken during transportation with probability p. Then, the number of effective doses follows binomial distribution $B(y, 1-p)$ for each arrival. Together, doses of vaccines arrive according to a compound Poisson process, which is in fact a $BMAP$ with

$$D_0 = -\hat{\lambda} + p^y\hat{\lambda}; \quad D_k = \begin{pmatrix} y \\ k \end{pmatrix} (1-p)^k p^{y-k}\hat{\lambda}, \quad \text{for} \quad k = 1, 2, ..., y. \tag{6.1}$$

We consider patients arriving at this clinic for vaccination according to a Poisson process with parameter $\hat{\mu}$. However, the demand for each patient depends on the patient's weight. While patients with weight in the normal range need only one dose, those with weight beyond this range need two doses. Suppose that the proportion of overweight patients in this area is r $(0 < r < 1)$. Then we have another compound Poisson process with its $BMAP$ expression given in equation (6.2) as follows,

$$\hat{D}_0 = -\hat{\mu}; \quad \hat{D}_1 = (1-r)\hat{\mu}; \quad \hat{D}_2 = r\hat{\mu}. \tag{6.2}$$

Note that the interarrival times for both patient and vaccine arrivals can be generalized to other distributions via an approximation with an appropriate Phase-Type distribution.

We suppose the vaccine will expire in v units of time after arrival and patient will abandon the system without serving after w units of time, which means the abandonment times are constants and $M = N = 2, \tilde{l}_1 = v, \hat{l}_1 = w$, $\eta_{k,1} = \dot{\eta}_{k,1} = 1$ and $\hat{\eta}_{k,1} = \dot{\hat{\eta}}_{k,1} = 1$. In this example, we assume that $\hat{\lambda} = 1, y = 10, p = 0.2, \hat{\mu} = 5, r = 0.3, v = 4, w = 1$. We also assume patients with different weights have the same abandonment times distribution and the distribution will not be changed after the first dose for overweight patients.

Demand	ω^o	$p^o_{B,F}$	$p^o_{BL,1}$	$p^o_{BL,>1}$	$E[W^o_{B,F}]$	$E[W^o_{B,L}]$	$E[W^o_B]$	$E[q^o_B(t)]$	$P\{W^o_{B,F} = 0\}$
	6.1420	**0.9449**	0.0000	0.0551	0.0398	1.0000	0.0927	0.6023	0.8601
Patient	ω_B	$p_{B,F}$	$p_{BL,1}$	$p_{BL,>1}$	$E[W_{B,F}]$	$E[W_{B,L}]$	$E[W_B]$	$E[q_B(t)]$	$P\{W_{B,F} = 0\}$
	4.7220	0.9444	0.0000	0.0556	0.0401	1.0000	0.0935	0.4674	0.8591
Dose	ω^o	$p^o_{S,F}$	$p^o_{SL,1}$	$p^o_{SL,>1}$	$E[W^o_{S,F}]$	$E[W^o_{S,L}]$	$E[W^o_S]$	$E[q^o_S(t)]$	$P\{W^o_{S,F} = 0\}$
	6.1420	**0.7678**	0.0092	0.2230	2.1719	4.0000	2.5965	20.7718	0.0689
Batch	ω_S	$p_{S,F}$	$p_{SL,1}$	$p_{SL,>1}$	$E[W_{S,F}]$	$E[W_{S,L}]$	$E[W_S]$	$E[q_S(t)]$	$P\{W_{S,F} = 0\}$
	0.6241	0.6241	0.0074	0.3685	2.2910	4.0000	2.9334	2.9334	0.0265

Table 3: Queueing quantities.

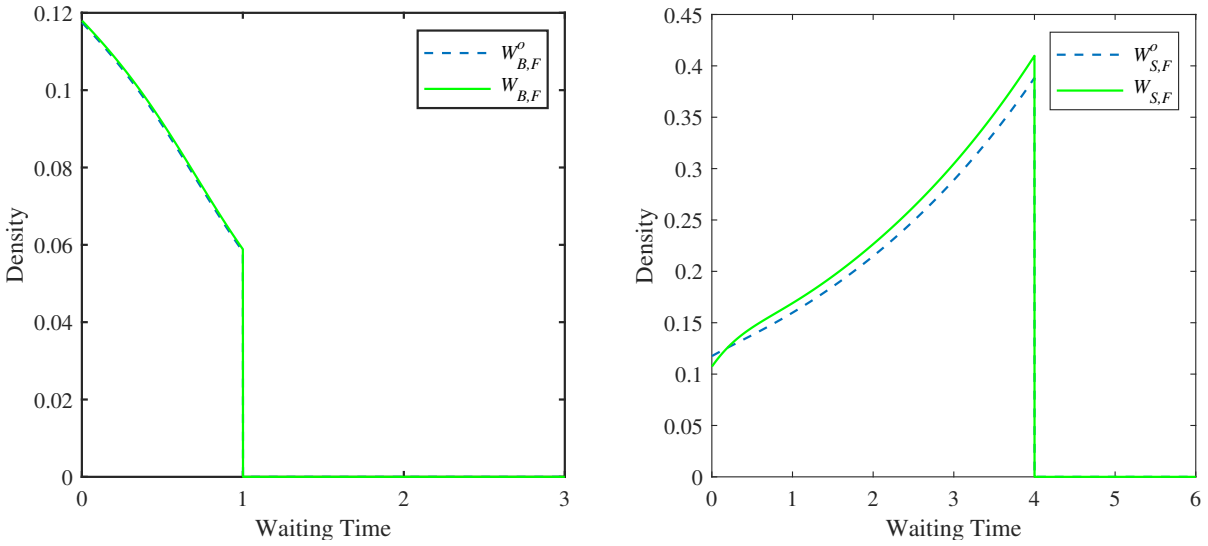

Figure 5: The stationary density functions of W^o_{BF},W_{BF},W^o_{SF} and W_{SF}.

We can find the performance of both sides of the system and the sojourn time distributions for both vaccine and patients as illustrated in Table 3 and Figure 5. From the results in Table 3, we can also see that 94.49% of patients' demands have been filled, however, only 76.78% of effective vaccine doses are being used in this example.

If the supplier wants to reduce the wastage of the vaccine but also maintain a high demand fill rate, the vaccine arrival rate should be reduced. Suppose the objective is to maximize the total matching probability for both sides (i.e., $(p^o_{B,F}+p^o_{S,F})/2$), the vaccine arrival rate should be around $\hat{\lambda} = 0.85$ (see Figure 6). When $\hat{\lambda} = 0.85$, 84.79% of vaccine doses are being used and 88.70% of patients' demands can be filled.

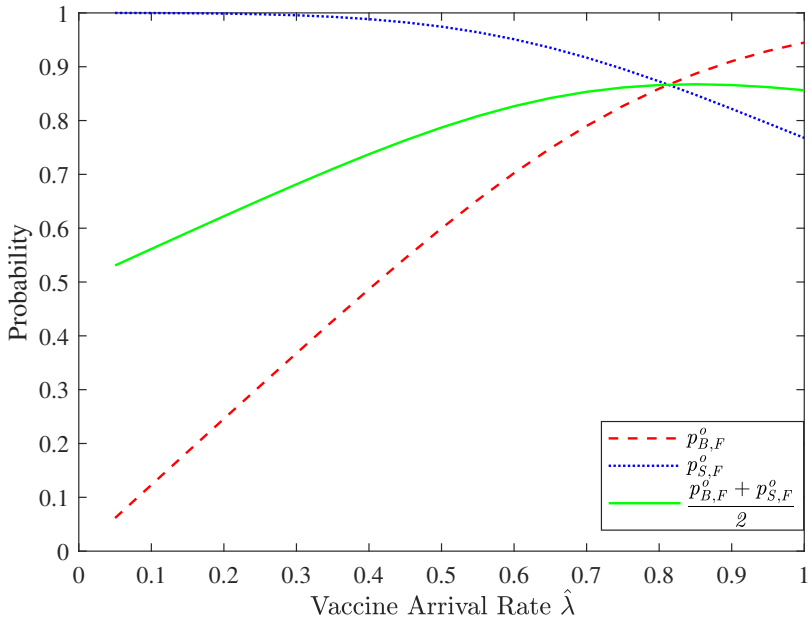

Figure 6: Matching probabilities with varying vaccine arrival rate.

6.2 Birth Dose Vaccines in West Africa

In this subsection, we use our queueing model to explore the performance of birth dose vaccines in West Africa (Miyahara et al. (2016)). We only consider Bacillus Calmette-Guerin (BCG) vaccine in this example and all parameters of our model can be obtained by the data provided in the study by Miyahara et al. (2016). An important metric in their study is the timely vaccination probability of birth dose vaccines, which is defined as the percentage of vaccination within 7 days of birth. Therefore, we use our model to estimate this metric and provide some insights about the vaccination policy.

The demand process in this example can be estimated by the birth rate within a certain region. Since we focus on the outreach program, we use the number of infants born in rural regions within the study area and assume that the demand process is a Poisson process. Based on the twining rate in that region, we get that the single birth rate is about 1.79 per day, the twin birth rate is around 0.03 per day, and multiple (greater than 2) birth rate is ignored. Infants after 7 days of birth can still be vaccinated, so we assume there is no

abandonment on this side (Note: Infants without vaccination are at high risk for contracting some dangerous diseases. So it is better to be late than no vaccination). Then we obtain

$$D_0 = -1.82; \quad D_1 = 1.79; \quad D_2 = 0.03 \tag{6.3}$$

The supply process (outreach program) is about once a week. The size of the BCG vaccine is 20 doses/vial. However, an opened vial of BCG has to be used within six hours. In practice, the delivery rate is not stable and depends on some other factors, such as the number of people, distance, and weather conditions. Since we have no information about the process, we consider three possible supply arrival processes (i.e., Poisson, Geometric renewal, and constant renewal), which are verified by data collected in Yang et al. (2014). We use a technique called Erlangization to approximate the constant interarrival times and geometric interarrival times (Asmussen et al. (2002); Ramaswami et al. (2008); He et al. (2022); Wu and He (2023)).

Using our queueing model, we obtained the timely vaccination probability and the wastage of the vaccine as shown in Table 4. Results in the table clearly indicate that the timely vaccination probability increases if the variation in the supply process of vaccine is decreasing. As such, the results suggest that a smooth supply of vaccine should be established to achieve a high timely vaccination rate. We would like to note that the wastage probability for all the cases is 0.35 due to the short expiring time of vaccine after a dose is open and the relatively small arrival rate of babies. We also want to note that timely vaccination probability should be one if the supply time is exactly 7 days. In Table 4, the timely vaccination probability for the approximate constant renewal process is less than one since there is variation in the approximation constant process.

	Supply Arrival Process					
Potential Vial Sizes	Poisson		Geometric Renewal		Approximate Constant Renewal	
	timely vaccination prob.	wastage prob.	timely vaccination prob.	wastage prob.	timely vaccination prob.	wastage prob.
20 doses/vial once per week	0.45	0.35	0.49	0.35	0.91	0.35
10 doses/vial twice per week	0.69	0.35	0.76	0.35	0.98	0.35

Table 4: Performance of birth dose vaccine (BCG) in West Africa. (Note: Approxiamte Constant Renewal is the Erlangization based approximation process to the constant renewal process.)

In this example, we show that our queueing model can approximately characterize the real vaccine program in outreach clinics in low- and middle-income countries. More data on cost-effectiveness is needed in order to

be able to make better vaccine policy recommendations, which are beyond the scope of this paper. However, we would like to comment that a vial of BCG vaccine is not open for few children to reduce vaccine wastage in many low- and middle-income countries (Thysen et al. (2021)), and timely vaccination is not only influenced by the delivery of vaccines but also by the awareness of parents of children, their lifestyle, and their level of education. In this study, we only explore the effect of delivery and try to find a better delivery pattern. For example, both 20 doses/vial and 10 doses/vial are available for BCG vaccine. In our example, we observe that 10 doses/vial option can significantly increase the timely vaccination probability, but if we want to keep the supply quantity constant, we need to increase the supply frequency to twice a week, but this will increase transportation and labor costs.

6.3 Sensitivity Analysis

We continue exploring the effects of the abandonment distributions and arrival processes on the queueing performance (e.g., matching probabilities) in this subsection.

We consider four different distributions with the same mean for each side. We sort the distributions by coefficients of variation (CV) in ascending order as, $0 = CV(\text{Constant}) < CV(\text{Erlang}) < 1 = CV(\text{Exponential}) < CV(\text{Hyperexponential})$. The hyperexponential distribution for patients is

$$\beta = [0.9, 0.1], \quad T = \begin{pmatrix} -9/2 & 0 \\ 0 & -1/8 \end{pmatrix}; \tag{6.4}$$

and the hyperexponential distribution for vaccines is

$$\beta = [0.8, 0.2], \quad T = \begin{pmatrix} -2 & 0 \\ 0 & -1/18 \end{pmatrix}. \tag{6.5}$$

We discretize all continuous distributions with $M = N = 10000$.

Matching Probabilities $p^o_{B,F}\|p^o_{S,F}$		Abandonment Distribution (Patients)			
		Constant=1	$Erlang(2,2)$	$Exp(1)$	Hyperexponential
Abandonment Distribution (Vaccines)	Constant=4	0.9449|0.7678	0.9407|0.7644	0.9380|0.7621	0.9279|0.7539
	$Erlang(2, 0.5)$	0.9004|0.7316	0.8915|0.7244	0.8856|0.7196	0.8632|0.7013
	$Exp(0.25)$	0.8687|0.7058	0.8558|0.6953	0.8472|0.6883	0.8136|0.6610
	Hyperexponential	0.7440|0.6045	0.7095|0.5765	0.6851|0.5566	0.5807|0.4718

Table 5: Comparison of different abandonment distributions.

From Table 5 we can see that the distribution of abandonment times has significant impacts on the queueing performance. If the coefficient of variation of the abandonment distribution of either side increases, the

matching probabilities are decreasing (i.e., vaccine coverage is decreasing while wastage is increasing). Intuitively, although the arrival processes and mean abandonment time are unchanged, increasing the variability of the abandonment times increases the probability that the two sides miss each other.

Next, we explore the effect of the arrival processes. To represent higher uncertainty of supply and demand, we use Markov-modulated Poisson processes ($MMPP$) to replace the Poisson processes without changing the total rates. Patients arrival can be affected by the weather, traffic, and many other factors. Therefore, we use the following MMPP,

$$\hat{D}_0 = \begin{pmatrix} -16 & 2 \\ 1 & -1.5 \end{pmatrix}; \hat{D}_1 = \begin{pmatrix} 9.8 & 0 \\ 0 & 0.35 \end{pmatrix}; \hat{D}_2 = \begin{pmatrix} 4.2 & 0 \\ 0 & 0.15 \end{pmatrix}. \tag{6.6}$$

In this process, with a probability of 1/3 the arrival rate is 14, with a probability of 2/3, the arrival rate is 0.5. It captures the non-stationary arrivals of patients to the clinic.

For the vaccine arrival process, we consider two situations. First, to capture the non-stationary arrivals of vaccines to remote areas, e.g., lower arrival rate to remote villages due to roads inaccessible after severe weather, we use the $MMPP$ in equation 6.7 as the arrival process. In this process, with a probability of 25% the arrival rate is 3, and with a probability of 75%, the arrival rate is 1/3.

$$\begin{aligned} D_0 &= \begin{pmatrix} -6 + 0.2^{10} \times 3 & 3 \\ 1 & -4/3 + 0.2^{10} \times \frac{1}{3} \end{pmatrix}; \\ D_k &= \begin{pmatrix} 3\begin{pmatrix} 10 \\ k \end{pmatrix} 0.8^k 0.2^{10-k} & 0 \\ 0 & \frac{1}{3}\begin{pmatrix} 10 \\ k \end{pmatrix} 0.8^k 0.2^{10-k} \end{pmatrix}, \quad \text{for} \quad k = 1, 2, ..., 10. \end{aligned} \tag{6.7}$$

Second, we consider an ideal situation with constant interarrival times for the vaccine arrival process. We use Erlang distribution $Erlang(m_b, m_b\hat{\lambda})$ to approximate the constant interarrival times. To obtain high accuracy, m_b needs to be relatively large (we consider $m_b = 50, 30, 10$ in this example). From Table 6, we can see that the system performance is getting better with the increase of m_b, which indicates that the situation with constant interarrival times for vaccine arrivals is the best case, but it may not be the case in reality, especially in remote areas of low- and middle-income countries.

Matching Probabilities $p^o_{B,F}\|p^o_{S,F}$		Arrival Process (Patients)	
		Poisson	*MMPP*
Arrival Process (Vaccines)	Renewal Process with $Erlang(50, 50)$	0.9999|0.8124	0.9630|0.7825
	Renewal Process with $Erlang(30, 30)$	0.9998|0.8124	0.9620|0.7817
	Renewal Process with $Erlang(10, 10)$	0.9992|0.8119	0.9569|0.7775
	Poisson	0.9449|0.7678	0.8882|0.7217
	MMPP	0.8993|0.7307	0.8448|0.6864

Table 6: Comparison of different arrival processes.

The sensitivity analysis result on the arrival process given in Table 6 also shows that the non-stationary Poisson arrival processes (i.e., $MMPP$) result in the worst system performance. This result is important because most previous studies assumed stationary Poisson arrival processes when analyzing the vaccine administration problem. Although the Poisson process is shown to be a good match to the panel data on patient arrivals in several regions of developing countries, other interarrival distributions (normally with high variance) are shown to better represent the arrival processes (Yang et al. (2014)). Therefore, The discrepancy between the performance quantities of the proposed model in terms of matching rate under more general distributions and the Poisson arrival process indicates that the existing analyses may overestimate the performance of vaccine administration in such regions. This highlights the need to develop more advanced methods, e.g., the proposed method and its extensions, to analyze the problem more accurately in such regions.

7 Conclusions

In this paper, we considered a double-sided queueing model with $BMAP$ and finite discrete abandonment times. The model is quite general as $BMAP$ can approximate any stochastic arrival process and the finite discrete distribution can approximate general distributions. Further, we assumed that the abandonment times of the customers on both sides depend on their position in the queue and their batch size, which is more general than those in the literature and makes the model be more realistic. We used the multi-layer $MMFF$ processes to analyze the queueing model and developed an efficient computational method to compute the queueing quantities such as the matching rates, fill rates, abandonment probabilities, distributions of sojourn times, and the mean queue lengths. These quantities are important not only in the queueing system but also in inventory management and financial markets.

We applied our model to a vaccine inventory system. Our model can capture the uncertainty in supply, demand and storage in the system by considering non-stationary arrival processes and abandonment distribution functions with varying coefficients of variation. We observed that system performance is sensitive to the level of uncertainty and higher uncertainty in supply, demand and storage leads to more wastage of vaccines and worse system performance. This indicates that there is an incentive for decision-makers to consider the uncontrollable uncertainty (e.g., patients abandonment and non-stationary arrivals) when designing the vaccine inventory system and to maintain a stable vaccine supply and inventory system in low- and middle-income countries.

There are some limitations in our vaccine inventory system application. First, we did not consider open vial wastage for multi-dose vials vaccine in the system while open vial wastage is a major contributor to the vaccine wastage (Lee et al. (2010)). However, our model can still be used to analyze the open vial wastage of the multi-dose vial vaccine system if we consider a more complicated vaccine expiration mechanism, which is left for future research. Second, we considered a continuous-time model without breaks, but the vaccination outreach sessions are normally 2-8 hours per day (Azadi et al. (2019)), although we can use *BMAP* to approximate the out-sessions periods of time, the computational complexity would become a problem as the state space increases. Future work will focus on developing a more realistic model and building an optimal decision model for the vaccine inventory system considering the high uncertainty.

Double-sided queues have become an increasingly interesting research topic in recent years with the emerging of the sharing economy, for example, ride-sharing, bicycle-sharing, online rental and online lending (Cohen (2018); Benjaafar and Hu (2020)). Further investigation of such queueing models remains of interest for future research. For example, a double-sided queueing model with priority has been studied by Diamant and Baron (2019), it may be an interesting direction to consider batch arrivals and priority in the double-sided queueing model. Based on the work of Adan and Weiss (2012) and Adan et al. (2018), the double-sided queueing model with a type-dependent matching mechanism could also be an interesting topic for future research.

References

Adan, I., I. Kleiner, R. Righter, G. Weiss. 2018. FCFS parallel service systems and matching models. *Performance Evaluation*, 127, 253–272.

Adan, I., G. Weiss. 2012. Exact FCFS matching rates for two infinite multitype sequences. *Operations Research*, 60 (2), 475–489.

Afèche, P., A. Diamant, J. Milner. 2014. Double-sided batch queues with abandonment: Modeling crossing networks. *Operations Research*, 62 (5), 1179–1201.

Ahn, S., A. L. Badescu, V. Ramaswami. 2007. Time dependent analysis of finite buffer fluid flows and risk models with a dividend barrier. *Queueing Systems*, 55 (4), 207–222.

Akan, M., O. Alagoz, B. Ata, F. S. Erenay, A. Said. 2012. A broader view of designing the liver allocation system. *Operations Research*, 60 (4), 757–770.

Asmussen, S. 1995. Stationary distributions for fluid flow models with or without Brownian noise. *Communications in Statistics. Stochastic Models*, 11 (1), 21–49.

Asmussen, S., F. Avram, M. Usabel. 2002. Erlangian approximations for finite-horizon ruin probabilities. *ASTIN Bulletin: The Journal of the IAA*, 32 (2), 267–281.

Asmussen, S., G. Koole. 1993. Marked point processes as limits of Markovian arrival streams. *Journal of Applied Probability*, 30 (2), 365–372.

Azadi, Z., H. Gangammanavar, S. Eksioglu. 2019. Developing childhood vaccine administration and inventory replenishment policies that minimize open vial wastage. *Annals of Operations Research*, pp. 1–33.

Badescu, A. 2008. Discussion of the discounted joint distribution of the surplus prior to ruin in a Sparre Andersen model. *North American Actuarial Journal*, 12 (2), 210–212.

Badescu, A., S. Drekic, D. Landriault. 2007a. Analysis of a threshold dividend strategy for a MAP risk model. *Scandinavian Actuarial Journal*, 2007 (4), 227–247.

Badescu, A., S. Drekic, D. Landriault. 2007b. On the analysis of a multi-threshold Markovian risk model. *Scandinavian Actuarial Journal*, 2007 (4), 248–260.

Badescu, A., D. Landriault. 2007. Moments of the discounted dividends in a threshold-type Markovian risk process. *Brazilian Journal of Probability and Statistics*, 21 (1), 13–25.

Badescu, A., D. Landriault. 2008. Recursive calculation of the dividend moments in a multi-threshold risk model. *North American Actuarial Journal*, 12 (1), 74–88.

Badescu, A., D. Landriault. 2009. Applications of fluid flow matrix analytic methods in ruin theory - a review. *Serie A: Matemáticas de la Revista de la Real Academia de Ciencias Exactas, Físicas y Naturales*, 103 (2), 353–372.

Bar-Lev, S. K., O. Boxma, B. Mathijsen, D. Perry. 2017. A blood bank model with perishable blood and demand impatience. *Stochastic Systems*, 7 (2), 237–263.

Bean, N., M. O'Reilly. 2008. Performance measures of a multi-layer Markovian fluid model. *Annals of Operations Research*, 160 (1), 99–120.

Benjaafar, S., M. Hu. 2020. Operations management in the age of the sharing economy: what is old and what is new? *Manufacturing & Service Operations Management*, 22 (1), 93–101.

Boxma, O. J., I. David, D. Perry, W. Stadje. 2011. A new look at organ transplantation models and double matching queues. *Probability in the Engineering and Informational Sciences*, 25 (2), 135–155.

Cohen, M. C. 2018. Big data and service operations. *Production and Operations Management*, 27 (9), 1709–1723.

Conolly, B., P. Parthasarathy, N. Selvaraju. 2002. Double-ended queues with impatience. *Computers & Operations Research*, 29 (14), 2053–2072.

De Boeck, K., C. Decouttere, N. Vandaele. 2019. Vaccine distribution chains in low-and middle-income countries: A literature review. *Omega*, p. 102097.

Diamant, A., O. Baron. 2019. Double-sided matching queues: Priority and impatient customers. *Operations Research Letters*, 47 (3), 219–224.

Dobbie, J. M. 1961. Letter to the editor—a doubled-ended queuing problem of Kendall. *Operations Research*, 9 (5), 755–757.

Giveen, S. M. 1963. A taxicab problem with time-dependent arrival rates. *SIAM Review*, 5 (2), 119–127.

Graves, S. C. 1982. The application of queueing theory to continuous perishable inventory systems. *Management Science*, 28 (4), 400–406.

He, Q.-M. 2014. *Fundamentals of Matrix-Analytic Methods*, volume 365. Springer.

He, Q.-M., B. Liu, H. Wu. 2022. Continuous approximations of discrete phase-type distributions and their applications to reliability models. *Performance Evaluation*, 154, 102284.

He, Q.-M., H. Wu. 2020. Multi-layer *MMFF* processes and the *MAP/PH/K+GI* queue: Theory and algorithm. *Queueing Models and Service Management*, 3 (1), 37–87.

Horváth, G. 2015. Efficient analysis of the *MMAP[K]/PH[K]/1* priority queue. *European Journal of Operational Research*, 246 (1), 128–139.

Horváth, G., B. Van Houdt. 2012. A multi-layer fluid queue with boundary phase transitions and its application to the analysis of multi-type queues with general customer impatience. *Quantitative Evaluation of Systems (QEST), 2012 Ninth International Conference on Quantitative Evaluation of Systems*, pp. 23–32.

Hsiao, A., S. N. Desai, V. Mogasale, J.-L. Excler, L. Digilio. 2017. Lessons learnt from 12 oral cholera vaccine campaigns in resource-poor settings. *Bulletin of the World Health Organization*, 95 (4), 303.

Karaesmen, I. Z., A. Scheller-Wolf, B. Deniz. 2011. Managing perishable and aging inventories: review and future research directions. *Planning production and inventories in the extended enterprise*, pp. 393–436.

Karp, C. L., D. Lans, J. Esparza, E. B. Edson, K. E. Owen, C. B. Wilson, P. M. Heaton, O. S. Levine, R. Rao. 2015. Evaluating the value proposition for improving vaccine thermostability to increase vaccine impact in low and middle-income countries. *Vaccine*, 33 (30), 3471–3479.

Kashyap, B. R. 1966. The double-ended queue with bulk service and limited waiting space. *Operations Research*, 14 (5), 822–834.

Kaspi, H., D. Perry. 1983. Inventory systems of perishable commodities. *Advances in applied probability*, pp. 674–685.

Kendall, D. G. 1951. Some problems in the theory of queues. *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 151–185.

Latouche, G., G. Nguyen. 2018. Analysis of fluid flow models. *Queueing Models and Service Management*, 1 (2), 1–30.

Ledford, H. 2020. How obesity could create problems for a COVID vaccine. *Nature*, 586 (7830), 488–489.

Lee, B. Y., B. A. Norman, T.-M. Assi, S.-I. Chen, R. R. Bailey, J. Rajgopal, S. T. Brown, A. E. Wiringa, D. S. Burke. 2010. Single versus multi-dose vaccine vials: an economic computational model. *Vaccine*, 28 (32), 5292–5300.

Liu, X. 2019. Diffusion approximations for double-ended queues with reneging in heavy traffic. *Queueing Systems*, 91 (1-2), 49–87.

Liu, X., Q. Gong, V. G. Kulkarni. 2015. Diffusion models for double-ended queues with renewal arrival processes. *Stochastic Systems*, 5 (1), 1–61.

Meini, B. 2013. On the numerical solution of a structured nonsymmetric algebraic Riccati equation. *Performance Evaluation*, 70 (9), 682–690.

Miyahara, R., M. Jasseh, P. Gomez, Y. Shimakawa, B. Greenwood, K. Keita, S. Ceesay, U. D'alessandro, A. Roca. 2016. Barriers to timely administration of birth dose vaccines in the gambia, west africa. *Vaccine*, 34 (29), 3335–3341.

Mofrad, M., L. M. Maillart, B. A. Norman, J. Rajgopal. 2018. Multi-dose vial administration with non-stationary demand and delayed service. *Operations Research for Health Care*, 19, 66–79.

Perry, D., M. Posner. 1990. Control of input and demand rates in inventory systems of perishable commodities. *Naval Research Logistics (NRL)*, 37 (1), 85–97.

Perry, D., W. Stadje. 1999. Perishable inventory systems with impatient demands. *Mathematical Methods of Operations Research*, 50 (1), 77–90.

Perry, D., W. Stadje. 2006. A controlled *M/G/1* workload process with an application to perishable inventory systems. *Mathematical Methods of Operations Research*, 64 (3), 415–428.

Pfizer 2020. COVID-19 vaccine U.S. distribution fact sheet. Available at `https://www.pfizer.com/news/hot-topics/covid_19_vaccine_u_s_distribution_fact_sheet`, accessed: 2020-12-05.

Ramaswami, V. 1999. Matrix analytic methods for stochastic fluid flows. *in: D. Smith, P. Hey (Eds.), Teletraffic Engineering in a Competitive World (Proceedings of the 16th International Teletraffic Congress)*, pp. 1019–1030.

Ramaswami, V., D. G. Woolford, D. A. Stanford. 2008. The Erlangization method for Markovian fluid flows. *Annals of Operations Research*, 160 (1), 215–225.

Ren, Q., H. Xiong, Y. Li, R. Xu, C. Zhu. 2009. Evaluation of an outside-the-cold-chain vaccine delivery strategy in remote regions of western China. *Public Health Reports*, 124 (5), 745–750.

Rogers, L. 1994. Fluid models in queueing theory and wiener-hopf factorization of Markov chains. *Annals of Applied Probability*, 4 (2), 390–413.

Takahashi, M., H. Ōsawa, T. Fujisawa. 2000. On a synchronization queue with two finite buffers. *Queueing Systems*, 36 (1-3), 107–123.

Takahashi, M., Y. Takahashi. 2000. Synchronization queue with two MAP inputs and finite buffers. In *Proceedings of the Third International Conference on Matrix Analytical Methods in Stochastic Models, Leuven, Belgium.*

Thysen, S. M., A. B. Fisker, S. Byberg, P. Aaby, P. Roy, R. White, U. Griffiths, R. C. Harris. 2021. Disregarding the restrictive vial-opening policy for bcg vaccine in guinea-bissau: impact and cost-effectiveness for tuberculosis mortality and all-cause mortality in children aged 0–4 years. *BMJ global health*, 6 (8), e006127.

UNICEF 2020. In photos: How vaccines reach the most remote places on earth. Available at `https://www.unicef.org.au/blog/unicef-in-action/october-2019/photos-vaccines-reach-most-remote-places-earth`, accessed: 2020-11-21.

Van Houdt, B. 2012. Analysis of the adaptive *MMAP[K]/PH[K]/1* queue: a multi-type queue with adaptive arrivals and general impatience. *European Journal of Operational Research*, 220 (3), 695–704.

WHO 2020. Immunization coverage. Available at `https://www.who.int/news-room/fact-sheets/detail/immunization-coverage`, accessed: 2020-11-22.

Wu, H., Q.-M. He. 2021. Double-sided queues with marked Markovian arrival processes and abandonment. *Stochastic Models*, 37 (1), 23–58.

Wu, H., Q.-M. He. 2023. Optimizing erlangization-based approximations for finite discrete distributions and discrete phase-type distributions. *Stochastic Models*, pp. 1–29.

Yang, W., M. Parisi, B. J. Lahue, M. J. Uddin, D. Bishai. 2014. The budget impact of controlling wastage with smaller vials: a data driven model of session sizes in Bangladesh, India (Uttar Pradesh), Mozambique, and Uganda. *Vaccine*, 32 (49), 6643–6648.

Zenios, S. A. 1999. Modeling the transplant waiting list: A queueing model with reneging. *Queueing Systems*, 31 (3-4), 239–251.

A Computational steps for the $MMFF$ processes

In this section, we present the computational steps for multi-layer $MMFF$ processes.

Step 1: We compute the set of basic quantities in equations (2.6), (2.7), (2.8).

Step 2: We compute two other sets of quantities for multi-layer $MMFF$ processes.

1. Within Layer n, for $n = 2, 3, ..., M + N - 1$, we have

- $\Psi_{+-}^{(l_n-l_{n-1})}$ is the transition matrix of the state of the underlying Markov chain $Q^{(n)}$ from an epoch that the fluid level $X(t)$ starts to increase from Border $n-1$ to the next first epoch that $X(t)$ returns to Border $n-1$ without reaching Border n;

- $\widehat{\Psi}_{-+}^{(l_n-l_{n-1})}$ is the transition matrix of the state of the underlying Markov chain $Q^{(n)}$ from an epoch that the fluid level $X(t)$ starts to decrease from Border n to the next first epoch that $X(t)$ returns to Border n without reaching Border $n-1$;

- $\Lambda_{++}^{(l_n-l_{n-1})}$ is the transition matrix of the state of the underlying Markov chain $Q^{(n)}$ for the process to go from Border $n-1$ to Border n before returning to Border $n-1$;

- $\widehat{\Lambda}_{--}^{(l_n-l_{n-1})}$ is the transition matrix of the state of the underlying Markov chain $Q^{(n)}$ for the process to go from Border n to Border $n-1$ before returning to Border n.

This set of quantities can be solved by the following equation if $\mu^{(n)} \neq 0$:

$$\begin{aligned}&\begin{pmatrix} \Lambda_{++}^{(l_n-l_{n-1})} & \Psi_{+-}^{(l_n-l_{n-1})} \\ \widehat{\Psi}_{-+}^{(l_n-l_{n-1})} & \widehat{\Lambda}_{--}^{(l_n-l_{n-1})} \end{pmatrix} \begin{pmatrix} I & \Psi^{(n)} e^{\mathcal{U}^{(n)}(l_n-l_{n-1})} \\ \widehat{\Psi}^{(n)} e^{\widehat{\mathcal{U}}^{(n)}(l_n-l_{n-1})} & I \end{pmatrix} \\ &= \begin{pmatrix} e^{\widehat{\mathcal{U}}^{(n)}(l_n-l_{n-1})} & \Psi^{(n)} \\ \widehat{\Psi}^{(n)} & e^{\mathcal{U}^{(n)}(l_n-l_{n-1})} \end{pmatrix}.\end{aligned} \tag{A.1}$$

2. $T_+^{(M)}$ is the transition of the underlying state from an epoch that the fluid flow level $X(t)$ starts to increase from Border M to the next first epoch that $X(t)$ returns to Border M; and $T_-^{(M)}$ is the transition of the underlying state from an epoch that the fluid flow level $X(t)$ starts to decrease from Border M to the next first epoch that $X(t)$ returns to Border M. Matrices $T_+^{(M)}$ and $T_-^{(M)}$ can be computed recursively as follows:

   - We have $T_+^{(M+N-1)} = \Psi^{(M+N)}$, and, for $n = M+1, M+2, ..., M+N-1$,

   $$T_+^{(n-1)} = \Psi_{+-}^{(l_n-l_{n-1})} + \Lambda_{++}^{(l_n-l_{n-1})}(P_{+b-}^{(n)} + P_{+b+}^{(n)} T_+^{(n)}) \left(I - \widehat{\Psi}_{-+}^{(l_n-l_{n-1})}(P_{+b-}^{(n)} + P_{+b+}^{(n)} T_+^{(n)})\right)^{-1} \widehat{\Lambda}_{--}^{(l_n-l_{n-1})}. \tag{A.2}$$

   - We have $T_-^{(1)} = \widehat{\Psi}^{(1)}$, and, for $n = 1, 2, ..., M-1$,

   $$T_-^{(n+1)} = \widehat{\Psi}_{-+}^{(l_{n+1}-l_n)} + \widehat{\Lambda}_{--}^{(l_{n+1}-l_n)}(P_{-b+}^{(n)} + P_{-b-}^{(n)} T_-^{(n)}) \left(I - \Psi_{+-}^{(l_{n+1}-l_n)}(P_{-b+}^{(n)} + P_{-b-}^{(n)} T_-^{(n)})\right)^{-1} \Lambda_{++}^{(l_{n+1}-l_n)}. \tag{A.3}$$

Step 3: We find the border probabilities $\mathbf{p}^{(M)}$ by constructing a censored continuous time Markov process $Q_{\mathbf{p}}^{(M)}$ such that $\mathbf{p}^{(M)}Q_{\mathbf{p}}^{(M)} = 0$ and $\mathbf{p}^{(M)}\mathbf{e} = 1$, and we have

$$Q_{\mathbf{p}}^{(M)} = Q_{bb}^{(M)} + \left(Q_{b+}^{(M)}, Q_{b-}^{(M)}\right) \begin{pmatrix} T_+^{(M)} & 0 \\ 0 & T_-^{(M)} \end{pmatrix} \left(I - \begin{pmatrix} 0 & P_{-b-}^{(M)} T_-^{(M)} \\ P_{+b+}^{(M)} T_+^{(M)} & 0 \end{pmatrix} \right)^{-1} \begin{pmatrix} P_{-bb}^{(M)} \\ P_{+bb}^{(M)} \end{pmatrix}. \tag{A.4}$$

.

Step 4: Compute the stationary joint density function $\pi_j^{(n)}(x)$, for $-\infty < x < \infty$, $j \in \mathcal{S}^{(n)}$ and $n = 1, 2, ..., M+N$, by Theorem 1 and the following set of linear equations,

$$\begin{aligned}
&(\boldsymbol{\pi}_+^{(M+1)}(l_M), \boldsymbol{\pi}_-^{(M)}(l_M)) \left(I - \begin{pmatrix} T_+^{(M)}(P_{-b+}^{(M)}, P_{-b-}^{(M)}) \\ T_-^{(M)}(P_{+b+}^{(M)}, P_{+b-}^{(M)}) \end{pmatrix} \right) = \mathbf{p}^{(M)}(Q_{b+}^{(M)}, Q_{b-}^{(M)}); \\
&(\boldsymbol{\pi}_+^{(2)}(l_1), \boldsymbol{\pi}_-^{(1)}(l_1)) \left(I - \begin{pmatrix} \Psi_{+-}^{(l_2-l_1)}(P_{-b+}^{(1)}, P_{-b-}^{(1)}) \\ \widehat{\Psi}^{(1)}(P_{+b+}^{(1)}, P_{+b-}^{(1)}) \end{pmatrix} \right) = \boldsymbol{\pi}_-^{(2)}(l_2) \widehat{\Lambda}_{--}^{(l_2-l_1)}(P_{-b+}^{(1)}, P_{-b-}^{(1)}); \\
&(\boldsymbol{\pi}_+^{(n+1)}(l_n), \boldsymbol{\pi}_-^{(n)}(l_n)) \left(I - \begin{pmatrix} \Psi_{+-}^{(l_{n+1}-l_n)}(P_{-b+}^{(n)}, P_{-b-}^{(n)}) \\ \widehat{\Psi}_{-+}^{(l_n-l_{n-1})}(P_{+b+}^{(n)}, P_{+b-}^{(n)}) \end{pmatrix} \right) \\
&\qquad = (\boldsymbol{\pi}_+^{(n)}(l_{n-1}), \boldsymbol{\pi}_-^{(n+1)}(l_{n+1})) \begin{pmatrix} \Lambda_{++}^{(l_n-l_{n-1})}(P_{+b+}^{(n)}, P_{+b-}^{(n)}) \\ \widehat{\Lambda}_{--}^{(l_{n+1}-l_n)}(P_{-b+}^{(n)}, P_{-b-}^{(n)}) \end{pmatrix}, \\
&\qquad\qquad \text{for } \ n = 2, ..., M-1, M+1, ..., M+N-2; \\
&(\boldsymbol{\pi}_+^{(M+N)}(l_{M+N-1}), \boldsymbol{\pi}_-^{(M+N-1)}(l_{M+N-1})) \left(I - \begin{pmatrix} \Psi^{(M+N)}(P_{-b+}^{(M+N-1)}, P_{-b-}^{(M+N-1)}) \\ \widehat{\Psi}_{-+}^{(l_{M+N-1}-l_{M+N-2})}(P_{+b+}^{(M+N-1)}, P_{+b-}^{(M+N-1)}) \end{pmatrix} \right) \\
&\qquad = \boldsymbol{\pi}_+^{(M+N-1)}(l_{M+N-2}) \Lambda_{++}^{(l_{M+N-1}-l_{M+N-2})}(P_{+b+}^{(M+N-1)}, P_{+b-}^{(M+N-1)}).
\end{aligned} \tag{A.5}$$

Step 5: By the law of total probability (i.e., $P\{-\infty < X(t) < \infty\} = 1$), we normalize the coefficients in the joint density function $(\mathbf{u}_+^{(n)}, \mathbf{u}_-^{(n)})$ and the border probabilities $\mathbf{p}^{(M)}$. The normalization factor is given by

$$c_{norm} = \mathbf{p}^{(M)}\mathbf{e} + \sum_{n=1}^{M+N} \mathbf{u}_+^{(n)} \mathcal{L}_{l_{n-1},l_n}^{\mathcal{K}^{(n)}} \left(I,\ \Psi^{(n)} \right) \mathbf{e} + \sum_{n=1}^{M+N} \mathbf{u}_-^{(n)} \widetilde{\mathcal{L}}_{l_{n-1},l_n}^{\widehat{\mathcal{K}}^{(n)}} \left(\widehat{\Psi}^{(n)},\ I \right) \mathbf{e}, \tag{A.6}$$

where $\mathbf{u}_+^{(1)} = 0$ and $\mathbf{u}_-^{(M+N)} = 0$, and

$$\begin{array}{rcl} \mathcal{L}_{l_{n-1},l_n}^{\mathcal{K}^{(n)}} & = & \int_{l_{n-1}}^{l_n} \exp(\mathcal{K}^{(n)}(x - l_{n-1}))\mathrm{d}x; \\ \widetilde{\mathcal{L}}_{l_{n-1},l_n}^{\widehat{\mathcal{K}}^{(n)}} & = & \int_{l_{n-1}}^{l_n} \exp(\widehat{\mathcal{K}}^{(n)}(l_n - x))\mathrm{d}x. \end{array} \tag{A.7}$$

Closed form solutions of $\mathcal{L}_{l_{n-1},l_n}^{\mathcal{K}^{(n)}}$ and $\widetilde{\mathcal{L}}_{l_{n-1},l_n}^{\widehat{\mathcal{K}}^{(n)}}$ can be found in Appendix B. Consequently, $\{\mathbf{u}_+^{(n)}, \mathbf{u}_-^{(n)}, \mathbf{p}^{(M)}\}$ are normalized as: $\mathbf{u}_+^{(n)} =: \mathbf{u}_+^{(n)}/c_{norm}$, $\mathbf{u}_-^{(n)} =: \mathbf{u}_-^{(n)}/c_{norm}$, and $\mathbf{p}^{(M)} =: \mathbf{p}^{(M)}/c_{norm}$.

B Evaluation of Integrals

Closed form expressions given in this Appendix are obtained in He and Wu (2020). We present them here for convenience and completeness. Define

$$\begin{array}{rcl} \mathcal{L}_{a,b}^{\mathcal{K}} & = & \int_a^b \exp(\mathcal{K}(x-a))\mathrm{d}x; \\ \widetilde{\mathcal{L}}_{a,b}^{\mathcal{K}} & = & \int_a^b \exp(\mathcal{K}(b-x))\mathrm{d}x; \\ \mathcal{M}_{a,b}^{\mathcal{K}} & = & \int_a^b x\exp(\mathcal{K}(x-a))\mathrm{d}x; \\ \widetilde{\mathcal{M}}_{a,b}^{\mathcal{K}} & = & \int_a^b x\exp(\mathcal{K}(b-x))\mathrm{d}x; \\ \mathcal{L}_{a,b}^{(\mathcal{K},D)} & = & \int_a^b \exp(\mathcal{K}(x-a))\left(I \otimes e^{D(x-a)} \otimes I\right)\mathrm{d}x; \\ \widetilde{\mathcal{L}}_{a,b}^{(\mathcal{K},D)} & = & \int_a^b \exp(\mathcal{K}(b-x))\widehat{\Psi}^{(n)}\left(I \otimes e^{D(x-a)} \otimes I\right)\mathrm{d}x. \end{array} \tag{B.1}$$

Assume that $-\infty < a < b < \infty$. If matrix $\mathcal{K}$ is invertible, we have

$$\begin{array}{rcl} \mathcal{L}_{a,b}^{\mathcal{K}} = \widetilde{\mathcal{L}}_{a,b}^{\mathcal{K}} & = & \mathcal{K}^{-1}(e^{\mathcal{K}(b-a)} - I); \\ \mathcal{M}_{a,b}^{\mathcal{K}} & = & \mathcal{K}^{-1}\left(\mathcal{K}^{-1} - aI + (bI - \mathcal{K}^{-1})e^{\mathcal{K}(b-a)}\right); \\ \widetilde{\mathcal{M}}_{a,b}^{\mathcal{K}} & = & \mathcal{K}^{-1}\left(-\mathcal{K}^{-1} - bI + (aI + \mathcal{K}^{-1})e^{\mathcal{K}(b-a)}\right), \end{array} \tag{B.2}$$

and $\mathcal{L}_{a,b}^{(\mathcal{K},D)}$, $\widehat{\mathcal{L}}_{a,b}^{(\mathcal{K},D)}$ satisfy the following Sylvester equations, respectively,

$$\begin{array}{l} \mathcal{K}\mathcal{L}_{a,b}^{(\mathcal{K},D)} + \mathcal{L}_{a,b}^{(\mathcal{K},D)}(I \otimes D \otimes I) = e^{\mathcal{K}(b-a)}\left(I \otimes e^{D(b-a)} \otimes I\right) - I; \\ \mathcal{K}\widetilde{\mathcal{L}}_{a,b}^{(\mathcal{K},D)} - \widetilde{\mathcal{L}}_{a,b}^{(\mathcal{K},D)}(I \otimes D \otimes I) = e^{\mathcal{K}(b-a)}\widehat{\Psi}^{(n)} - \widehat{\Psi}^{(n)}\left(I \otimes e^{D(b-a)} \otimes I\right). \end{array} \tag{B.3}$$

If matrix $\mathcal{K}$ is singular. Let $\mathbf{v}_L$ and $\mathbf{v}_R$ be the left and right eigenvectors, corresponding to eigenvalue zero, of $\mathcal{K}$, i.e., $\mathbf{v}_L\mathcal{K} = 0$ and $\mathcal{K}\mathbf{v}_R = 0$, and are normalized by $\mathbf{v}_L\mathbf{e} = 1$ and $\mathbf{v}_L\mathbf{v}_R = 1$. It can be shown that

$\mathcal{K}-\mathbf{v}_R\mathbf{v}_L$ is invertible. We have

$$\begin{aligned}
\mathcal{L}^{\mathcal{K}}_{a,b}=\widetilde{\mathcal{L}}^{\mathcal{K}}_{a,b} &= (\mathcal{K}-\mathbf{v}_R\mathbf{v}_L)^{-1}\left(e^{\mathcal{K}(b-a)}-I\right)+(b-a)\mathbf{v}_R\mathbf{v}_L;\\
\mathcal{M}^{\mathcal{K}}_{a,b} &= (\mathcal{K}-\mathbf{v}_R\mathbf{v}_L)^{-1}\left(be^{\mathcal{K}(b-a)}-aI\right)+\frac{(b^2-a^2)}{2}\mathbf{v}_R\mathbf{v}_L\\
&\quad -(\mathcal{K}-\mathbf{v}_R\mathbf{v}_L)^{-2}\left(e^{\mathcal{K}(b-a)}-I\right)+(b-a)\mathbf{v}_R\mathbf{v}_L;\\
\widetilde{\mathcal{M}}^{\mathcal{K}}_{a,b} &= (\mathcal{K}-\mathbf{v}_R\mathbf{v}_L)^{-1}\left(ae^{\mathcal{K}(b-a)}-bI\right)+\frac{(b^2-a^2)}{2}\mathbf{v}_R\mathbf{v}_L\\
&\quad +(\mathcal{K}-\mathbf{v}_R\mathbf{v}_L)^{-2}\left(e^{\mathcal{K}(b-a)}-I\right)-(b-a)\mathbf{v}_R\mathbf{v}_L,
\end{aligned}\tag{B.4}$$

and $\mathcal{L}^{(\mathcal{K},D)}_{a,b}$, $\widetilde{\mathcal{L}}^{(\mathcal{K},D)}_{a,b}$ satisfy the following Sylvester equations, respectively,

$$\begin{aligned}
&(\mathcal{K}-\mathbf{v}_R\mathbf{v}_L)\mathcal{L}^{(\mathcal{K},D)}_{a,b}+\mathcal{L}^{(\mathcal{K},D)}_{a,b}(I\otimes D\otimes I)=e^{\mathcal{K}(b-a)}\left(I\otimes e^{D(b-a)}\otimes I\right)-I-\mathbf{v}_R\mathbf{v}_L L_1;\\
&(\mathcal{K}-\mathbf{v}_R\mathbf{v}_L)\widetilde{\mathcal{L}}^{(\mathcal{K},D)}_{a,b}-\widetilde{\mathcal{L}}^{(\mathcal{K},D)}_{a,b}(I\otimes D\otimes I)=e^{\mathcal{K}(b-a)}\widehat{\Psi}^{(n)}-\widehat{\Psi}^{(n)}\left(I\otimes e^{D(b-a)}\otimes I\right)-\mathbf{v}_R\mathbf{v}_L\widehat{\Psi}^{(n)}L_1,
\end{aligned}\tag{B.5}$$

where $L_1=I\otimes\left(\int_a^b e^{D(x-a)}\mathrm{d}x\right)\otimes I=I\otimes\left((e^{D(b-a)}-I-(b-a)\mathbf{e}\boldsymbol{\theta}_b)(D-\mathbf{e}\boldsymbol{\theta}_b)^{-1}\right)\otimes I$.

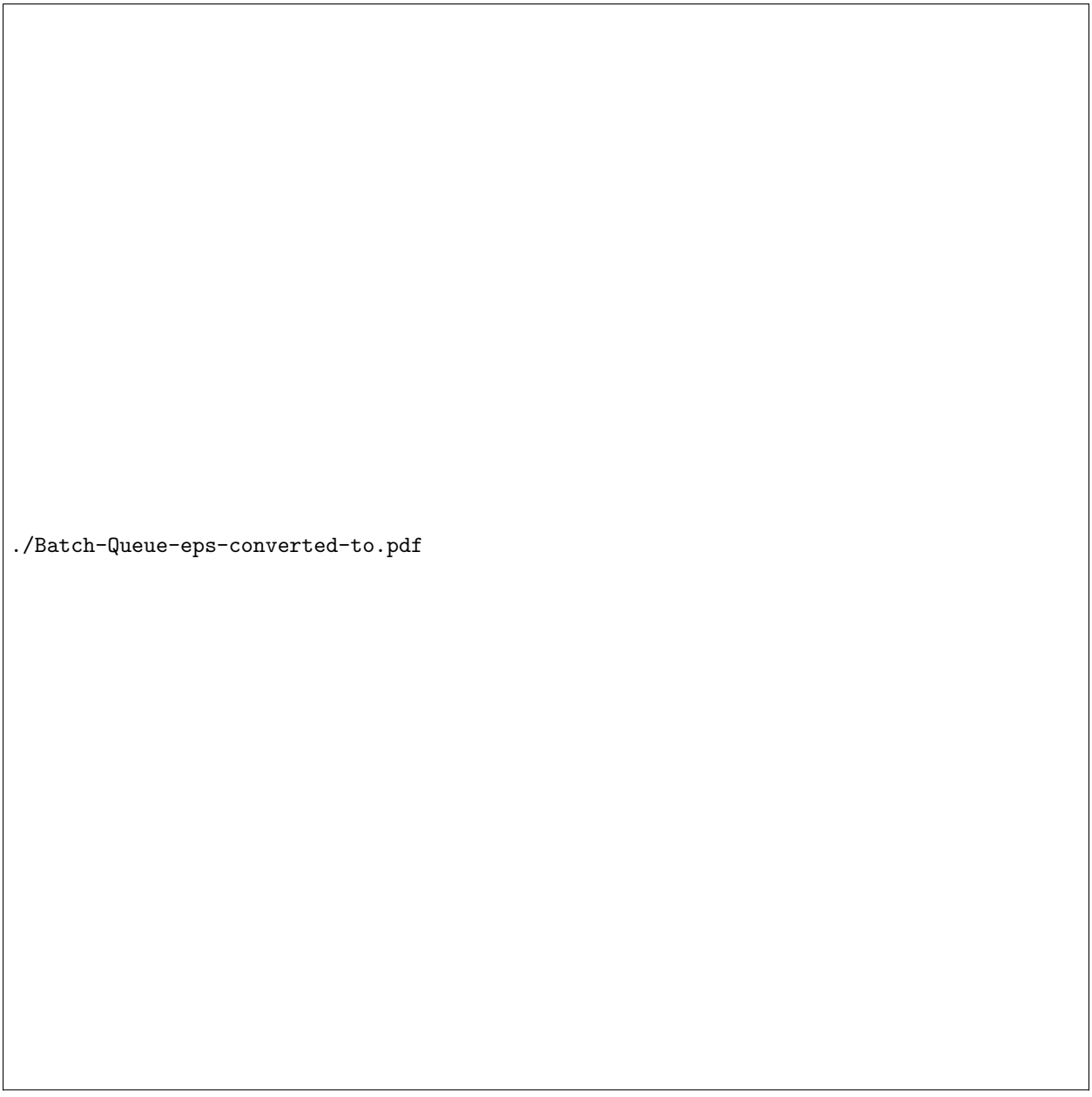

Figure 4: Diagram of the vaccine inventory system.