Analysis of Stochastic Models through Multi-Layer Markov Modulated Fluid Flow Processes

by

Haoran Wu

A thesis

presented to the University of Waterloo

in fulfillment of the

thesis requirement for the degree of

Doctor of Philosophy

in

Management Sciences

Waterloo, Ontario, Canada, 2021

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner:	Małgorzata O'Reilly Associate Professor Discipline of Mathematics, University of Tasmania
Supervisor(s):	Qi-Ming He Professor Dept. of Management Sciences, University of Waterloo
	Fatih Safa Erenay Associate Professor Dept. of Management Sciences, University of Waterloo
Internal Member:	Elizabeth Jewkes Professor Emeritus Dept. of Management Sciences, University of Waterloo
	Stanko B. Dimitrov Associate Professor Dept. of Management Sciences, University of Waterloo
Internal-External Member:	David Landriault Professor Dept. of Statistics and Actuarial Science, University of Waterloo

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

This thesis contains joint work with my supervisors, Dr. Qi-Ming He and Dr. Fatih Safa Erenay.

Chapters 3 and 4 are largely from [66], with adjustments to the computational steps, proofs and numerical examples.

Chapter 5 is largely from [116], with adjustments to notation for consistency within this thesis.

Chapter 6 has been incorporated in a paper that is currently under revision. This paper is co-authored by myself, Dr. Qi-Ming He and Dr. Fatih Safa Erenay.

My contributions to all of the work are major and I am the solo author of Chapters 1, 2 and 7.

Abstract

This thesis is concerned with the multi-layer Markov modulated fluid flow ($MMFF$) processes and their applications to queueing systems with customer abandonment.

For the multi-layer $MMFF$ processes, we review and refine the theory on the joint distribution of the multi-layer $MMFF$ processes and develop an easy to implement algorithm to calculate the joint distribution. Then, we apply the theory to three quite general queueing systems with customer abandonment to show the applicability of this approach and obtain a variety of queueing quantities, such as the customer abandonment probabilities, waiting times distributions and mean queue lengths.

The first application is the $MAP/PH/K+GI$ queue. The $MMFF$ approach and the count-server-for-phase ($CSFP$) method are combined to analyze this multi-server queueing system with a moderately large number of servers. An efficient and easy-to-implement algorithm is developed for the performance evaluation of the $MAP/PH/K+GI$ queueing model. Some of the queueing quantities such as waiting time distributions of the customers abandoning the queue at the head of the waiting queue are difficult to derive through other methods.

Then the double-sided queues with marked Markovian arrival processes ($MMAP$) and abandonment are studied. Multiple types of inputs and finite discrete abandonment times make this queueing model fairly general. Three age processes related to the inputs are defined and then converted into a multi-layer $MMFF$ process. A number of aggregate queueing quantities and quantities for individual types of inputs are obtained by the $MMFF$ approach, which can be useful for practitioners to design stochastic systems such as ride-hailing platforms and organ transplantation systems.

The last queueing model is the double-sided queues with batch Markovian arrival processes ($BMAP$) and abandonment, which arise 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. The abandonment time of a customer depends on the batch size and the position in the queue of the customer. Similar to the previous double-sided queueing model, a multi-layer $MMFF$ process related to some

age processes is constructed. A number of queueing quantities including matching rates, fill rates, sojourn times and queue length for both sides of the system are derived. This queueing model is used to analyze a vaccine inventory system as a case study in the thesis.

Overall, this thesis studies the joint stationary distribution of the multi-layer $MMFF$ processes and shows the power of this approach in dealing with complex queueing systems. Four algorithms are presented to help practitioners to design stochastic systems and researchers do numerical experiments.

Acknowledgements

I would like to express my profound gratitude to my supervisors, Professors Qi-Ming He and Fatih Safa Erenay. I received from them valuable advice, support, encouragement and more. Professor Qi-Ming He opened the gate to the theory on $MMFF$ processes and expanded my knowledge in the fields of applied probability, operations research and machine learning. My weekly meetings with Professor Fatih Safa Erenay broaden my research topics in healthcare and provided me precious opportunities to collaborate with other researchers.

Many thanks to the members of my Ph.D. committee, Professors Małgorzata O'Reilly, Elizabeth Jewkes, Stanko B. Dimitrov and David Landriault for their insightful comments and useful suggestions that improved this thesis. I also express my gratitude to Professor Stanko B. Dimitrov for allowing me to use his computing resource to do numerical experiments.

During my Ph.D. study, I worked on a healthcare project in collaboration with Mayo Clinic. I am grateful for those discussions with Dr. Brian Crum, Dr. Kalyan Pasupathy and Dr. Osman Ozaltin. They shared invaluable insights and experience in the field of healthcare.

Special thanks to my Ph.D. colleagues, Kiefer J. Burgess, Krishna Sabareesh Rajangom, Zhenggao Wu, Aliaa Alnaggar, Esma Akgün, Hsiu-Chuan Chang and Bo Lin. It was a great pleasure working with them.

Last but not least, I would like to thank my family for all the support, encouragement and love. I am greatly indebted to my girlfriend, Linjun He, who has always brought me joy, care and love.

Dedication

In memory of my father, who had always been a source of inspiration.

To my mother and my grandparents, for all their love, patience and support.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background and Motivation

The status of many important systems in our life fluctuates up and down in a continuous state space over time given the underlying conditions, e.g., the stock price, the water level in a reservoir and the data in a buffer for telecommunication systems. As an illustration, take a simple example of two seasons with constant water volume changing to a water reservoir. Specifically, the water level in a reservoir is constantly increasing during the wet season while the water level is constantly decreasing during the dry season. It can be seen from this example that such a system has an uncountable state space (e.g., water level), and the change of state is controlled by the underlying conditions (e.g., seasons). If the transition between the wet season and the dry season is dynamic and stochastic, evaluating the performance of the system becomes an interesting problem and cannot be easily solved by standard Markov chain methods. On the other hand, there are always more than two underlying conditions in real-life systems and the status of the systems may also provide feedback to the underlying conditions, thus the performance evaluation of such systems is a challenging task. Classical methods are not powerful enough to analyze such complicated stochastic systems. Therefore, we introduce multi-layer Markov modulated fluid flow ($MMFF$) processes and demonstrate its three applications to queueing models in this thesis.

Markov modulated fluid flow ($MMFF$) processes are two dimensional Markov processes and have been useful modelling tools for representing many real-life systems (e.g, dams, telecommunication and risk models) and analysing some other complex stochastic models. In the area of telecommunication systems, $MMFF$ processes have been successfully applied to analyze the system performance [53, 85, 120]. In the area of risk analysis, $MMFF$ processes have been used to find some ruin-related quantities, e.g., the time until ruin, the surplus before ruin and the deficit at ruin [20, 26]. More recently, $MMFF$ processes have been used in analyzing the performance of hydro-power generation systems [30], maintenance in continuously deteriorating systems [104], energy harvesting IoT systems [109] and SIR epidemic models [105]. An extensive literature review of $MMFF$ processes will be provided in the next chapter. In general, $MMFF$ processes still have tremendous application potential in many areas. In this thesis, we focus on the area of queueing theory, where $MMFF$ processes and their generalizations can be used to analyze complicated queueing models, especially the queueing models with customer abandonment.

Queues with customer abandonment are potentially very important, as many real-life situations in service systems and industries can be modelled as such queueing models. For instance, individuals may feel impatient when waiting for service, and perishable products may expire after a period of time (e.g., Pfizer-BioNTech COVID-19 vaccine can be stored for five days at refrigerated 2-8^{o}C conditions [94]). Other examples can be found in the emergency department in hospitals. For instance, a waiting patient may decide to leave without being seen or transfer to another health care facility because his/her health condition changes after a period of time [35, 57]. Therefore, the phenomenon of customer abandonment has been incorporated in the study of queueing systems to improve the accuracy of queueing analysis and to make queueing models more practical. In this thesis, our first application is a queueing model with multiple servers and general abandonment time.

Double-sided queues are a special type of queueing models in which each demands service from the other. Double-sided queues with customer abandonment have gained a lot of attention with the emerging of ride-hailing online platforms and sharing economy in recent years. In the ride-hailing system, both passengers and drivers can abandon the system without being paired after waiting for a long time. The time for online pairing

is very short and negligible if both sides are available, so the waiting is often due to the imbalance between the demand and supply of the two sides. For instance, the demand for ride-hailing during peak hours is often difficult to meet. Therefore, the performance evaluation of such systems can help platforms make proper decisions and further provide better service and achieve higher profits. In this thesis, we use $MMFF$ processes to analyze a double-sided queueing model with multiple types of customers and abandonment.

In trading systems, 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 the customer with the unmatched or remaining orders can abandon the system with a general (discrete) abandonment time. Those matching and abandonment features can also be seen in inventory systems. Inspired by the studies of perishable inventory systems and crossing networks trading systems, we introduce and analyze a double-sided queueing model with batch arrivals and customer abandonment.

1.2 Methodology and Basic Ideas

$MMFF$ processes are two-dimensional stochastic processes. The first dimension, i.e., the fluid level, is a piece-wise linear stochastic process in which the fluid changing rate is modulated by the second dimension, which is an underlying continuous time Markov chain. The changing rate can be positive, negative or zero. Figure 1.1 (a) plots a sample path of the fluid level of a classical $MMFF$ process. However, classical (single-layer) $MMFF$ processes are not appropriate tools to do stationary analysis, since the process can never reach a steady state. In order to evaluate the stationary performance of the queueing models with customer abandonment, multi-layer $MMFF$ processes have to be used.

Multi-layer Markov modulated fluid flow (multi-layer $MMFF$) processes, as a generalization of $MMFF$ processes, make the fluid changing rate and the underlying Markov chain depend on the fluid layers. As such, the fluid changing rate and the underlying Markov chain can be different for different layers of the fluid level, which are divided by border lines. Under certain conditions, multi-layer $MMFF$ processes have stationary dis-

tribution. Figure 1.1 (b) plots a sample path of the fluid level of a typical multi-layer $MMFF$ process with two (dashed) border lines at $l_1 = 0, l_2 = 3$ respectively and three layers $((-\infty, 0), (0, 3)$ and $(3, +\infty))$.

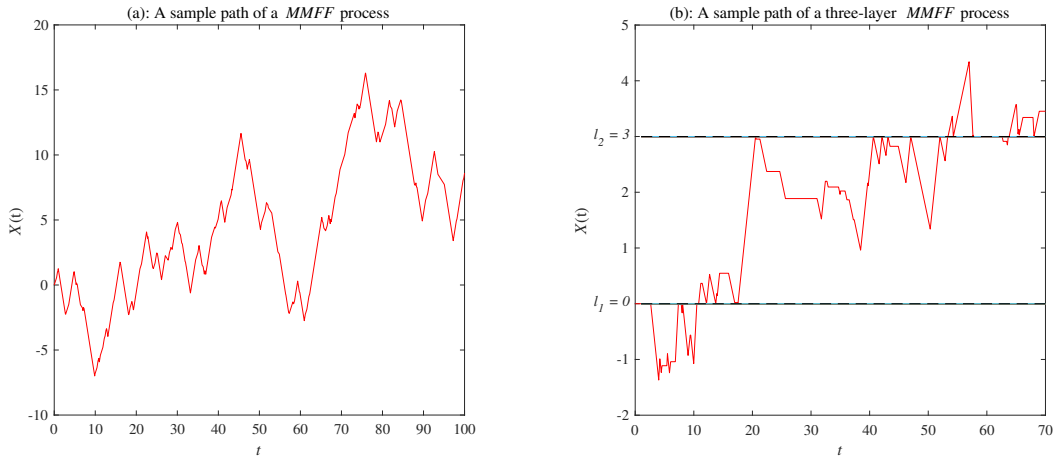

Figure 1.1: Sample paths of $MMFF$ processes

Multi-layer $MMFF$ processes were investigated extensively in the past decade, and have been applied in areas such as queueing theory (e.g., [68, 70, 110]) and risk analysis (e.g., [21, 22]). A comprehensive analysis considering all possible transitions on borders and effective and efficient algorithms for the joint stationary distribution are still needed to the best of our knowledge. Multi-layer $MMFF$ processes are complicated stochastic processes with complicated solutions for a number of basic quantities. They may not be a convenient tool to analyze simple stochastic systems, such as $M/M/1$ queueing system, but they are powerful for the investigation of complicated stochastic systems, such as the $MAP/PH/K + GI$ queueing system in Chapter 4 and double-sided queues with abandonment in Chapter 5 and Chapter 6.

For general queueing models with abandonment, the basic idea of our approach consists of three steps.

- First, we introduce a multi-layer $MMFF$ process associated with the age process (i.e., the time spent in the system of the customer at the head of the queue) of the

queueing model. Basically, we can turn an age process into a corresponding $MMFF$ process by introducing some fictitious periods. These border lines in multi-layer $MMFF$ processes can be modelled as the abandonment time points.

- Second, we use algorithms to find the joint stationary distribution of the corresponding $MMFF$ process, and censor out those fictitious periods to get the joint stationary distribution of the age process.

- Last, we use the joint stationary distribution of the age process to find a number of queueing quantities.

Regarding the basic idea of double-sided queues with abandonment, we track the ages of both sides by a single *age process* because the two sides of the queueing model can never co-exist in the system at the same time. We introduce a multi-layer $MMFF$ process associated with the age process and with a border line being 0. Thus, the age above 0 is for one side and the age below 0 (i.e., flipped age) is for the counterpart. Since we can get the joint stationary distribution of the age process, queueing quantities for individual types ($MMAP$ model) or order level ($BMAP$ model) can be obtained by considering the underlying states.

1.3 Scope of the Thesis

Based on the motivation and methodology introduced above, we will first review and refine the basic theory on multi-layer $MMFF$ processes. Thus, our first goal is to find the joint stationary distribution of multi-layer $MMFF$ processes, and develop an easy to implement algorithm to calculate the joint stationary distribution and some basic quantities. Then we apply this approach to analyze three queueing models with customer abandonment as shown in Figure 1.2. More specifically,

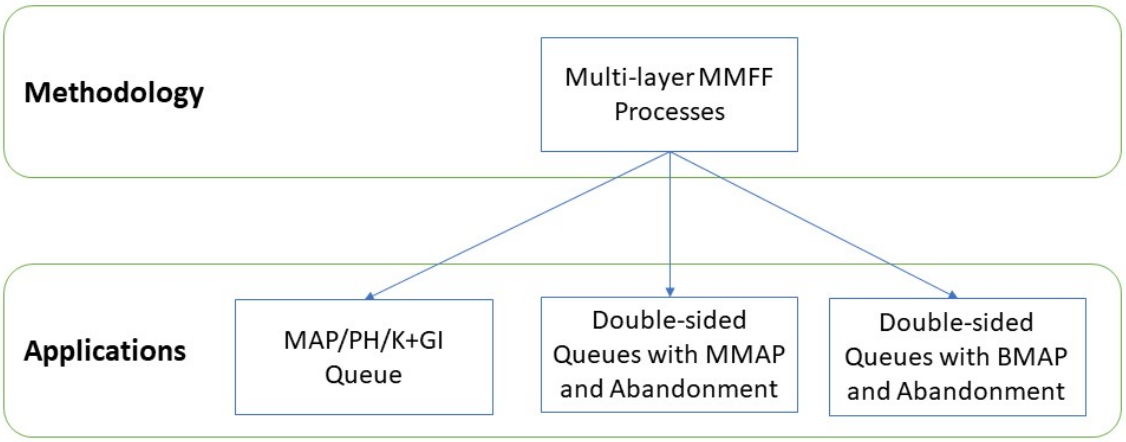

Figure 1.2: Scope of the thesis

1. $MAP/PH/K+GI$ *queueing model*: This queueing model is a very general queueing model as the Markovian arrival processes (MAP) can approximate any stochastic arrival processes and Phase-type (PH) distribution can approximate any non-negative random variables. As we mentioned in the basic ideas in the previous section, we use the border lines to model the abandonment time, which means the model can handle finite discrete abandonment times, then we can use the discrete distribution to approximate general distributions with a large number of border lines in the model. In order to analyze the queueing model with a moderately large number of servers, we combine the multi-layer $MMFF$ approach and the count-server-for-phase ($CSFP$) method in [64] to reduce the number of states and make the algorithm more efficient. In addition, we use this approach to find a bunch of interesting queueing quantities, which are difficult to find by other methods. For example, the abandonment probability of the customer at the head of the queue and the waiting time distribution of the customer abandoned at the head of the queue.

2. *Double-sided queues with* $MMAP$ *and abandonment*: In this double-sided queueing system, we use marked Markovian arrival processes ($MMAP$) to model multiple types of input. Different types of input have different abandonment time distributions, which makes the model fairly general compared with existing literature. A number of interesting quantities, such as the matching rates/probabilities, waiting

times and queue lengths for both sides can be obtained. These quantities for individual types of inputs can also be obtained, which can be useful for the analysis and design of, for instance, the ride-hailing platform.

3. *Double-sided queues with $BMAP$ and abandonment*: In this queueing system, we use batch Markovian arrival processes ($BMAP$) to model batch arrivals. We consider different abandonment time distributions for different batch sizes. In addition, the abandonment time distribution can be changed for a particular batch as partial matching may happen. Such a system can be used to analyze the performance of financial, inventory and health care systems. We apply this queueing system to a vaccine inventory model and obtain a number of queueing quantities and some insights, e.g., the fill rates of orders, and the effects of abandonment time distributions on the system performance.

1.4 Organization of the Thesis

The rest of the thesis is organized as follows. In Chapter 2, we give a brief literature review on matrix-analytic methods, $MMFF$ processes, queueing models with customer abandonment and double-sided queues. In Chapter 3, we review and refine the theory on multi-layer $MMFF$ processes and develop a computational algorithm for the joint stationary distribution. In Chapter 4, we apply multi-layer $MMFF$ processes to the $MAP/PH/K + GI$ queue and develop an algorithm for computing a variety of queueing quantities. Chapter 5 and Chapter 6 apply multi-layer $MMFF$ processes to two double-sided abandonment queues with marked arrivals and batch arrivals respectively. Chapter 7 concludes this thesis. Two important lemmas and the notations tables are collected in Appendices

Chapter 2

Literature Review

In this chapter, we review the literature related to our research problems. Section 2.1 briefly introduces three basic tools of matrix-analytic methods: Markovian arrival processes (MAP) in 2.1.1, Phase-type distributions in 2.1.2, and Quasi-Birth-and-Death (QBD) processes in 2.1.3; as well as the count-server-for-phase method in 2.1.4. Section 2.2 reviews the Markov modulated fluid flow ($MMFF$) processes while Section 2.3 reviews existing literature about queueing models with customer abandonment. Finally, Section 2.4 discusses the double-sided queues in the literature.

2.1 Matrix-Analytic Methods

Matrix-analytic methods ($MAMs$), as a set of powerful tools to analyze stochastic models, were first introduced by Marcel F. Neuts in the 1970s. Since then, $MAMs$ have been widely applied to analyze a variety of stochastic models in operations research, management science, risk/insurance and telecommunication. Early important works have been summarized in [78, 90, 91]. In terms of works on matrix-analytic methods and queueing theory, we refer to [38, 62]. In this section, we review three basic components of $MAMs$ (i.e., Markovian arrival processes, Phase-type distribution and Quasi-Birth-Death processes) and the count-server-for-phase algorithm, which will be extensively used in our research

problems. $MAMs$ have nice probabilistic interpretations and numerical tractability. As will be shown, the analysis of $MMFF$ processes by $MAMs$ involves several important and probabilistically interpretable matrices, which enhance our understanding of the stochastic models.

2.1.1 Markovian Arrival Processes

$MAPs$ were first introduced in [89], and then become a set of indispensable tools in stochastic modeling. Markovian arrival process (MAP) is a generalization of Poisson process and it keeps many useful properties of the Poisson process (e.g., Markovian property) because of the underlying Markov chain. Meanwhile, $MAPs$ can approximate any stochastic counting process arbitrarily closely. Formal definitions of the continuous time $MAPs$ are given in [62]. The basic idea is as follows.

Define an underlying continuous time Markov chain $\{I(t), t \geq 0\}$ with infinitesimal generator $D = (d_{(i,j)})$ of order m. Decompose D into matrices $\{D_0, D_1\}$, where $D_0 = (d_{0,(i,j)})$ and $D_1 = (d_{1,(i,j)})$, and all the elements of the two matrices are nonnegative except the diagonal elements of D_0 (i.e., $d_{0,(i,i)}$), which are negative. Then (D_0, D_1) defines MAP $\{(N(t), I(t)), t \geq 0\}$ with $N(0) = 0$. In the MAP, there are two ways to generate arrivals.

1. For phase i, define a Poisson process with arrival rate $d_{1,(i,i)} > 0$, for $i = 1, 2, ..., m$. The Poisson process is turned on, if $I(t) = i$; and is turned off, otherwise. If $I(t) = i$ and an arrival from the imposed Poisson process occurs, $N(t)$ increases by one, for $i = 1, 2, ..., m$.

2. At the end of each stay in state i, with probability $d_{0,(i,j)}/(-d_{(i,i)})$, $I(t)$ transits from phase i to j and $N(t)$ remains the same (i.e., without an arrival); and, with probability $d_{1,(i,j)}/(-d_{(i,i)})$, $I(t)$ transits from phase i to j and $N(t)$ increases by one (i.e., with an arrival), for $i \neq j$, and $i, j = 1, ..., m$.

There are two important generalizations of the Markovian arrival processes: batch Markovian arrival processes ($BMAPs$) [83] and marked Markovian arrival processes ($MMAPs$) [65]. $BMAPs$ can be used to model the arrival of a group of customers, which will be used

in Chapter 6 in the double-sided queues with $BMAP$ and abandonment, while $MMAPs$ can be used to model the arrival of different types of customers, which will be used in Chapter 5 in the double-sided queues with $MMAP$ and abandonment.

A $BMAP$ with matrix representation $(D_0, D_1, ..., D_K)$ is a two-dimensional continuous-time Markov process $\{(N(t), I(t)) : t \geq 0\}$ with state space $\{(n, i) : n \geq 0, i \in \{1, 2..., m\}\}$, $N(0) = 0$ and infinitesimal generator

$$G = \begin{pmatrix} D_0 & D_1 & ... & D_K & & \\ & D_0 & D_1 & ... & D_K & \\ & & D_0 & D_1 & \ddots & \ddots \\ & & & \ddots & \ddots & \ddots \end{pmatrix}, \tag{2.1.1}$$

where D_0 is a square matrix of order m with nonnegative off-diagonal elements and negative diagonal elements and D_k, $k = 1, 2, ..., K$, are nonnegative square matrices of order m, and matrix $D = D_0 + D_1 + ... + D_K$ is an irreducible infinitesimal generator.

Note that $N(t)$ is the number of arrivals by time t and $I(t)$ is the state of the underlying Markov chain at time t. The maximum batch size is K, and D_k means the transition rates with the arrival of batch size k, where $k = 0, 1, ..., K$.

$MMAP$ are actually generalizations of $BMAP$ with different interpretations to the matrix representations. For $BMAP$, matrix D_k means the transition rates with arrivals of batch size k. For $MMAP$, with the same matrix representation $(D_0, D_1, ...D_K)$, the subscript $k > 0$ of D_k represents type k arrivals. Let $N_k(t)$ be the number of type k arrivals by time t, for $k = 1, 2, ..., K$, then the continuous time Markov chain $\{(N_1(t), N_2(t), ..., N_K(t), I(t)), t \geq 0\}$ becomes an $MMAP$.

For more details on $MAPs$, we refer to [62].

2.1.2 Phase-Type Distributions

PH-distributions, as a distribution class, were initially introduced in [88]. A PH-distribution is the distribution of the time until absorption of an absorbing state of a finite-state continuous time Markov chain, which is usually called the underlying Markov chain of the

PH-distribution. If we keep tracking the state of the underlying Markov chain, we can know the residual time of absorption, and the residual time has a PH-distribution as well. This is the partial memoryless or Markovian property of PH-distribution. In general, define a continuous time Markov chain $\{I(t), t \geq 0\}$ with $m+1$ states and infinitesimal generator

$$Q = \begin{pmatrix} T & \mathbf{T^0} \\ 0 & 0 \end{pmatrix}, \tag{2.1.2}$$

where T is a subgenerator of order m and $\mathbf{T^0} = -T\mathbf{e}$.

Assume the Markov chain will be absorbed into state $m+1$ with probability one. Then the absorption time of state $m+1$ of the continuous time Markov chain, denoted by $X = \min\{t : I(t) = m+1\}$, has a phase-type distribution, given that the initial distribution of the Markov chain is $(\boldsymbol{\alpha}, 1 - \boldsymbol{\alpha}\mathbf{e})$. The distribution function of X is given by

$$P\{X \leq t\} = 1 - \boldsymbol{\alpha}\exp(Tt)\mathbf{e}. \tag{2.1.3}$$

The set of PH-distributions is closed under a number of operations (e.g., "min", "max", "+"), which is called closure properties. The closure properties play a key role in the application of PH-distributions in queueing systems.

For more details on PH-distributions, we refer to [62, 78].

2.1.3 Quasi-Birth-and-Death Processes

Quasi-Birth-and-Death Processes ($QBDs$), as generalizations of Birth-and-Death Processes, are two-dimensional Markov Processes and the transitions are skipfree to the left and to the right. For discrete time QBD, define the transition probability matrix

$$P = \begin{pmatrix} A_{0,0} & A_{0,1} & & & \\ A_{1,0} & A_{1,1} & A_0 & & \\ & A_2 & A_1 & A_0 & \\ & & \ddots & \ddots & \ddots \end{pmatrix}, \tag{2.1.4}$$

where $A_{0,0}, A_{0,1}, A_{1,0}, A_{1,1}, A_0, A_1, A_2$ are nonnegative matrices, $A_{0,0}$ is of order m_0 by m_0, $A_{0,1}$ is of order m_0 by m, $A_{1,0}$ is of order m by m_0, and others are of order m by m. They have to satisfy $(A_0+A_1+A_2)\mathbf{e} = \mathbf{e}$, $A_{1,0}\mathbf{e}+(A_{1,1}+A_0)\mathbf{e} = \mathbf{e}$ and $A_{0,0}\mathbf{e}+A_{0,1}\mathbf{e} = \mathbf{e}$. Then we can define a level independent QBD process $\{(X_k, J_k), k = 0, 1, 2, ...\}$ as a two-dimensional process with state space $\{\{(0,1),(0,2),...,(0,m_0)\} \cup \{\{1,2,...\} \times \{1,2,...,m\}\}$, where X_k is the level variable and J_k is the phase variable.

To analyze QBD processes, we need to first compute two basic matrices R and G. We can find the limiting probabilities of the QBD processes by the matrix-geometric solution with matrix R. Matrix G provides first passage probabilities. Both matrices have probabilistic interpretation and are very important in matrix-analytic methods([78, 90, 91]). They can also be further applied to $GI/M/1$ (skipfree to the right process) and $M/G/1$ (skipfree to the left process) type Markov chains (refer to [71, 90, 91, 96]).

Although QBD processes are not directly used in our research in this thesis, the algorithmic probability philosophy of QBD and $MMFF$ processes is the same. In paper [98], the relationship between these two methods was discussed and a new matrix-analytic method was developed to analyze the $MMFF$ processes. Our research follows this idea. In Chapter 3, the analysis of $MMFF$ processes also starts from several basic matrix quantities (i.e., Ψ, $\mathcal{U}$ and $\mathcal{K}$).

For more details on $QBDs$ and related structured Markov Chains, we refer to [62, 78, 91].

2.1.4 Count-Server-for-Phase Method

In this subsection, we briefly review the count-server-for-phase ($CSFP$) method, which is used to reduce the state space in the $MAP/PH/K + GI$ queue in Chapter 4.

One biggest drawback of $MAMs$ is the curse of dimensionality. In the $MAP/PH/K$ queueing system, if the order of MAP is m_a and the order of PH is m_s, we can use a straightforward method, called track-phase-for-server ($TPFS$), to generate the state space of the system, which results in a huge state space of $m_a m_s^K$ for the resulting Markov chain, which can be very big if the number of servers K is large.

The $TPFS$ method tracks the phase of each server and thus can handle nonidentical multiple servers queueing models. The construction procedure is to find the Kronecker product of the transition matrices of all servers. This procedure is easy to implement and understand, although it would result in a huge number of states. However, if servers in the queueing model are independent and identical, it is not necessary to know the phase of each server. Since all servers are identical with the same Markov chain, we just need to know the number of Markov chains that are in each phase, which leads to the $CSFP$ method. For the $MAP/PH/K$ queueing system described above, the resulting state space by $CSFP$ method is $m_a(K + m_s - 1)!/(K!(m_s - 1)!)$, which is significantly smaller than $m_a m_s^K$. We use an example in [63] to illustrate the significant difference between those two methods.

Example 2.1 [63] In a $MAP/PH/K$ queueing system, if $m_a = 1$ and $m_s = 2$, the numbers of states of the resulting Markov chains by the $TPFS$ and $CSFP$ methods are given in Table 2.1 for various K values.

K	2	4	6	8	10	15	20	30
$TPFS$	4	16	64	256	1024	32768	1048576	1073741824
$CSFP$	3	5	7	9	11	16	21	31

Table 2.1: Comparison of the numbers of states for $TPFS$ and $CSFP$

The $CSFP$ method was first formally introduced to solve a continuous time Markov chain by Ramaswami in [97]. The algorithm for the discrete case was later introduced by He and Alfa [63]. This approach has been applied to queueing models with multiple servers for decades [9, 17, 39, 67, 76, 99, 111]. The construction procedure of the $CSFP$ method is complicated and challenging, we refer to [64] for more details and several important subroutines.

2.2 Markov Modulated Fluid Flow Processes

The history of Markov modulated fluid flow ($MMFF$) processes can date back to the 1950s. These models were initially used for dam control. After the 1980s, the popularity

of telecommunication systems results in more studies in $MMFF$ (e.g., [11, 12, 103]). Since the classical $MMFF$ processes can approach positive infinity, negative infinity, or both (depending on the mean drift rate), they do not have stationary distributions. In telecommunication systems, the research focuses on the stationary distribution of the buffer content, thus Markov modulated fluid queues ($MMFQ$), the truncated version of $MMFF$, were introduced and the stationary distributions were obtained by differential equations solution techniques in the early works. In [103], $MMFF$ processes were defined and some basic quantities were obtained. By using Wiener-Hopf factorization, basic matrices such as Ψ, for the state change at some regenerative epochs, and $\mathcal{U}$, for the state change as the fluid level reaches a new low level, were obtained. By using time-reversed Markov processes, the joint stationary distributions of the fluid level and the state of the underlying Markov chain were obtained for $MMFQ$s. We use $MMFF$ for $MMFQ$ with the understanding that stationary distributions exist under a certain restriction.

Paper [98] discovered a relationship between the basic quantities for $MMFF$ processes and the basic matrix G for discrete time QBD processes in $MAMs$, which led to a new method for computing basic quantity Ψ, in addition to the classical method of solving a quadratic Riccati equation. Paper [98] also found a new approach to compute the joint stationary distribution by the crossing numbers of the fluid level, which led to another basic matrix $\mathcal{K}$. Those basic matrices are the most important matrices of $MMFF$ processes. Since then, the study of $MMFF$ processes attracted the attention of many researchers and a large number of papers appeared with various applications including; i) In matrix-analytic methods: [5, 6, 7, 43, 44, 45, 46]; ii) In risk analysis: [4, 14, 18, 20, 21, 22, 23, 24, 25]; iii) In queueing theory: [68, 70, 110]; and iv) In the theory of $MMFF$ processes (e.g., two-stage $MMFF$ processes, first passage times, two dimensional $MMFF$ processes, the Yaglom limit and fluid network): [28, 29, 31, 32, 33, 87, 92].

Multi-layer $MMFF$ processes are natural extensions of the classical $MMFF$ processes, which were first defined in [45] as a truncated classical $MMFF$ process from both above and below. The paper extended existing results on first passage probabilities and the joint stationary distribution. It turns out that multi-layer $MMFF$ processes can be analyzed in a similar way, although the solution process is more involved and the presentation of results can be tedious. Since then, more studies on multi-layer $MMFF$ processes and their

applications in queueing theory have followed.

- The basic theory for the analysis on multi-layer $MMFF$ was established in [45, 46], especially those that are related to the joint stationary distribution of the processes. We review and refine the theory on the joint stationary distribution, and present the theory and related algorithm in a systematic form in Chapter 3. The multi-layer $MMFF$ processes, in their full scale, have been analyzed in [32], but their paper focused on the transient analysis only.

- Multi-layer $MMFF$ processes have also been applied to multi-threshold MAP risk models ([21, 22, 23, 24]). The basic idea is to assume that the insurer pays dividends at different rates and collects net premiums at different rates when the surplus level resides in different surplus layers. The joint discounted density of the surplus prior to ruin and the deficit at ruin is obtained in their research.

- The theory on multi-layer $MMFF$ processes has also been applied to queueing models in the past decade ([68, 70, 110]). Paper [110] investigated a single server queue with multiple types of customers and customer abandonment, and obtained quantities related to customer abandonment and waiting times. Paper [68] analyzed a single server queue with multiple types of customers with service priority.

Our work on the queueing model in Chapter 4 is close to that in [110]. We consider a queue with many servers and customer abandonment, and extend the analysis to cover more queueing quantities (e.g., different types of abandonment probabilities and waiting times, and the queue length).

2.3 Queueing Models with Customer Abandonment

Queueing models with customer abandonment play an essential role in the design of many stochastic systems such as call centres [55, 56]. Customer abandonment means that a customer, having joined the queue, decided to leave without service after a maximal waiting time (i.e., abandonment time), which may be a constant value or follow a distribution. The

impact of customers' abandonment time distribution was empirically studied in [84]. Next, we summarize the literature according to the abandonment time assumptions.

- One case is the abandonment time being a fixed constant. Early papers usually assume the arrival processes are Poisson and service time follows an exponential distribution. We refer to [36, 61] for the early work with analytic solutions. In terms of approximation techniques, we refer to [117, 118] for examples. Approximation techniques can deal with non-exponential arrival rates and service rates.

  In terms of matrix-analytic methods, we refer the readers to [41, 67, 75]. Specifically, [41] introduced a method to analyze the $MAP/M/K$ queue with constant abandonment time; [75] used the same method to analyze $the M/PH/1$ queue with constant abandonment time; and [67] investigate $M/PH/K$ queue with constant abandonment time. Unfortunately, the method is failed to be applied to the $MAP/PH/K$ queue with customer abandonment, due to the lack of commutability of some matrices.

- In other papers, the abandonment time is assumed to be distributed in accordance with a specific distribution, most of the time, exponential distribution. We refer to [10, 101, 102] for the early work. Paper [114] compared the results of constant abandonment time and an adjusted exponential abandonment time. Paper [112] used the matrix geometric method to derive the steady-state probabilities of the queueing system with exponential abandonment time.

- There are also general abandonment time distributed assumptions in the literature. Few analytic solutions can be found in the literature. Paper [115] developed an algorithm to compute approximations for the performance measures. Then, such queueing systems have been studied by approximation techniques extensively in the last decade with the increasing power of computers. (e.g., [47, 48, 49, 72, 73]).

$MMFF$ processes have been proven to be an effective tool in analyzing queueing models [68, 70, 110]. The basic idea of the approach is to introduce the workload/age process of the queueing systems and find the corresponding Markov modulated fluid flow process by transforming jumps to skip-free periods. We first find the stationary distribution of

the fluid flow process, then obtain the stationary distribution of the workload/age by censoring, and last investigate some other queueing quantities. Furthermore, queueing models with general customer abandonment time distribution have also been studied in [70, 110] by multi-layer $MMFF$ processes. The basic idea is to use the borders to represent the abandonment time points. However, their research focuses on single server queues, and only obtain some queueing quantities (e.g., waiting time distributions and abandonment probabilities).

In Chapter 4, we investigate multi-server queueing systems with general abandonment time distribution (i.e., the $MAP/PH/K+GI$ queue) by the multi-layer $MMFF$ processes. We also apply the $CSFP$ method to reduce the state space so that the algorithm can handle systems with up to one hundred (identical) servers.

2.4 Double-Sided Queues

The double-sided queue, also being called matching queue or synchronization queue, is a queueing model that entities in each queue demands service from those in the other queue. The model was first proposed in [74] as a passenger-taxi service system where passengers come to a taxi-station to take taxis. When a passenger arrives, if there is an available taxi, the passenger takes it and they both leave the taxi-station immediately; otherwise, the passenger joins a single queue of passengers and waits for a taxi. When a taxi arrives, if there is a waiting passenger, the taxi takes the passenger and they both leave the taxi-station immediately; otherwise, the taxi joins a single queue of taxis and waits for a passenger. Further, some literature studied double-sided queues with customer abandonment in which passengers and taxis will leave if their patience runs out.

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

- **Taxi-station system:** The double-sided queueing model was first proposed by [74] as a taxi-station system where passengers and taxis are matched with each other.

- **Organ transplant system:** Patients and donated organs are waiting to match with each other while the health state of the patients and the quality of the organs deteriorate, and may abandon the system without matching. Papers [8, 37, 121] analyzed this kind of organ transplant waiting systems in the United States.

- **Perishable inventory system:** Inventory model with abandonment can be found in perishable inventory systems. Paper [93] studied a model with finite waiting space and impatient demands. Paper [27] studied a blood bank model with perishable blood and impatient demand.

- **Financial market:** Double-sided queueing model has been most recently studied by [3] as a crossing networks trading system, which has batch order arrivals for both sides.

Paper [51] analyzed a double-sided queue with priority and impatience. Papers [81, 82] analyzed such double-sided queues by using diffusion models. Similar models in manufacturing systems are called kitting systems, which have been investigated extensively [50, 95, 106]. Again, their models usually assume that the arrivals of patients or customers form a Poisson process. In [1, 2], they considered matching models with multiple types of customers with a general matching rule such that whether or not a customer can match an opposite customer depends on customers' types. However, they assumed Poisson arrival process and did not consider customer abandonment.

Matrix-analytic methods ($MAMs$) have been applied to analyze double-sided queues. For example, [107] introduced a finite space double-sided queueing model with a phase-type (PH) distribution for the interarrival times for one side. Paper [108] further analyzed a finite space double-sided queueing model with MAP arrival processes, using the quasi-birth-and-death (QBD) method. On the other hand, neither [107] nor [108] considered customer abandonment in their models. Our models are close to the model in [108] but with infinite waiting space and finite discrete abandonment time distributions for both sides.

Instead of using QBD, we use $MMFF$ processes to analyze our double-sided queueing models. In queueing applications, $MMFF$ processes are usually constructed from the age

process of the customer at the head of the queue or the virtual work-load in the queueing system. Analysis of the constructed $MMFF$ processes leads to computational methods for queueing quantities. Similar to the studies in [52, 110] and Chapter 4, the analysis approach of the double-sided queueing models is also based on the age process. We use the method developed in Chapter 3 to analyze the corresponding multi-layer $MMFF$ process constructed from the age process.

Chapter 3

Multi-Layer *MMFF* Processes

In this chapter, we review and refine the basic theory on multi-layer $MMFF$ processes and develop an algorithm to compute the joint density/distribution function and basic performance quantities (e.g., mean, variance). In Section 3.1, we first introduce the classical (i.e., single-layer) $MMFF$ processes and some basic quantities; then, we define the multi-layer $MMFF$ processes in Section 3.2; the joint stationary distribution is obtained in Section 3.3; and an algorithm is summarized in Section 3.4 with several numerical examples. Section 3.5 concludes this chapter.

3.1 Classical *MMFF* Processes and Basic Quantities

The classical $MMFF$ processes (i.e., single-layer $MMFF$ processes) are two dimensional stochastic processes $\{(X(t), \phi(t)), t \geq 0\}$ in which the changing rate of the piece-wise linear fluid level ($\{X(t), t \geq 0\}$) is modulated by a finite state continuous-time Markov chain ($\{\phi(t), t \geq 0\}$). Specifically, we have

- $\{\phi(t), t \geq 0\}$ is a continuous-time irreducible Markov chain on finite state space $\mathcal{S}$ with infinitesimal generator Q.
- The fluid level $\{X(t), t \geq 0\}$ is controlled by $\phi(\cdot)$ such that the value of $X(t)$ changes linearly at rate $c_{\phi(t)}$. The changing rate c_i of the fluid level can be positive, negative,

or zero. We put the rates into vectors $\mathbf{c} = \{c_i, i \in \mathcal{S}\}$. For convenience, the state space $\mathcal{S}$ has to be partitioned into three subsets according to the sign of c_i as follows:

$$\begin{aligned} \mathcal{S}_+ &= \{i \in \mathcal{S} : \ c_i > 0\}; \\ \mathcal{S}_- &= \{i \in \mathcal{S} : \ c_i < 0\}; \\ \mathcal{S}_0 &= \{i \in \mathcal{S} : \ c_i = 0\}. \end{aligned} \tag{3.1.1}$$

We further divide $\mathbf{c}$, according the signs of its elements, and the infinitesimal generator Q of the underlying Markov chain as

$$\mathbf{c} = (\mathbf{c}_+, \ \mathbf{c}_-, \ 0); Q = \begin{matrix} \mathcal{S}_+ \\ \mathcal{S}_- \\ \mathcal{S}_0 \end{matrix} \begin{pmatrix} Q_{++} & Q_{+-} & Q_{+0} \\ Q_{-+} & Q_{--} & Q_{-0} \\ Q_{0+} & Q_{0-} & Q_{00} \end{pmatrix}. \tag{3.1.2}$$

Given the generator Q and fluid changing rates $\mathbf{c}$, the mean drift of the fluid flow process in steady state is

$$\zeta = \boldsymbol{\alpha}\mathbf{c} \tag{3.1.3}$$

where $\boldsymbol{\alpha}$ is the stationary distribution of Q. Note that we also put fluid changing rates in diagonal matrices for computational convenience as

$$C_+ = \text{diag}(\mathbf{c}_+); C_- = -\text{diag}(\mathbf{c}_-). \tag{3.1.4}$$

With the above definition, $X(t)$ is controlled by $\phi(t)$ explicitly as

$$X(t) = X(0) + \int_0^t c_{\phi(s)} \mathrm{d}s, \quad \text{or} \quad \frac{\mathrm{d}X(t)}{\mathrm{d}t} = c_{\phi(t)}, \quad \text{for} \ \ t \geq 0. \tag{3.1.5}$$

Based on the above equations, the fluid level $\{X(t), t \geq 0\}$ can approach positive infinity, negative infinity, or both (depending on the mean drift rate ζ), so the classical $MMFF$ processes do not have limiting probabilities. Intuitively, when t tends to infinity, the process will tend to $+\infty$ if $\zeta > 0$; the process will tend to $-\infty$ if $\zeta < 0$; and if $\zeta = 0$, the process is null-recurrent and $|X(t)| \to \infty$. It has been shown mathematical rigorously that the three limits hold with probability one ([13]).

Although there are no limiting probabilities for classical $MMFF$ processes, some basic quantities related to the classical $MMFF$ processes are essential for the steady state analysis of the multi-layer $MMFF$ processes in the following sections. Next, we discuss three basic quantities in $MMFF$ processes.

If $\phi(t) \in \mathcal{S}_0$, the fluid flow level $X(t)$ remains the same. This fact causes some technical inconvenience when computing the basic quantities. The issue can be resolved by censoring the time periods that $\phi(t)$ is in $\mathcal{S}_0$. The censored underlying Markov chain is defined by

$$T = \begin{pmatrix} T_{++} & T_{+-} \\ T_{-+} & T_{--} \end{pmatrix} = \begin{pmatrix} Q_{++} & Q_{+-} \\ Q_{-+} & Q_{--} \end{pmatrix} + \begin{pmatrix} Q_{+0} \\ Q_{-0} \end{pmatrix} (-Q_{00})^{-1}(Q_{0+}, Q_{0-}). \tag{3.1.6}$$

In the rest of this section, we work with both infinitesimal generators T and Q.

Matrices Ψ and $\widehat{\Psi}$ are the most fundamental quantities in the analysis of $MMFF$ processes. All other quantities and distribution functions can be built upon Ψ and $\widehat{\Psi}$. In order to define Ψ and $\widehat{\Psi}$, we first introduce some embedded regenerative processes in $\{(X(t), \phi(t)), t \geq 0\}$. Define, $\delta_0 = \inf\{t > 0 : \ X(t) > 0\}$, and, for $n > 0$,

$$\begin{array}{rcl} \theta_n & = & \inf\{t > \delta_{n-1} : \ X(t) = 0\}, \\ \delta_n & = & \inf\{t > \theta_n : \ X(t) > 0\}, \end{array} \tag{3.1.7}$$

which are called regenerative epochs (see Figure 3.1). For example, $\{(X(\theta_n), \phi(\theta_n)), n = 1, 2, ...\}$ is a regenerative process with state space $\{0\} \times \mathcal{S}_-$. The fluid level is above 0 in intervals (δ_n, θ_{n+1}) and below 0 in intervals (θ_n, δ_n). Then we have the definitions of every elements in the matrices Ψ and $\widehat{\Psi}$:

$$\begin{array}{rcll} \Psi_{i,j} & = & P\{\theta_{n+1} - \delta_n < \infty, \ \phi(\theta_{n+1}) = j \mid \phi(\delta_n) = i\}, & \text{for } i \in \mathcal{S}_+, \ j \in \mathcal{S}_-; \\ \widehat{\Psi}_{i,j} & = & P\{\delta_n - \theta_n < \infty, \ \phi(\delta_n) = j \mid \phi(\theta_n) = i\}, & \text{for } i \in \mathcal{S}_-, \ j \in \mathcal{S}_+. \end{array} \tag{3.1.8}$$

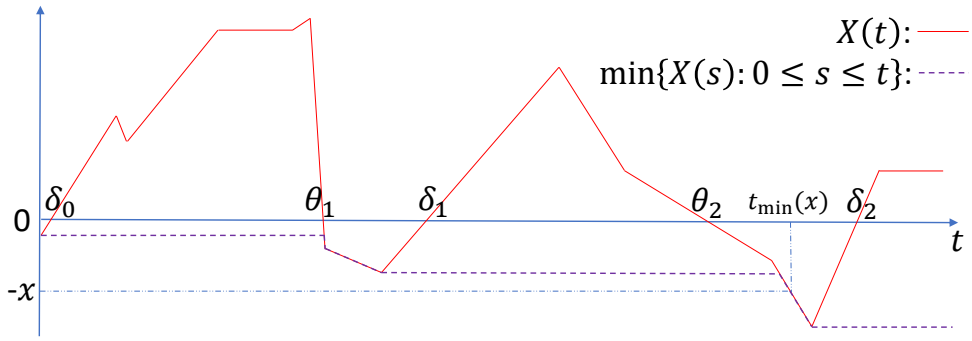

Figure 3.1: δ_n, θ_n, $t_{\min}(x)$ and $\min\{X(s) : 0 \le s \le t\}$

Observe that Ψ ($\widehat{\Psi}$) is the transition of the state of the underlying Markov chain Q from an epoch that the fluid flow level $X(t)$ starts to increase (decrease) from zero to the next first epoch that $X(t)$ returns to zero.

Lemma 3.1. *([103]) Matrices Ψ and $\widehat{\Psi}$ are the minimal nonnegative solution to the following quadratic Riccati equations, respectively:*

$$\begin{aligned} &C_+^{-1}T_{+-} + C_+^{-1}T_{++}\Psi + \Psi C_-^{-1}T_{--} + \Psi C_-^{-1}T_{-+}\Psi = 0; \\ &C_-^{-1}T_{-+} + C_-^{-1}T_{--}\widehat{\Psi} + \widehat{\Psi}C_+^{-1}T_{++} + \widehat{\Psi}C_+^{-1}T_{+-}\widehat{\Psi} = 0. \end{aligned} \tag{3.1.9}$$

The computation of Ψ and $\widehat{\Psi}$ is critical for obtaining all other quantities. Numerous papers addressed the issue. We use the Newton's method developed in [58] to obtain the minimal nonnegative solution to the quadratic Riccati equations. The algorithm is briefly discussed in Appendix A. For more details and algorithms for Ψ and $\widehat{\Psi}$, please refer to [58, 59, 77, 86, 98].

Second, we consider matrices $\mathcal{U}$ and $\widehat{\mathcal{U}}$, which are defined as

$$\begin{aligned} \mathcal{U} &= C_-^{-1}T_{--} + C_-^{-1}T_{-+}\Psi; \\ \widehat{\mathcal{U}} &= C_+^{-1}T_{++} + C_+^{-1}T_{+-}\widehat{\Psi}. \end{aligned} \tag{3.1.10}$$

Define, for $x \ge 0$,

$$\begin{aligned} t_{\min}(x) &= \min\{t : \ X(t) = -x\}; \\ i_{\min}(x) &= \phi(t_{\min}(x)). \end{aligned} \tag{3.1.11}$$

We interpret $i_{\min}(x)$ as the phase of the underlying Markov chain at the first time epoch that $X(t)$ reaches $-x$.

Lemma 3.2. *([12]) If $\zeta \leq 0$, $\{i_{\min}(x), x \geq 0\}$ is a continuous time Markov chain with infinitesimal generator $\mathcal{U}$. If $\zeta > 0$, then $\{i_{\min}(x), x \geq 0\}$ is an absorption Markov chain with state space $\mathcal{S}_- \cup \{\Delta\}$, where Δ is defined as an absorption state, and infinitesimal generator*

$$\begin{array}{c} \mathcal{S}_- \\ \Delta \end{array} \begin{pmatrix} \mathcal{U} & -\mathcal{U}\mathbf{e} \\ 0 & 0 \end{pmatrix}, \tag{3.1.12}$$

where $\mathbf{e}$ is a column vector of ones.

If we define, for $x \geq 0$, (See the dash line in Figure 3.1)

$$x_{\min}(t) = \min\{X(s) : \ 0 \leq s \leq t\}; \tag{3.1.13}$$

we can find the minimum of the fluid flow process by matrix $\mathcal{U}$. If $\zeta \leq 0$, the fluid can go to $-\infty$, thus the minimum must be $-\infty$. If $\zeta > 0$, the minimum must be finite. Assume that $X(0) = 0$ and $\phi(0)$ has a distribution $(\boldsymbol{\beta}_+, \boldsymbol{\beta}_-)$, then $-x_{\min}(\infty)$ has a phase-type distribution with representation $\left((\boldsymbol{\beta}_+, \boldsymbol{\beta}_-) \begin{pmatrix} \Psi \\ I \end{pmatrix}, \mathcal{U}\right)$.

Similarly, the same idea can be applied to $\widehat{\mathcal{U}}$, which can be interpreted as a continuous time Markov chain related to the maximum of the fluid flow process.

Third, we consider matrices $\mathcal{K}$ and $\widehat{\mathcal{K}}$, which are defined as

$$\begin{aligned} \mathcal{K} &= C_+^{-1} T_{++} + \Psi C_-^{-1} T_{-+}; \\ \widehat{\mathcal{K}} &= C_-^{-1} T_{--} + \widehat{\Psi} C_+^{-1} T_{+-}. \end{aligned} \tag{3.1.14}$$

Matrix $\mathcal{K}$ is associated with numbers of visits to certain fluid level and state during some first passage periods. Without loss of generality, we assume that $X(0) = 0$ and $\phi(0) = i$.

- For $i \in \mathcal{S}_+$ and $x > 0$, we define $(N_+(x))_{i,j}$ as the mean number of visits of the process $(X(t), \phi(t))$ to state (x, j) before $X(t)$ returns to zero. $(N_+(x))_{i,j}$ can be

further divided into $\{(N_{++}(x))_{i,j}, (N_{+-}(x))_{i,j}\}$ with $j \in \mathcal{S}_+$ (called *up-crossings*) or $j \in \mathcal{S}_-$ (called *down-crossings*), respectively.

- For $i \in \mathcal{S}_-$ and $x < 0$, we define $(N_-(x))_{i,j}$ as the mean number of visits of the process $(X(t), \phi(t))$ to state (x, j) before $X(t)$ returns to zero. $(N_-(x))_{i,j}$ can be further divided into $\{(N_{-+}(x))_{i,j}, (N_{--}(x))_{i,j}\}$ with $j \in \mathcal{S}_+$ (called *up-crossings*) or $j \in \mathcal{S}_-$ (called *down-crossings*), respectively.

Lemma 3.3. *([98]) For $x > 0$, we have i) $N_{++}(x) = \exp\{\mathcal{K}x\}$; and ii) $N_{+-}(x) = N_{++}(x)\Psi = \exp\{\mathcal{K}x\}\Psi$. For $x < 0$, we have iii) $N_{--}(x) = \exp\{\widehat{\mathcal{K}}(-x)\}$; and iv) $N_{-+}(x) = N_{--}(x)\widehat{\Psi} = \exp\{\widehat{\mathcal{K}}(-x)\}\widehat{\Psi}$.*

The three sets of basic quantities are summarized in Table 3.1. References are given in Table 3.1 for further reading on the basic quantities. Using the basic quantities, the joint stationary distribution of the multi-layer $MMFF$ process can be found in a closed form.

	Solutions	Intuitive Interpretation	Paper
Ψ $(\widehat{\Psi})$	Equation (3.1.9)	Ψ $(\widehat{\Psi})$ contains the transition probabilities of the state of underlying Markov chain Q from an epoch that $X(t)$ starts to increase (decrease) from 0 to the next first epoch that $X(t)$ returns to 0.	[77] [86] [98] [103]
$\mathcal{U}$ $(\widehat{\mathcal{U}})$	Equation (3.1.10)	$\mathcal{U}$ $(\widehat{\mathcal{U}})$ contains the transition rates of the state of the underlying Markov chain Q when the fluid level reaches a new lower (higher) point.	[12]
$\mathcal{K}$ $(\widehat{\mathcal{K}})$	Equation (3.1.14)	$\mathcal{K}$ $(\widehat{\mathcal{K}})$ contains the numbers of visits to certain fluid level and state during some first passage periods.	[98]

Table 3.1: Basic quantities for $MMFF$ processes

There are some relationships between the mean drift ζ and the basic matrices. These relationships are important for derivation and numerical computation.

Lemma 3.4. *([12, 98, 103]) The relationships between ζ and basic quantities are as follows.*

- *If $\zeta > 0$, then we have i) $\Psi\mathbf{e} < \mathbf{e}$ and $\widehat{\Psi}\mathbf{e} = \mathbf{e}$; ii) $\mathcal{U}\mathbf{e} < 0$ and $\widehat{\mathcal{U}}\mathbf{e} = 0$; and iii) $\mathcal{K}$ is singular and $\widehat{\mathcal{K}}$ is non-singular.*

- *If* $\zeta = 0$*, then we have i)* $\Psi\mathbf{e} = \mathbf{e}$ *and* $\widehat{\Psi}\mathbf{e} = \mathbf{e}$*; ii)* $\mathcal{U}\mathbf{e} = 0$ *and* $\widehat{\mathcal{U}}\mathbf{e} = 0$*; and iii)* $\mathcal{K}$ *and* $\widehat{\mathcal{K}}$ *are singular.*

- *If* $\zeta < 0$*, then we have i)* $\Psi\mathbf{e} = \mathbf{e}$ *and* $\widehat{\Psi}\mathbf{e} < \mathbf{e}$*; ii)* $\mathcal{U}\mathbf{e} = 0$ *and* $\widehat{\mathcal{U}}\mathbf{e} < 0$*; and iii)* $\mathcal{K}$ *is non-singular and* $\widehat{\mathcal{K}}$ *is singular.*

For extensions to multi-layer $MMFF$ processes, we need quantities when the processes constrained to an interval, say (a, b). Similar to matrices $(N_+(x))_{i,j}$ and $(N_-(x))_{i,j}$, we define, for $a < x < b$,

- $(N_+^{(a,b)}(x))_{i,j}$ be the expected number of visits to state (x, j) before the process reaches to level a or level b, given that the process started in (a, i) for $i \in \mathcal{S}_+$ (See Figure 3.2). $N_+^{(a,b)}(x)$ can be divided into two sub-blocks $N_{++}^{(a,b)}(x)$ for up-crossings and $N_{+-}^{(a,b)}(x)$ for down-crossings according to $j \in \mathcal{S}_+$ or $j \in \mathcal{S}_-$, respectively.

- $(\widehat{N}_-^{(a,b)}(x))_{i,j}$ be the expected number of visits to state (x, j) before the process reaches to level b or level a, given that the process started in (b, i) for $i \in \mathcal{S}_-$. $\widehat{N}_-^{(a,b)}(x)$ can be divided into two sub-blocks $\widehat{N}_{-+}^{(a,b)}(x)$ for up-crossings and $\widehat{N}_{--}^{(a,b)}(x)$ for down-crossings according to $j \in \mathcal{S}_+$ or $j \in \mathcal{S}_-$, respectively.

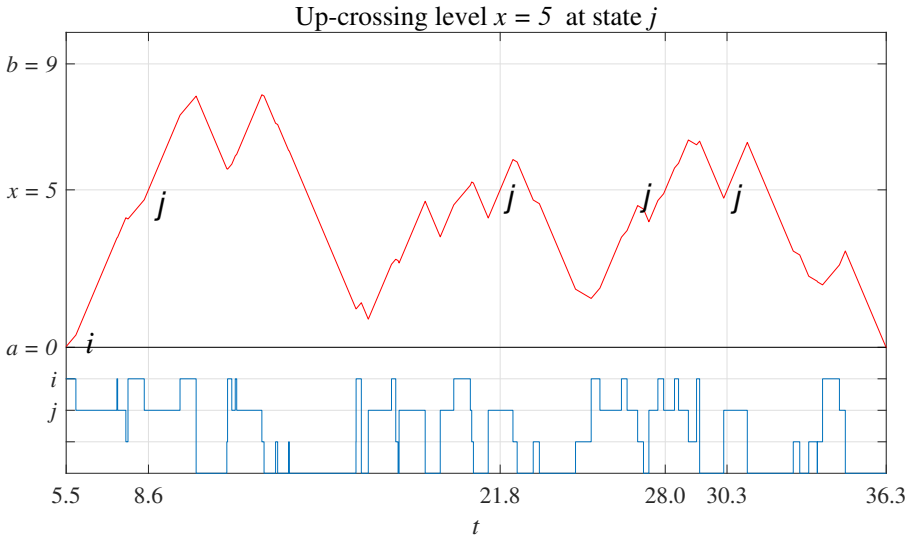

Figure 3.2: Up-crossings of level x, starting from level a, before visiting level a or level b

Lemma 3.5. *([45]) For $a < x < b$, we have*

$$\begin{pmatrix} I & e^{\mathcal{K}(b-a)}\Psi \\ e^{\widehat{\mathcal{K}}(b-a)}\widehat{\Psi} & I \end{pmatrix} \begin{pmatrix} N_+^{(a,b)}(x) \\ \widehat{N}_-^{(a,b)}(x) \end{pmatrix} = \begin{pmatrix} e^{\mathcal{K}(x-a)} & 0 \\ 0 & e^{\widehat{\mathcal{K}}(b-x)} \end{pmatrix} \begin{pmatrix} I & \Psi \\ \widehat{\Psi} & I \end{pmatrix}. \quad (3.1.15)$$

The first matrix on the left hand side in the above equation is invertible if $\zeta \neq 0$.

For the first passage probabilities from one fluid level to another (e.g., from a to b or vice versa), we define, for $a < b$,

- $\Psi_{+-}^{(b-a)}$ is defined similar to Ψ except that the process does not reach fluid level b and the process starts in fluid level a; $\widehat{\Psi}_{-+}^{(b-a)}$ is defined similar to $\widehat{\Psi}$ except that the process does not reach fluid level a and the process starts in fluid level b.
- $\Lambda_{++}^{(b-a)}$ is defined as the probabilities for the process to go from level a to level b before returning to level a. $\widehat{\Lambda}_{--}^{(b-a)}$ is defined as the probabilities for the process to go from level b to level a before returning to level b.

Lemma 3.6. *([45]) The matrices of first passage probabilities satisfy the following equations:*

$$\begin{pmatrix} \Lambda_{++}^{(b-a)} & \Psi_{+-}^{(b-a)} \\ \widehat{\Psi}_{-+}^{(b-a)} & \widehat{\Lambda}_{--}^{(b-a)} \end{pmatrix} \begin{pmatrix} I & \Psi e^{\mathcal{U}(b-a)} \\ \widehat{\Psi} e^{\widehat{\mathcal{U}}(b-a)} & I \end{pmatrix} = \begin{pmatrix} e^{\widehat{\mathcal{U}}(b-a)} & \Psi \\ \widehat{\Psi} & e^{\mathcal{U}(b-a)} \end{pmatrix}. \quad (3.1.16)$$

The second matrix on the left-hand-side of the above equation is invertible if $\zeta \neq 0$.

Lemmas 3.5 and 3.6 are developed for $MMFF$ processes with only one layer in their paper, but they are the foundation of the analysis of multi-layer $MMFF$ processes and will be used repeatedly in the following sections.

3.2 Definition of Multi-Layer *MMFF* Processes

The multi-layer $MMFF$ processes were first introduced in [32]. As a generalization of classical $MMFF$ processes, multi-layer $MMFF$ processes are fluid flow processes in which the

changing rate of the fluid level is modulated by layer-dependent continuous-time Markov chains. A *multi-layer Markov modulated fluid flow (MMFF) process* $\{(X(t), \phi(t)), t \geq 0\}$ consists of following four parts.

1. **Borders and Layers**:

   - There are $N+1$ constants such that $N \geq 1$ and $l_0 = -\infty < l_1 < ... < l_N = \infty$, to be called Borders.
   - Borders form N intervals (l_0, l_1), (l_1, l_2), ..., and (l_{N-1}, l_N), to be called Layer 1, 2, ..., and N, respectively.

2. **Generator and fluid changing rates within Layer n, i.e, $l_{n-1} < X(t) < l_n$, for $n = 1, ..., N$**: (In this part, a classical $MMFF$ process is defined for each layer.)

   - $\{\phi(t), t \geq 0\}$ 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 $c^{(n)}_i$ of fluid level change can be positive, negative, or zero. We put the rates into vectors $\mathbf{c}^{(n)} = \{c^{(n)}_i, i \in \mathcal{S}^{(n)}\}$. 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:

$$\begin{aligned} \mathcal{S}^{(n)}_+ &= \{i \in \mathcal{S}^{(n)} : \ c^{(n)}_i > 0\}; \\ \mathcal{S}^{(n)}_- &= \{i \in \mathcal{S}^{(n)} : \ c^{(n)}_i < 0\}; \\ \mathcal{S}^{(n)}_0 &= \{i \in \mathcal{S}^{(n)} : \ c^{(n)}_i = 0\}. \end{aligned} \tag{3.2.1}$$

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

$$\mathbf{c}^{(n)} = (\mathbf{c}^{(n)}_+, \ \mathbf{c}^{(n)}_-, \ 0); Q^{(n)} = \begin{matrix} \mathcal{S}^{(n)}_+ \\ \mathcal{S}^{(n)}_- \\ \mathcal{S}^{(n)}_0 \end{matrix} \begin{pmatrix} Q^{(n)}_{++} & Q^{(n)}_{+-} & Q^{(n)}_{+0} \\ Q^{(n)}_{-+} & Q^{(n)}_{--} & Q^{(n)}_{-0} \\ Q^{(n)}_{0+} & Q^{(n)}_{0-} & Q^{(n)}_{00} \end{pmatrix}. \tag{3.2.2}$$

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

$$\zeta^{(n)} = \boldsymbol{\alpha}^{(n)}\mathbf{c}^{(n)}, \tag{3.2.3}$$

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

$$C_+^{(n)} = \text{diag}(\mathbf{c}_+^{(n)}); C_-^{(n)} = \text{diag}(\mathbf{c}_-^{(n)}). \tag{3.2.4}$$

3. **Generator on Border n, i.e, $X(t) = l_n$, for $n = 1, ..., N-1$:**

   - $X(t)$ remains at l_n during the period that $\phi(t)$ is in $\mathcal{S}_b^{(n)}$ with sub-generator $Q_{bb}^{(n)}$, 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 $Q_{bb}^{(n)}\mathbf{e} + Q_{b+}^{(n)}\mathbf{e} + Q_{b-}^{(n)}\mathbf{e} = 0$, where $\mathbf{e}$ is the column vector of ones with an appropriate size.

4. **Transitions when reaching Border n, for $n = 1, ..., N-1$:**

   - If $X(t)$ reaches l_n from below, the process $\{\phi(t), t \geq 0\}$ can switch from $\mathcal{S}_+^{(n)}$ to $\mathcal{S}_-^{(n)}$ (i.e., reflecting back to Layer n) with probability $P_{+b-}^{(n)}$; $\mathcal{S}_+^{(n+1)}$ (i.e., passing Border n to Layer $(n+1)$) with probability $P_{+b+}^{(n)}$; $\mathcal{S}_b^{(n)}$ (i.e., entering Border n) with probability $P_{+bb}^{(n)}$.
   - If $X(t)$ reaches l_n from above, the process $\{\phi(t), t \geq 0\}$ can switch from $\mathcal{S}_-^{(n+1)}$ to $\mathcal{S}_+^{(n+1)}$ (i.e., reflecting back to Layer $(n+1)$) with probability $P_{-b+}^{(n)}$; $\mathcal{S}_-^{(n)}$ (i.e., passing Border n to Layer n) with probability $P_{-b-}^{(n)}$; $\mathcal{S}_b^{(n)}$ (i.e., entering Border n) with probability $P_{-bb}^{(n)}$.
   - Note that $P_{+b+}^{(n)}\mathbf{e} + P_{+b-}^{(n)}\mathbf{e} + P_{+bb}^{(n)}\mathbf{e} = \mathbf{e}$; and $P_{-b+}^{(n)}\mathbf{e} + P_{-b-}^{(n)}\mathbf{e} + P_{-bb}^{(n)}\mathbf{e} = \mathbf{e}$.
   - We shall call a border i) a *sticky border* if $\mathcal{S}_b^{(n)}$ is nonempty; ii) a *crossable border* if one of $P_{-b-}^{(n)}$ and $P_{+b+}^{(n)}$ is nonzero; and iii) a *reflective border* if one of $P_{-b+}^{(n)}$ and $P_{+b-}^{(n)}$ is nonzero.

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

$$X(t) = X(0) + \int_0^t c_{\phi(s)}^{(L(X(s)))} \mathrm{d}s, \quad \text{or} \quad \frac{\mathrm{d}X(t)}{\mathrm{d}t} = c_{\phi(t)}^{(L(X(t)))}, \tag{3.2.5}$$

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

The classical $MMFF$ processes are obviously special cases of multi-layer $MMFF$ processes when there is only one layer. Another special case is the classical Markov modulated fluid queue ($MMFQ$) with $N = 2$ and the fluid level truncated at Border $l_1 = 0$.

In general, the multi-layer $MMFF$ process does not have the independent incremental property, and its evolutions in individual layers interact with each other through the borders. On the other hand, it evolves conditionally independently within individual layers. This observation implies that the process can be analyzed separately in individual layers and then all layers are combined together. The study of the process within individual layers is equivalent to that of the classical $MMFF$ process, thus the basic quantities introduced in Section 3.1 are essential for the analysis of multi-layer $MMFF$ processes. Since the generator and fluid changing rates are different for individual layers, we add the superscript "(n)" to the basic quantities for Layer n as $\Psi^{(n)}$, $\widehat{\Psi}^{(n)}$, $\mathcal{U}^{(n)}$, $\widehat{\mathcal{U}}^{(n)}$, $\mathcal{K}^{(n)}$ and $\widehat{\mathcal{K}}^{(n)}$.

Example 3.1. Parameters of a multi-layer $MMFF$ process with $N = 3$ are presented in Table 3.2. In this example, all borders ($l_1 = -2$ and $l_2 = 3$) are sticky, reflective and crossable, which means the fluid can enter the borders and cumulate mass at that level. The generator and fluid changing rates for individual layers are quite different. The basic quantities for this example can be found in Table 3.3. The numerical result satisfies Lemma 3.4. The sample paths can be found in Figure 3.3.

Borders / Layers	**Parameters**
Border ($L_3 = \infty$)	Not defined
Layer 3	$\mathbf{c}^{(3)} = (1.5,\ -3,\ -10,\ 0);\quad Q^{(3)} = \begin{pmatrix} -3 & 1.5 & 1 & 0.5 \\ 2 & -3 & 0 & 1 \\ 2 & 1 & -4 & 1 \\ 0.5 & 0.5 & 0 & -1 \end{pmatrix}.$
Border ($L_2 = 3$)	$Q^{(2)}_{bb} = \begin{pmatrix} -3 & 1 \\ 1 & -2 \end{pmatrix};\quad Q^{(2)}_{b+} = \begin{pmatrix} 1.5 \\ 0.3 \end{pmatrix};\quad Q^{(2)}_{b-} = \begin{pmatrix} 0.5 \\ 0.7 \end{pmatrix};$ $P^{(2)}_{+b+} = \begin{pmatrix} 0.1 \end{pmatrix};\quad P^{(2)}_{+b-} = \begin{pmatrix} 0.4 \end{pmatrix};\quad P^{(2)}_{+bb} = \begin{pmatrix} 0.3 & 0.2 \end{pmatrix};$ $P^{(2)}_{-b+} = \begin{pmatrix} 0.3 \\ 0.3 \end{pmatrix};\quad P^{(2)}_{-b-} = \begin{pmatrix} 0.1 \\ 0.1 \end{pmatrix};\quad P^{(2)}_{-bb} = \begin{pmatrix} 0.2 & 0.4 \\ 0.4 & 0.2 \end{pmatrix}.$
Layer 2	$\mathbf{c}^{(2)} = (5,\ -2.5,\ 0,\ 0);\quad Q^{(2)} = \begin{pmatrix} -2 & 0.5 & 0.5 & 1 \\ 1 & -2 & 1 & 0 \\ 0 & 1 & -1 & 0 \\ 1 & 0 & 0 & -1 \end{pmatrix}.$
Border ($L_1 = -2$)	$Q^{(1)}_{bb} = \begin{pmatrix} -1 \end{pmatrix};\quad Q^{(1)}_{b+} = \begin{pmatrix} 0.2 \end{pmatrix};\quad Q^{(1)}_{b-} = \begin{pmatrix} 0.8 \end{pmatrix};$ $P^{(1)}_{+b+} = \begin{pmatrix} 0.3 \\ 0.5 \end{pmatrix};\quad P^{(1)}_{+b-} = \begin{pmatrix} 0.3 \\ 0.1 \end{pmatrix};\quad P^{(1)}_{+bb} = \begin{pmatrix} 0.4 \\ 0.4 \end{pmatrix};$ $P^{(1)}_{-b+} = \begin{pmatrix} 0.2 \end{pmatrix};\quad P^{(1)}_{-b-} = \begin{pmatrix} 0.3 \end{pmatrix};\quad P^{(1)}_{-bb} = \begin{pmatrix} 0.5 \end{pmatrix}.$
Layer 1	$\mathbf{c}^{(1)} = (20,\ 3,\ -2,\ 0);\quad Q^{(1)} = \begin{pmatrix} -1 & 0 & 1 & 0 \\ 0 & -1 & 0 & 1 \\ 1 & 2 & -5 & 2 \\ 1 & 0 & 1 & -2 \end{pmatrix}.$
Border ($L_0 = -\infty$)	Not defined

Table 3.2: Parameters of Example 3.1

Quantities	Layer $n=1$	Layer $n=2$	Layer $n=3$
$\zeta^{(n)}$	10.6667	1.2500	-1.6905
$\Psi^{(n)}$	$\begin{pmatrix} 0.0257 \\ 0.0766 \end{pmatrix}$	$\begin{pmatrix} 0.5 \end{pmatrix}$	$\begin{pmatrix} 0.5929 & 0.4071 \end{pmatrix}$
$\widehat{\Psi}^{(n)}$	$\begin{pmatrix} 0.5518 & 0.4482 \end{pmatrix}$	$\begin{pmatrix} 1 \end{pmatrix}$	$\begin{pmatrix} 0.4024 \\ 0.1895 \end{pmatrix}$
$\mathcal{U}^{(n)}$	$\begin{pmatrix} -1.8977 \end{pmatrix}$	$\begin{pmatrix} -0.2000 \end{pmatrix}$	$\begin{pmatrix} -0.3393 & 0.3393 \\ 0.2982 & -0.2982 \end{pmatrix}$
$\widehat{\mathcal{U}}^{(n)}$	$\begin{pmatrix} -0.0224 & 0.0224 \\ 0.2586 & -0.2586 \end{pmatrix}$	$\begin{pmatrix} 0 \end{pmatrix}$	$\begin{pmatrix} -1.2375 \end{pmatrix}$
$\mathcal{K}^{(n)}$	$\begin{pmatrix} -0.0243 & 0.0257 \\ 0.2433 & -0.2567 \end{pmatrix}$	$\begin{pmatrix} 0 \end{pmatrix}$	$\begin{pmatrix} -1.2375 \end{pmatrix}$
$\widehat{\mathcal{K}}^{(n)}$	$\begin{pmatrix} -1.8977 \end{pmatrix}$	$\begin{pmatrix} -0.2000 \end{pmatrix}$	$\begin{pmatrix} -0.3638 & 0.2683 \\ 0.3711 & -0.2736 \end{pmatrix}$
$\Psi_{+-}^{(l_n-l_{n-1})}$	NA	$\begin{pmatrix} 0.3873 \end{pmatrix}$	NA
$\widehat{\Psi}_{-+}^{(l_n-l_{n-1})}$	NA	$\begin{pmatrix} 0.7746 \end{pmatrix}$	NA
$\Lambda_{++}^{(l_n-l_{n-1})}$	NA	$\begin{pmatrix} 0.6127 \end{pmatrix}$	NA
$\widehat{\Lambda}_{--}^{(l_n-l_{n-1})}$	NA	$\begin{pmatrix} 0.2254 \end{pmatrix}$	NA

Table 3.3: Basic quantities for Example 3.1

3.3 Joint Stationary Distribution

In this section, we present the solution for the joint stationary distribution of the fluid level and the state of the underlying Markov chain. Define, for $-\infty < x < \infty$,

$$\begin{aligned} p_j^{(n)} &= \lim_{t\to\infty} P\{X(t)=l_n,\ \phi(t)=j \mid X(0),\phi(0)\}, \quad \text{for } j\in\mathcal{S}_b^{(n)},\ n=1,2,...,N-1; \\ g_j^{(n)}(x) &= \lim_{t\to\infty} P\{X(t)<x,\ \phi(t)=j \mid X(0),\phi(0)\}, \quad \text{for } j\in\mathcal{S}^{(n)},\ n=1,2,...,N; \\ \pi_j^{(n)}(x) &= \frac{\mathrm{d}g_j^{(n)}(x)}{\mathrm{d}x}, \qquad\qquad\qquad\qquad\qquad\quad \text{for } j\in\mathcal{S}^{(n)},\ n=1,2,...,N. \end{aligned} \tag{3.3.1}$$

Let,

$$\begin{array}{rcll} \mathbf{p}^{(n)} & = & (p_j^{(n)} : \ j \in \mathcal{S}_b^{(n)}), & \text{for } n = 1, 2, ..., N-1; \\ \boldsymbol{\pi}^{(n)}(x) & = & (\pi_j^{(n)}(x) : \ j \in \mathcal{S}^{(n)}), & \text{for } n = 1, 2, ..., N \ \text{ and } -\infty < x < \infty. \end{array} \tag{3.3.2}$$

Our analysis consists of three steps. Step 1: We use semi-Markov chain theory to get the relationship between the density function $\boldsymbol{\pi}^{(n)}(x)$, the border probabilities $\{\mathbf{p}^{(1)}, ..., \mathbf{p}^{(N-1)}\}$, and an integral of a conditional density function in Subsection 3.3.1; Step 2: Construct a censored CTMC to find the border probabilities $\{\mathbf{p}^{(1)}, ..., \mathbf{p}^{(N-1)}\}$ and use the border probabilities to get the coefficients of the density function in Subsection 3.3.2; Step 3 (Subsection 3.3.3): Put things together to find the closed form expressions for the stationary joint density function.

3.3.1 Density Function and Level-Crossing Numbers

Let $f_j(x,t)$ be the density at the state (x,j) at time t, given the initial state $(X(0), \phi(0))$, and define two taboo conditional density functions as follows

- $\gamma_{k,j}^{(n)}(l_{n-1}, x, t)$ be the taboo conditional density of (x,j) at time t, avoiding both Border l_{n-1} and Border l_n in the time interval $(0,t)$, given that the initial state is (l_{n-1}, k), for $l_{n-1} < x < l_n$;
- $\gamma_{k,j}^{(n)}(l_n, x, t)$ be the taboo conditional density of (x,j) at time t, avoiding both Border l_{n-1} and Border l_n in the time interval $(0,t)$, given that the initial state is (l_n, k), for $l_{n-1} < x < l_n$.

We note that $f_j(x,t)h \approx P\{x < X(t) < x+h, \phi(t) = j\}$ for initial condition $(X(0), \phi(0))$, and $\gamma_{k,j}^{(n)}(y, x, t)h$ is approximately the taboo conditional probability that the fluid level is in $(x, x+h)$ at time t.

For $l_{n-1} < x < l_n$, we condition on the state at which the process is either in Border l_{n-1} or l_n for the last time before reaching state (x,j) at time t. After that time point, denoted as $t-\tau$, the process will be between the two borders until it reaches (x,j) at t (see

Figure 3.3). At the point $t-\tau$, the fluid level either touches one of the borders and enters into the interval (l_{n-1}, l_n) or goes from one of the two borders into the interval (l_{n-1}, l_n), a total of six cases. The corresponding probabilities for the occurrence for the six cases are given approximately as follows.

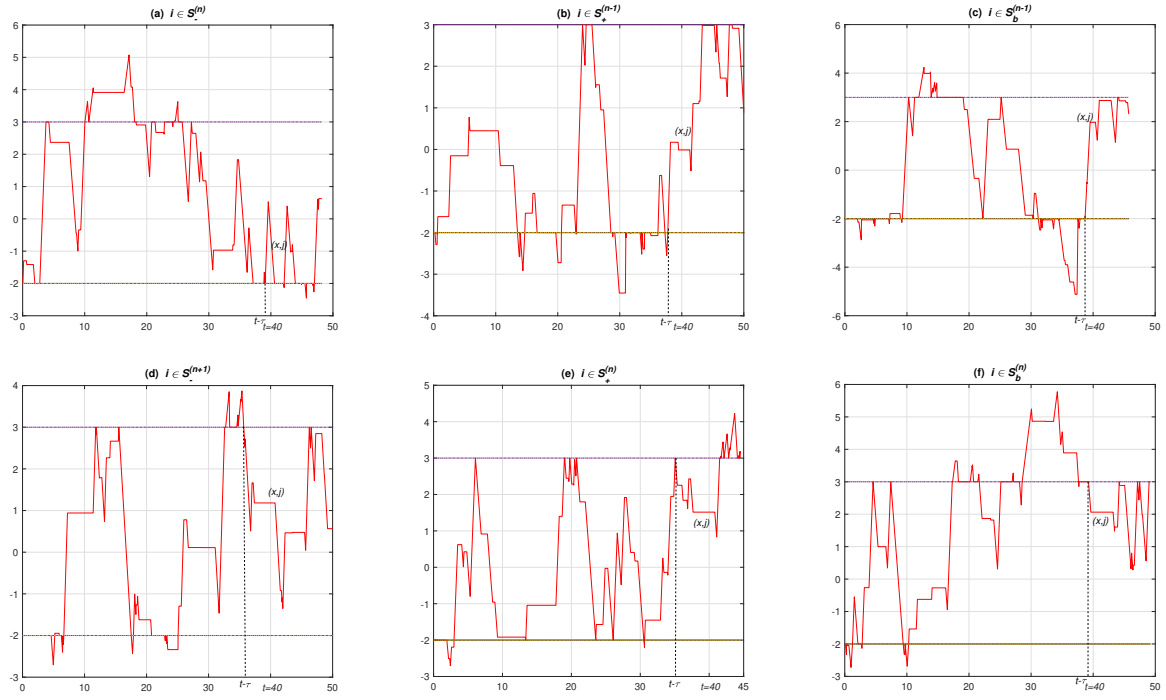

Figure 3.3: The fluid process in $(0,t)$ with $(X(t),\phi(t)) = (x,j)$ and the time epoch $t-\tau$

1. Approaching Border l_{n-1} from above in state $i \in \mathcal{S}_-^{(n)}$, the probability is given by $\frac{P_{X(0),\phi(0)}\{l_{n-1}<X(t-\tau)<l_{n-1}+c_i\mathrm{d}\tau,\phi(t-\tau)=i\}}{c_i^{(n)}\mathrm{d}\tau} c_i^{(n)}\mathrm{d}\tau$, which can be written in density function as: $f_i(l_{n-1}+,t-\tau)c_i^{(n)}\mathrm{d}\tau$. Then the process can be reflected at Border l_{n-1} at epoch $t-\tau$ with (matrix) probability $P_{-b+}^{(n-1)}$. (Remark: In state i, when the time elapses $\mathrm{d}\tau$ units, the fluid level changes by $c_i\mathrm{d}\tau$. That is why we need to use $c_i\mathrm{d}\tau$, instead of only $\mathrm{d}\tau$ in the expression.) (See Figure 3.3(a).)

2. Approaching Border l_{n-1} from below in state $i \in \mathcal{S}_+^{(n-1)}$, the probability is given by $\frac{P_{X(0),\phi(0)}\{l_{n-1}-c_i\mathrm{d}\tau<X(t-\tau)<l_{n-1},\phi(t-\tau)=i\}}{c_i^{(n-1)}\mathrm{d}\tau} c_i^{(n-1)}\mathrm{d}\tau$, which can be written in density func-

tion as $f_i(l_{n-1}-, t-\tau)c_i^{(n-1)}\mathrm{d}\tau$. Then the process can upcross Border l_{n-1} at epoch $t-\tau$ with (matrix) probability $P_{+b+}^{(n-1)}$. (See Figure 3.3(b).)

3. Leaving Border l_{n-1} from state $i \in \mathcal{S}_b^{(n-1)}$, the probability is given by $p_i^{(n-1)}$. Then the process can enter Layer n at epoch $t-\tau$ with (matrix) probability $Q_{b+}^{(n-1)}\mathrm{d}\tau$. (See Figure 3.3(c).)

4. Approaching Border l_n from above in state $i \in \mathcal{S}_-^{(n+1)}$, the probability is given by $\frac{P_{X(0),\phi(0)}\{l_n < X(t-\tau) < l_n + c_i\mathrm{d}\tau, \phi(t-\tau)=i\}}{c_i^{(n+1)}\mathrm{d}\tau} c_i^{(n+1)}\mathrm{d}\tau$, which can be written in density function as $f_i(l_n+, t-\tau)c_i^{(n+1)}\mathrm{d}\tau$. Then the process can downcross Border l_n at epoch $t-\tau$ with (matrix) probability $P_{-b-}^{(n)}$. (See Figure 3.3(d).)

5. Approaching Border l_n from below in state $i \in \mathcal{S}_+^{(n)}$, the probability is given by $\frac{P_{X(0),\phi(0)}\{l_n - c_i\mathrm{d}\tau < X(t-\tau) < l_n, \phi(t-\tau)=i\}}{c_i^{(n)}\mathrm{d}\tau} c_i^{(n)}\mathrm{d}\tau$, which can be written in density function as $f_i(l_n-, t-\tau)c_i^{(n)}\mathrm{d}\tau$. Then the process can be reflected at Border l_n at epoch $t-\tau$ with (matrix) probability $P_{+b-}^{(n)}$. (See Figure 3.3(e).)

6. Leaving Border l_n from state $i \in \mathcal{S}_b^{(n)}$, the probability is given by $p_i^{(n)}$. Then the process can enter Layer n at epoch $t-\tau$ with (matrix) probability $Q_{b-}^{(n)}\mathrm{d}\tau$. (See Figure 3.3(f).)

Using the arguments given in [46], given $(X(0), \phi(0))$, and conditioning on the state

change (i.e., $i \to k$) at epoch $t-\tau$, we have, for $l_{n-1} < x < l_n$,

$$
\begin{aligned}
& f_j(x,t)h \\
= & \sum_{i\in\mathcal{S}_-^{(n)}} \sum_{k\in\mathcal{S}_+^{(n)}} \int_0^t f_i(l_{n-1}+, t-\tau) c_i^{(n)} (P_{-b+}^{(n-1)})_{i,k} \gamma_{k,j}^{(n)}(l_{n-1}, x, \tau) h \mathrm{d}\tau \\
& + \sum_{i\in\mathcal{S}_+^{(n-1)}} \sum_{k\in\mathcal{S}_+^{(n)}} \int_0^t f_i(l_{n-1}-, t-\tau) c_i^{(n-1)} (P_{+b+}^{(n-1)})_{i,k} \gamma_{k,j}^{(n)}(l_{n-1}, x, \tau) h \mathrm{d}\tau \\
& + \sum_{i\in\mathcal{S}_b^{(n-1)}} \sum_{k\in\mathcal{S}_+^{(n)}} \int_0^t p_i^{(n-1)} (Q_{b+}^{(n-1)})_{i,k} \gamma_{k,j}^{(n)}(l_{n-1}, x, \tau) h \mathrm{d}\tau \\
& + \sum_{i\in\mathcal{S}_-^{(n+1)}} \sum_{k\in\mathcal{S}_-^{(n)}} \int_0^t f_i(l_n+, t-\tau) c_i^{(n+1)} (P_{-b-}^{(n)})_{i,k} \gamma_{k,j}^{(n)}(l_n, x, \tau) h \mathrm{d}\tau \\
& + \sum_{i\in\mathcal{S}_+^{(n)}} \sum_{k\in\mathcal{S}_-^{(n)}} \int_0^t f_i(l_n-, t-\tau) c_i^{(n)} (P_{+b-}^{(n)})_{i,k} \gamma_{k,j}^{(n)}(l_n, x, \tau) h \mathrm{d}\tau \\
& + \sum_{i\in\mathcal{S}_b^{(n)}} \sum_{k\in\mathcal{S}_-^{(n)}} \int_0^t p_i^{(n)} (Q_{b-}^{(n)})_{i,k} \gamma_{k,j}^{(n)}(l_n, x, \tau) h \mathrm{d}\tau \\
& + g_j(x,t)h + o(h),
\end{aligned}
\tag{3.3.3}
$$

where $g_j(x,t)$ is the conditional density such that the fluid level is always in Layer n in $(0,t)$. Recall that $f_j(x,t)h \approx P\{x < X(t) < x+h, \phi(t) = j\}$ for initial condition $(X(0), \phi(0))$, and $\gamma_{j,k}^{(n)}(y,x,t)h$ is approximately the taboo conditional probability that the fluid level is in $(x, x+h)$ at time t. We have the term $o(h)$ because in a short period of time $h/c_j^{(n)}$, there can still be more than one transitions occurring. The sum of the probabilities of all those events is $o(h)$.

We assume that $\zeta^{(1)} > 0$, $\zeta^{(N)} < 0$, and the process is irreducible. Then the stochastic process is ergodic. Consequently, the joint stationary distribution exists, is given by the limit of Equation (3.3.3), and is independent of the initial status at $t = 0$. Letting $h \to 0$ and $t \to \infty$ in Equation (3.3.3), in matrix form, we obtain:

Theorem 3.1. *([66]) We assume that $\zeta^{(1)} > 0$, $\zeta^{(N)} < 0$, and the process is irreducible. Then the joint stationary distribution exists. For $l_{n-1} < x < l_n$ and $n = 1, 2, ..., N$, we*

have

$$\boldsymbol{\pi}^{(n)}(x) = \mathbf{w}_L^{(n)} \int_0^\infty \gamma^{(n)}(l_{n-1}, x, s)\mathrm{d}s + \mathbf{w}_U^{(n)} \int_0^\infty \gamma^{(n)}(l_n, x, s)\mathrm{d}s, \tag{3.3.4}$$

where

$$\begin{aligned} \mathbf{w}_L^{(n)} &= \boldsymbol{\pi}_-^{(n)}(l_{n-1})C_-^{(n)}P_{-b+}^{(n-1)} + \boldsymbol{\pi}_+^{(n-1)}(l_{n-1})C_+^{(n-1)}P_{+b+}^{(n-1)} + \mathbf{p}^{(n-1)}Q_{b+}^{(n-1)}; \\ \mathbf{w}_U^{(n)} &= \boldsymbol{\pi}_-^{(n+1)}(l_n)C_-^{(n+1)}P_{-b-}^{(n)} + \boldsymbol{\pi}_+^{(n)}(l_n)C_+^{(n)}P_{+b-}^{(n)} + \mathbf{p}^{(n)}Q_{b-}^{(n)}. \end{aligned} \tag{3.3.5}$$

(Note: For notational convenience, we have added $\gamma^{(1)}(l_0, x, s) = 0$ *and* $\gamma^{(M+N)}(l_N, x, s) = 0$ *to the above equation. Recall that the underlying Markov chain* $\{\phi(t), t \geq 0\}$ *is irreducible when the fluid level is in a certain layer.)*

According to Theorem 3.1, to find the joint stationary distribution, we still need the following sets of border probabilities and coefficients in vector form and the two integrals in the above expression:

1. $\{\mathbf{p}^{(n)},\ n = 0, 1, 2, ..., N\}$; (Note that $\mathbf{p}^{(0)} = \mathbf{p}^{(N)} = 0$.)
2. $\{\mathbf{w}_L^{(n)}, \mathbf{w}_U^{(n)}\ n = 1, 2, ..., N\}$;(Note that $\mathbf{w}_L^{(1)} = \mathbf{w}_U^{(N)} = 0$.)
3. $\int_0^\infty \gamma^{(n)}(l_{n-1}, x, s)\mathrm{d}s$ and $\int_0^\infty \gamma^{(n)}(l_n, x, s)\mathrm{d}s$, for $n = 1, 2, ..., N$.

We find those two sets of vectors in the next subsection. The integrals can be calculated by Lemma 3.7.

Lemma 3.7. *([66]) Matrices of the integrals satisfy the following equation:*

$$\begin{aligned} &\begin{pmatrix} I & e^{\mathcal{K}^{(n)}b_n}\Psi^{(n)} \\ e^{\widehat{\mathcal{K}}^{(n)}b_n}\widehat{\Psi}^{(n)} & I \end{pmatrix} \begin{pmatrix} \int_0^\infty \gamma^{(n)}(l_{n-1}, x, s)\mathrm{d}s \\ \int_0^\infty \gamma^{(n)}(l_n, x, s)\mathrm{d}s \end{pmatrix} \\ &= \begin{pmatrix} e^{\mathcal{K}^{(n)}(x-l_{n-1})} & 0 \\ 0 & e^{\widehat{\mathcal{K}}^{(n)}(l_n-x)} \end{pmatrix} \begin{pmatrix} (C_+^{(n)})^{-1} & \Psi^{(n)}(C_-^{(n)})^{-1} & \Gamma^{(n)} \\ \widehat{\Psi}^{(n)}(C_+^{(n)})^{-1} & (C_-^{(n)})^{-1} & \widehat{\Gamma}^{(n)} \end{pmatrix}, \end{aligned} \tag{3.3.6}$$

where $b_n = l_n - l_{n-1}$, denote the width of the n-th layer, and

$$\begin{aligned} \Gamma^{(n)} &= \left((C_+^{(n)})^{-1}Q_{+0}^{(n)} + \Psi^{(n)}(C_-^{(n)})^{-1}Q_{-0}^{(n)}\right)(-Q_{00}^{(n)})^{-1}; \\ \widehat{\Gamma}^{(n)} &= \left(\widehat{\Psi}^{(n)}(C_+^{(n)})^{-1}Q_{+0}^{(n)} + (C_-^{(n)})^{-1}Q_{-0}^{(n)}\right)(-Q_{00}^{(n)})^{-1}. \end{aligned} \tag{3.3.7}$$

If $\zeta^{(n)} \neq 0$, the first matrix on the left hand side of Equation (3.3.6) is invertible.

Lemma 3.7 is derived from Lemma 3.5 by multiplying the number of visits ($N_+^{(l_{n-1},l_n)}(x)$ and $\widehat{N}_-^{(l_{n-1},l_n)}(x)$) with the time length to generate one unit of fluid ($(C_+^{(n)})^{-1}$ and $(C_-^{(n)})^{-1}$), with the consideration of $\mathcal{S}_0^{(n)}$ (related to $\Gamma^{(n)}$ and $\widehat{\Gamma}^{(n)}$).

3.3.2 Border Probabilities and Coefficients

In this subsection, we want to find out the border probabilities $\{\mathbf{p}^{(1)}, ..., \mathbf{p}^{(N-1)}\}$ and the coefficients $\{\mathbf{w}_L^{(n)}, \mathbf{w}_U^{(n)}\ n = 1, 2, ..., N\}$. For that purpose, we have three steps:

1. Construct an embedded discrete time Markov chain with the border states as absorption states to find out which border it will enter after the process leaving a border;
2. Build a continuous time Markov chain by censoring out the periods that the original $MMFF$ process is between borders to find out the border probabilities $\{\mathbf{p}^{(1)}, ..., \mathbf{p}^{(N-1)}\}$;
3. Use the border probabilities and the embedded discrete time Markov chain to find out the coefficients $\{\mathbf{w}_L^{(n)}, \mathbf{w}_U^{(n)}\ n = 1, 2, ..., N\}$.

Step 1: We construct a discrete time Markov chain such that the border states are absorption states. We first define two fictitious sets of states for the n-th border: i) a set of states for leaving the border by increasing the fluid level: which is $\mathcal{S}_+^{(n+1)}$; and ii) a set of states for leaving the border by decreasing the fluid level: which is $\mathcal{S}_-^{(n)}$. Plus the border states $\mathcal{S}_b^{(n)}$, we have three sets of states associated with each border. We arrange the states in the order: $(\mathcal{S}_-^{(1)}, \mathcal{S}_+^{(2)}, \mathcal{S}_-^{(2)}, \mathcal{S}_+^{(3)}, ..., \mathcal{S}_-^{(N-1)}, \mathcal{S}_+^{(N)}, \mathcal{S}_b^{(1)}, ..., \mathcal{S}_b^{(N-1)})$. The embedded discrete time Markov chain is defined at the time epochs the $MMFF$ process is

leaving (e.g., up-crossing, down-crossing, reflecting, and entering) a border. The transition probability matrix $\mathcal{D}$ of the Markov chain has the following structure:

$$\mathcal{D} = \begin{pmatrix} A & B \\ 0 & I \end{pmatrix}, \tag{3.3.8}$$

where matrix A contains all the transition blocks from $\{\mathcal{S}_-^{(1)}, \mathcal{S}_+^{(2)}, \mathcal{S}_-^{(2)}, \mathcal{S}_+^{(3)}, ..., \mathcal{S}_-^{(N-1)}, \mathcal{S}_+^{(N)}\}$ to themselves, and matrix B contains all the transition blocks from $\{\mathcal{S}_-^{(1)}, \mathcal{S}_+^{(2)}, \mathcal{S}_-^{(2)}, \mathcal{S}_+^{(3)}, ..., \mathcal{S}_-^{(N-1)}, \mathcal{S}_+^{(N)}\}$ to $\{\mathcal{S}_b^{(1)}, ..., \mathcal{S}_b^{(N-1)}\}$. The transition blocks in A and B can be expressed explicitly by the basic quantities as follows.

- From $\mathcal{S}_-^{(n)}$ (i.e., the set below the n-th border), the process can
  1. return to itself (i.e., $\mathcal{S}_-^{(n)}$) with (matrix) probability $\widehat{\Psi}_{-+}^{(l_n - l_{n-1})} P_{+b-}^{(n)}$, (Note: If $n = 1$ (below) or $n = N$ (above), we should use the unbounded $\widehat{\Psi}$ and Ψ to replace $\widehat{\Psi}_{-+}^{(l_n - l_{n-1})}$ and $\Psi_{+-}^{(l_n - l_{n-1})}$, respectively.)
  2. go to the set above the n-th border (i.e., $\mathcal{S}_+^{(n+1)}$) with probability $\widehat{\Psi}_{-+}^{(l_n - l_{n-1})} P_{+b+}^{(n)}$,
  3. enter the n-th border (i.e., $\mathcal{S}_b^{(n)}$) with probability $\widehat{\Psi}_{-+}^{(l_n - l_{n-1})} P_{+bb}^{(n)}$,
  4. go to the set above the $(n-1)$-st border (i.e., $\mathcal{S}_+^{(n)}$) with probability $\widehat{\Lambda}_{--}^{(l_n - l_{n-1})} P_{-b+}^{(n-1)}$,
  5. go to the set below the $(n-1)$-st border (i.e., $\mathcal{S}_-^{(n-1)}$) with probability $\widehat{\Lambda}_{--}^{(l_n - l_{n-1})} P_{-b-}^{(n-1)}$, and
  6. enter the $(n-1)$-st border (i.e., $\mathcal{S}_b^{(n-1)}$) with probability $\widehat{\Lambda}_{--}^{(l_n - l_{n-1})} P_{-bb}^{(n-1)}$.
- From $\mathcal{S}_+^{(n+1)}$ (i.e., the set above the n-th border), the process can
  1. return to itself (i.e., $\mathcal{S}_+^{(n+1)}$) with probability $\Psi_{+-}^{(l_{n+1} - l_n)} P_{-b+}^{(n)}$,
  2. go to the set below the n-th border (i.e., $\mathcal{S}_-^{(n)}$) with probability $\Psi_{+-}^{(l_{n+1} - l_n)} P_{-b-}^{(n)}$,
  3. enter the n-th border (i.e., $\mathcal{S}_b^{(n)}$) with probability $\Psi_{+-}^{(l_{n+1} - l_n)} P_{+bb}^{(n)}$,
  4. go to the set above the $(n+1)$-st border (i.e., $\mathcal{S}_+^{(n+2)}$) with probability $\Lambda_{++}^{(l_{n+1} - l_n)} P_{+b+}^{(n+1)}$,
  5. go to the set below the $(n+1)$-st border (i.e., $\mathcal{S}_-^{(n+1)}$) with probability $\Lambda_{++}^{(l_{n+1} - l_n)} P_{+b-}^{(n+1)}$, and

6. enter the $(n+1)$-st border (i.e., $\mathcal{S}_b^{(n+1)}$) with probability $\Lambda_{++}^{(l_{n+1}-l_n)} P_{+bb}^{(n+1)}$.

The absorption probabilities from those "leaving border" sets to the border sets can be obtained by

$$(I-A)^{-1}B = \begin{array}{c} \\ \vdots \\ \mathcal{S}_-^{(m)} \\ \mathcal{S}_+^{(m+1)} \\ \vdots \end{array} \begin{array}{c} \mathcal{S}_b^{(n)} \\ \begin{pmatrix} \vdots \\ \cdots \quad H_{-b}^{(m,n)} \quad \cdots \\ \cdots \quad H_{+b}^{(m,n)} \quad \cdots \\ \vdots \end{pmatrix} \end{array}, \tag{3.3.9}$$

where $H_{-b}^{(m,n)}$ contains the probabilities that the first border entered by the original $MMFF$ process, after leaving the m-th border by decreasing in the set $\mathcal{S}_-^{(m)}$, is $\mathcal{S}_b^{(n)}$, and $H_{+b}^{(m,n)}$ contains the probabilities that the first border entered by the original $MMFF$ process, after leaving the m-th border by increasing in the set $\mathcal{S}_+^{(m+1)}$, is $\mathcal{S}_b^{(n)}$.

Step 2: We build a continuous time Markov chain $\mathcal{Q}_{\mathbf{p}}$ by censoring out the periods that the original $MMFF$ process is between borders. Thus, the state space of $\mathcal{Q}_{\mathbf{p}}$ constitutes (only) all the border states $\mathcal{S}_b^{(1)} \cup \mathcal{S}_b^{(2)} \cup ... \cup \mathcal{S}_b^{(N-1)}$. The infinitesimal generator $\mathcal{Q}_{\mathbf{p}}$ can be divided into blocks as follow:

$$\mathcal{Q}_{\mathbf{p}} = \begin{array}{c} \\ \mathcal{S}_b^{(m)} \end{array} \begin{array}{c} \mathcal{S}_b^{(n)} \\ \begin{pmatrix} \vdots \\ \cdots \quad Q_{m,n} \quad \cdots \\ \vdots \end{pmatrix} \end{array}, \tag{3.3.10}$$

where, for $m, n = 1, 2, ..., N-1$,

$$Q_{m,n} = \begin{cases} Q_{bb}^{(m)} + Q_{b-}^{(m)} H_{-b}^{(m,m)} + Q_{b+}^{(m)} H_{+b}^{(m,m)}, & \text{if } m = n; \\ Q_{b-}^{(m)} H_{-b}^{(m,n)} + Q_{b+}^{(m)} H_{+b}^{(m,n)}, & \text{if } m \neq n. \end{cases} \tag{3.3.11}$$

Let $\mathbf{p} = (\mathbf{p}^{(1)}, \mathbf{p}^{(2)}, ..., \mathbf{p}^{(N-1)})$, we have the following result.

Lemma 3.8. *Vector* $\mathbf{p}$ *is proportional to the steady state probability of the process with generator* $\mathcal{Q}_{\mathbf{p}}$.

Proof. We mimic the proof given by Theorem 2.2 in [45], in which the case with only one sticky border is considered. We denote by $\hat{\phi}(t)$ the censored process of $\phi(t)$ observed only when the fluid is in borders. We define $\hat{\theta}_0 = \inf\{t \geq 0 : \phi(t) \in \mathcal{S}_b^{(m)}, m = 1, 2, ..., N-1\}$, and for $n \geq 0$, $\hat{\delta}_n = \inf\{t > \hat{\theta}_n : \ \phi(t) \notin \mathcal{S}_b^{(m)}, m = 1, 2, ..., N-1\}$, $\hat{\theta}_{n+1} = \inf\{t > \hat{\delta}_n : \phi(t) \in \mathcal{S}_b^{(m)}, m = 1, 2, ..., N-1\}$. Define $\Theta_n = \sum_{i=0}^{n}(\hat{\delta}_i - \hat{\theta}_i)$ for $n \geq 0$. Then the process $\hat{\phi}(t)$ evolves in the interval (Θ_{n-1}, Θ_n) exactly like $\phi(t)$ in the interval $(\hat{\theta}_n, \hat{\delta}_n)$. Vector $\mathbf{p}$ is proportional to the steady state probability of the process $\hat{\phi}(t)$ (see [54]). By the construction process of $\mathcal{Q}_{\mathbf{p}}$ given above, we can see that $\mathcal{Q}_{\mathbf{p}}$ is the generator of $\hat{\phi}(t)$, which completes the proof. □

We can first solve the linear system $\mathbf{p}\mathcal{Q}_{\mathbf{p}} = 0$ and $\mathbf{pe} = 1$ for vector $\mathbf{p}$. But vector $\mathbf{p}$ is not the actual border probabilities. We shall further normalize $\mathbf{p}$ to get the actual border probabilities, which will be discussed later.

Step 3: The computation of the coefficients $(\mathbf{w}_L^{(n)}, \mathbf{w}_U^{(n)})$ requires the border probabilities and the matrix A in the embedded discrete time Markov chain in Equation (3.3.8).

Lemma 3.9. *Let* $\boldsymbol{w} = (\mathbf{w}_U^{(1)}, \mathbf{w}_L^{(2)}, \mathbf{w}_U^{(2)}, \mathbf{w}_L^{(3)}, ..., \mathbf{w}_U^{(N-1)}, \mathbf{w}_L^{(N)})$, *we have*

$$\boldsymbol{w} = \boldsymbol{w}A + (\mathbf{p}^{(1)}Q_{b-}^{(1)}, \mathbf{p}^{(1)}Q_{b+}^{(1)}, \mathbf{p}^{(2)}Q_{b-}^{(2)}, ..., \mathbf{p}^{(N-1)}Q_{b-}^{(N-1)}, \mathbf{p}^{(N-1)}Q_{b+}^{(N-1)}). \tag{3.3.12}$$

Proof. First, based on the definition of matrix A, Equation (3.3.12) can be written as a set of linear equations as follows, for $n = 1, 2, ..., N-1$,

$$\begin{aligned}
\mathbf{w}_U^{(n)} &= \mathbf{w}_L^{(n)}\Lambda_{++}^{(l_n - l_{n-1})}P_{+b-}^{(n)} + \mathbf{w}_U^{(n)}\widehat{\Psi}_{-+}^{(l_n - l_{n-1})}P_{+b-}^{(n)} + \mathbf{w}_L^{(n+1)}\Psi_{+-}^{(l_{n+1} - l_n)}P_{-b-}^{(n)} \\
&\quad + \mathbf{w}_U^{(n+1)}\widehat{\Lambda}_{--}^{(l_{n+1} - l_n)}P_{-b-}^{(n)} + \mathbf{p}^{(n)}Q_{b-}^{(n)}; \\
\mathbf{w}_L^{(n+1)} &= \mathbf{w}_L^{(n)}\Lambda_{++}^{(l_n - l_{n-1})}P_{+b+}^{(n)} + \mathbf{w}_U^{(n)}\widehat{\Psi}_{-+}^{(l_n - l_{n-1})}P_{+b+}^{(n)} + \mathbf{w}_L^{(n+1)}\Psi_{+-}^{(l_{n+1} - l_n)}P_{-b+}^{(n)} \\
&\quad + \mathbf{w}_U^{(n+1)}\widehat{\Lambda}_{--}^{(l_{n+1} - l_n)}P_{-b+}^{(n)} + \mathbf{p}^{(n)}Q_{b+}^{(n)},
\end{aligned} \tag{3.3.13}$$

where $\mathbf{w}_L^{(1)} = 0$, $\mathbf{w}_U^{(N)} = 0$.

Essentially, we need to prove these two equations in (3.3.13). For the first equation, by definition in Equation (3.3.5), we have

$$\mathbf{w}_U^{(n)} = \boldsymbol{\pi}_-^{(n+1)}(l_n)C_-^{(n+1)}P_{-b-}^{(n)} + \boldsymbol{\pi}_+^{(n)}(l_n)C_+^{(n)}P_{+b-}^{(n)} + \mathbf{p}^{(n)}Q_{b-}^{(n)}. \tag{3.3.14}$$

We use Equation (3.3.4) in Theorem 3.1 to give the expression of the density limits $\boldsymbol{\pi}_{-}^{(n+1)}(l_n)$ and $\boldsymbol{\pi}_{+}^{(n)}(l_n)$, we have

$$\begin{aligned}
\boldsymbol{\pi}_{-}^{(n+1)}(l_n) &= \mathbf{w}_L^{(n+1)} \int_0^\infty \gamma^{(n+1)}(l_n, l_n, s)\mathrm{d}s + \mathbf{w}_U^{(n+1)} \int_0^\infty \gamma^{(n+1)}(l_{n+1}, l_n, s)\mathrm{d}s, \\
\boldsymbol{\pi}_{+}^{(n)}(l_n) &= \mathbf{w}_L^{(n)} \int_0^\infty \gamma^{(n)}(l_{n-1}, l_n, s)\mathrm{d}s + \mathbf{w}_U^{(n)} \int_0^\infty \gamma^{(n)}(l_n, l_n, s)\mathrm{d}s.
\end{aligned} \tag{3.3.15}$$

Then, we replace these two density limits in (3.3.14), we have

$$\begin{aligned}
\mathbf{w}_U^{(n)} \quad = \quad & \left(\mathbf{w}_L^{(n+1)} \int_0^\infty \gamma^{(n+1)}(l_n, l_n, s)\mathrm{d}s + \mathbf{w}_U^{(n+1)} \int_0^\infty \gamma^{(n+1)}(l_{n+1}, l_n, s)\mathrm{d}s \right) C_-^{(n+1)} P_{-b-}^{(n)} \\
& + \left(\mathbf{w}_L^{(n)} \int_0^\infty \gamma^{(n)}(l_{n-1}, l_n, s)\mathrm{d}s + \mathbf{w}_U^{(n)} \int_0^\infty \gamma^{(n)}(l_n, l_n, s)\mathrm{d}s \right) C_+^{(n)} P_{+b-}^{(n)} \\
& + \mathbf{p}^{(n)} Q_{b-}^{(n)}.
\end{aligned} \tag{3.3.16}$$

Next, we need to evaluate the integrals in the above equation. By the definition of the taboo conditional density functions, we have

$$\begin{aligned}
&\gamma_{k,j}^{(n)}(l_n, l_n, s)h \\
&\quad \approx P\{l_n < X(s) < l_n + h, \phi(t) = j, l_{n-1} < X(t) < l_n, t \in (0, s) | X(0) = l_n, \phi(0) = k\}.
\end{aligned} \tag{3.3.17}$$

If $\phi(s) = j$ when the process approaching Border l_n at time s, the fluid level changing rate is $c_j^{(n)}$. Suppose the time elapses $\mathrm{d}s$, the first return probability from $(X(0) = l_n, \phi(0) = k)$ to $(X(s) = l_n, \phi(s) = j)$ at time s without touching Border l_{n-1} is $\gamma_{k,j}^{(n)}(l_n, l_n, s)c_j^{(n)}\mathrm{d}s$. Integrating the probability with respect to the time s from 0 to ∞ and in matrix form we have $\int_0^\infty \gamma^{(n)}(l_n, l_n, s)\mathrm{d}sC_+^{(n)}$, which gives us the first return probabilities to Border l_n from Border l_n, without touching Border l_{n-1}. It turns out that this integral is equivalent to $\widehat{\Psi}_{-+}^{(l_n - l_{n-1})}$.

Similarly, the integral $\int_0^\infty \gamma^{(n)}(l_{n-1}, l_n, s)\mathrm{d}sC_+^{(n)}$ is equivalent to $\Lambda_{++}^{(l_n - l_{n-1})}$. The integrals $\int_0^\infty \gamma^{(n+1)}(l_n, l_n, s)\mathrm{d}sC_-^{(n+1)}$ and $\int_0^\infty \gamma^{(n+1)}(l_{n+1}, l_n, s)\mathrm{d}sC_-^{(n+1)}$ respectively give us the first return probabilities to Border l_n from Border l_n, without touching Border l_{n+1}, and the first passage probabilities to Border l_n from Border l_{n+1}, without returning to Border l_{n+1}. In matrix form, they are equivalent to $\Psi_{+-}^{(l_{n+1} - l_n)}$ and $\widehat{\Lambda}_{--}^{(l_{n+1} - l_n)}$, respectively. This leads to

the desired first equation in (3.3.13).

The second equation in (3.3.13) can be obtained in the same way, details are omitted. □

3.3.3 Closed Form Expressions

With the assumption that $\zeta^{(1)} > 0$ and $\zeta^{(N)} < 0$, the stationary distribution of the $MMFF$ process exists. The expressions for the density functions and distribution functions can be found with all the vectors in place. We define, for $n = 1, 2, ..., N$,

$$(\mathbf{u}_+^{(n)},\ \mathbf{u}_-^{(n)}) = (\mathbf{w}_L^{(n)},\ \mathbf{w}_U^{(n)}) \begin{pmatrix} I & e^{\mathcal{K}^{(n)} b_n} \Psi^{(n)} \\ e^{\widehat{\mathcal{K}}^{(n)} b_n} \widehat{\Psi}^{(n)} & I \end{pmatrix}^{-1}, \tag{3.3.18}$$

recall $b_n = l_n - l_{n-1}$, denote the width of the n-th layer.

Combining Equation (3.3.18) and Lemma 3.7, we obtain a closed form expression of the joint density function.

Theorem 3.2. *([66]) We assume that $\zeta^{(1)} > 0$, $\zeta^{(N)} < 0$, and $\zeta^{(n)} \neq 0$[1] , for $n = 2, ..., N-1$. For $n = 1, 2, ..., N$, we have, for $l_{n-1} < x < l_n$, the joint density function*

$$\begin{aligned} \boldsymbol{\pi}^{(n)}(x) \ &= \ \mathbf{u}_+^{(n)} e^{\mathcal{K}^{(n)}(x - l_{n-1})} ((C_+^{(n)})^{-1},\ \Psi^{(n)} (C_-^{(n)})^{-1},\ \Gamma^{(n)}) \\ &\quad + \mathbf{u}_-^{(n)} e^{\widehat{\mathcal{K}}^{(n)}(l_n - x)} (\widehat{\Psi}^{(n)} (C_+^{(n)})^{-1},\ (C_-^{(n)})^{-1},\ \widehat{\Gamma}^{(n)}). \end{aligned} \tag{3.3.19}$$

Now, we construct the joint stationary distribution function. Let $\mathbf{G}^{(n)}(x) = \int_{l_{n-1}}^{x} \boldsymbol{\pi}^{(n)}(x) \mathrm{d}x$. We obtain, for $l_{n-1} < x < l_n$ and $n = 1, 2, ..., N$,

$$\begin{aligned} \mathbf{G}^{(n)}(x) \ &= \ \mathbf{u}_+^{(n)} \int_{l_{n-1}}^{x} e^{\mathcal{K}^{(n)}(y - l_{n-1})} \mathrm{d}y \left((C_+^{(n)})^{-1},\ \Psi^{(n)} (C_-^{(n)})^{-1},\ \Gamma^{(n)} \right) \\ &\quad + \mathbf{u}_-^{(n)} \int_{l_{n-1}}^{x} e^{\widehat{\mathcal{K}}^{(n)}(l_n - y)} \mathrm{d}y \left(\widehat{\Psi}^{(n)} (C_+^{(n)})^{-1},\ (C_-^{(n)})^{-1},\ \widehat{\Gamma}^{(n)} \right). \end{aligned} \tag{3.3.20}$$

[1]We note that results for the case with $\zeta^{(n)} = 0$ for some $n = 2, 3, ..., N-1$ are much more involved. We choose not to consider that case. Yet it is an interesting topic for future research.

Finally, we need to normalize the coefficients in the joint density function $\{\mathbf{u}_+^{(n)}, \mathbf{u}_-^{(n)}\}$ and the border probabilities $\mathbf{p}^{(n)}$. By the law of total probability, the normalization factor is given by

$$\begin{aligned} c_{norm} &= \sum_{n=1}^{N-1} \mathbf{p}^{(n)}\mathbf{e} + \sum_{n=1}^{N} \mathbf{u}_+^{(n)} \int_{l_{n-1}}^{l_n} e^{\mathcal{K}^{(n)}(y-l_{n-1})}\mathrm{d}y \left((C_+^{(n)})^{-1}, \Psi^{(n)}(C_-^{(n)})^{-1}, \Gamma^{(n)}\right)\mathbf{e} \\ &\quad + \sum_{n=1}^{N} \mathbf{u}_-^{(n)} \int_{l_{n-1}}^{l_n} e^{\widehat{\mathcal{K}}^{(n)}(l_n-y)}\mathrm{d}y \left(\widehat{\Psi}^{(n)}(C_+^{(n)})^{-1},\ (C_-^{(n)})^{-1},\ \widehat{\Gamma}^{(n)}\right)\mathbf{e}. \end{aligned} \tag{3.3.21}$$

Consequently, we have $\mathbf{u}_+^{(n)} =: \mathbf{u}_+^{(n)}/c_{norm}$, $\mathbf{u}_+^{(n)} =: \mathbf{u}_-^{(n)}/c_{norm}$ and $\mathbf{p}^{(n)} =: \mathbf{p}^{(n)}/c_{norm}$

Many quantities of interest can then be obtained. For example, the m-th moment of the (steady state) fluid level can be obtained as:

$$\begin{aligned} \mathbb{E}[X^m(t)] &= \sum_{n=1}^{N-1} l_n^m \mathbf{p}^{(n)}\mathbf{e} + \sum_{n=1}^{N} \int_{l_{n-1}}^{l_n} x^m \mathrm{d}\mathbf{G}^{(n)}(x)\mathbf{e} \\ &= \sum_{n=1}^{N-1} l_n^m \mathbf{p}^{(n)}\mathbf{e} + \sum_{n=1}^{N} \mathbf{u}_+^{(n)} \int_{l_{n-1}}^{l_n} y^m e^{\mathcal{K}^{(n)}(y-l_{n-1})}\mathrm{d}y \left((C_+^{(n)})^{-1}, \Psi^{(n)}(C_-^{(n)})^{-1}, \Gamma^{(n)}\right)\mathbf{e} \\ &\quad + \sum_{n=1}^{N} \mathbf{u}_-^{(n)} \int_{l_{n-1}}^{l_n} y^m e^{\widehat{\mathcal{K}}^{(n)}(l_n-y)}\mathrm{d}y \left(\widehat{\Psi}^{(n)}(C_+^{(n)})^{-1},\ (C_-^{(n)})^{-1},\ \widehat{\Gamma}^{(n)}\right)\mathbf{e}. \end{aligned} \tag{3.3.22}$$

Then, the mean and variance of the (steady state) fluid level can be easily found respectively by $\mathbb{E}[X(t)]$ and $\mathbb{E}[X^2(t)] - (\mathbb{E}[X(t)])^2$.

The integrals in Equations (3.3.20), (3.3.21), and (3.3.22) can be evaluated by using expressions in Lemma B.1 given in Appendix B.

3.4 Algorithm 1 and Numerical Examples

In this section, we summarize the computation steps for computing the density function, distribution function, and the mean fluid level in Algorithm 1, then we present some

numerical examples.

Algorithm 1: The joint stationary distribution of multi-layer $MMFF$ processes

1. Input Parameters: $\{l_0 = -\infty, l_1, ..., l_{N-1}, l_N = \infty\}$, $\{Q^{(n)}, C_+^{(n)}, C_-^{(n)}\}$, $n = 1, 2, ..., N$, and $\{P_{+b+}^{(n)}, P_{+b0}^{(n)}, P_{+b-}^{(n)}, P_{-b+}^{(n)}, P_{-b0}^{(n)}, P_{-b-}^{(n)}, Q_b^{(n)}, Q_{b+}^{(n)}, Q_{b-}^{(n)}\}$, for $n = 1, 2, ..., N-1$;

2. Compute $\{\Psi^{(n)}, \mathcal{K}^{(n)}, \mathcal{U}^{(n)}, \widehat{\Psi}^{(n)}, \widehat{\mathcal{K}}^{(n)}, \widehat{\mathcal{U}}^{(n)}\}$ for $\{Q^{(n)}, C_+^{(n)}, C_-^{(n)}\}$ by using the algorithm in Appendix A and equations in Section 3.1, for $n = 1, 2, ..., N$; Compute $\{\Gamma^{(n)}, \widehat{\Gamma}^{(n)}\}$ by Equation (3.3.7) for $n = 1, 2, ..., N$;

3. Compute $\{\Psi_{+-}^{(l_n - l_{n-1})}, \widehat{\Psi}_{-+}^{(l_n - l_{n-1})}, \Lambda_{++}^{(l_n - l_{n-1})}, \widehat{\Lambda}_{--}^{(l_n - l_{n-1})}\}$ for $\{Q^{(n)}, C_+^{(n)}, C_-^{(n)}\}$, for $n = 1, 2, ..., N$, by using Equation (3.1.16);

4. Construct matrix A and B (Equation (3.3.8)). Compute $\{H_{-b}^{(m,n)}, H_{+b}^{(m,n)}\}$ for $m, n = 1, 2, ..., N$ by using Equation (3.3.9);

5. Construct $\mathcal{Q}_{\mathbf{p}}$ by using Equations (3.3.10) and (3.3.11); Solve linear system $\mathbf{p}\mathcal{Q}_{\mathbf{p}} = 0$ and $\mathbf{pe} = 1$ for $\{\mathbf{p}_1, ..., \mathbf{p}_{N-1}\}$;

6. Compute $\{\mathbf{w}_U^{(n)}, \mathbf{w}_L^{(n)},\ n = 1, 2, ..., N\}$ by Lemma 3.9;

7. Compute $\{\mathbf{u}_+^{(n)}, \mathbf{u}_-^{(n)},\ n = 1, 2, ..., N\}$ by Equation (3.3.18);

8. Compute c_{norm} by using Equation (3.3.21) and Lemma B.1;

9. Use c_{norm} to normalize $\{\mathbf{p}^{(n)}, n = 1, 2, ..., N-1\}$ and $\{\mathbf{u}_+^{(n)}, \mathbf{u}_-^{(n)},\ n = 1, 2, ..., N\}$;

10. Use the updated vectors and Lemma B.1 to compute the stationary distribution function (Equation (3.3.20)), density function (Equation (3.3.19)), and the moments of the steady state fluid level (Equation (3.3.22)).

Example 3.1. (continued) Applying the algorithm to Example 3.1, we obtained the density function of the fluid level (See Figure 3.4 (a)) and calculated the mean fluid level at $\mathbb{E}[X(t)] = 1.3443$ and the variance of the fluid level $Var[X(t)] = 4.2857$. For this three-layer $MMFF$ process, the density function changes drastically at the two borders.

If we make all the borders in Example 3.1 non-sticky and change the passing and reflecting probabilities a little bit. The density function of the fluid level is in Figure 3.4(b). Since generators and fluid changing rates in layers are unchanged, all the basic quantities in Table 3.3 remain the same, but the mean and variance of the fluid level become $\mathbb{E}[X(t)] = 1.1927$ and $Var[X(t)] = 3.4793$.

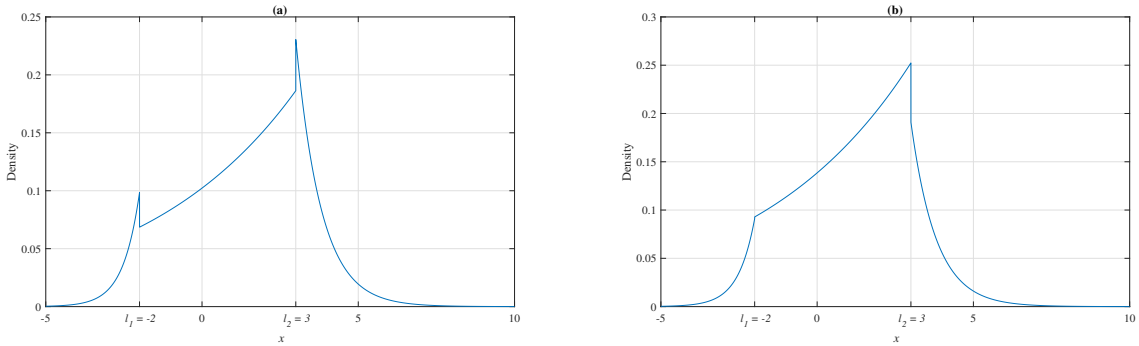

Figure 3.4: The density functions of two multi-layer $MMFF$ processes

Example 3.2. To demonstrate the ability of our algorithm to handle a moderately large number of layers, We present an example with $N = 102$. All borders ($l_1 = -50, l_2 = -49, ..., l_{50} = -1, l_{51} = 0, l_{52} = 1, ..., l_{101} = 50$) are sticky, reflective and crossable. The generator and fluid changing rates for individual layers and borders are presented in Table 3.4. The mean and variance of the fluid level are $\mathbb{E}[X(t)] = -0.1785$ and $Var[X(t)] = 67.8507$. Figure 3.5 shows the variety of the density functions that can be generated by multi-layer $MMFF$ processes.

Borders / Layers	**Parameters**
Layer n, for $n = 1, 2, ..., N$	$\mathbf{c}^{(n)} = (33 - 0.3n,\ 23 - 0.2n,\ -10 - 0.3n,\ -5 - 0.4n,\ 0,\ 0);$ $Q^{(n)} =$ $\begin{pmatrix} -1-0.1n & 0 & 0.5+0.3n & 0 & 0.5+0.7n & 0 \\ 0 & -1-0.3n & 0 & 0.5+0.1n & 0.5+0.2n & 0 \\ 1+0.1n & 0 & -2-0.2n & 0 & 1+0.1n & 0 \\ 0 & 1+0.2n & 0 & -2-0.5n & 0 & 1+0.3n \\ 0 & 1+0.1n & 1+0.1n & 0 & -2-0.2n & 0 \\ 1+0.1n & 0 & 0 & 1+0.1n & 0 & -2-0.2n \end{pmatrix}.$
Border n, for $n = 1, 2, ..., N-1$	$Q_{bb}^{(n)} = \begin{pmatrix} -2n \end{pmatrix};\ Q_{b+}^{(n)} = \begin{pmatrix} 0.5n & 0.5n \end{pmatrix};\ Q_{b-}^{(n)} = \begin{pmatrix} 0.5n & 0.5n \end{pmatrix};$ $P_{+bb}^{(n)} = \begin{pmatrix} 0.2 \\ 0.2 \end{pmatrix};\ P_{+b+}^{(n)} = \begin{pmatrix} 0.2 & 0.2 \\ 0.2 & 0.2 \end{pmatrix};\ P_{+b-}^{(n)} = \begin{pmatrix} 0.2 & 0.2 \\ 0.2 & 0.2 \end{pmatrix};$ $P_{-bb}^{(n)} = \begin{pmatrix} 0.2 \\ 0.2 \end{pmatrix};\ P_{-b+}^{(n)} = \begin{pmatrix} 0.2 & 0.2 \\ 0.2 & 0.2 \end{pmatrix};\ P_{-b-}^{(n)} = \begin{pmatrix} 0.2 & 0.2 \\ 0.2 & 0.2 \end{pmatrix}.$

Table 3.4: Parameters of Example 3.2

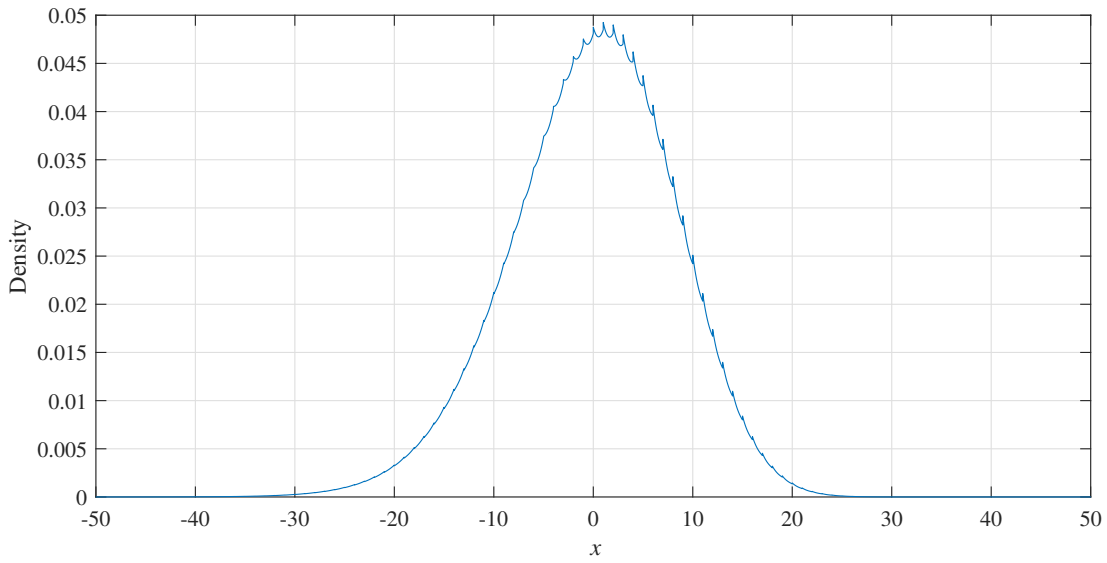

Figure 3.5: The density function of the multi-layer $MMFF$ process in Example 3.2

3.5 Summary

In this chapter, we discuss the basic theory on multi-layer $MMFF$ processes in which the fluid changing rate can be any real value and all borders can be sticky, reflective and crossable. We develop Algorithm 1 to compute the joint stationary distribution. This algorithm works well for small/moderate size problems. For large size problems, the algorithm has to be modified in order to reduce the size of the state space. For instance, computation of border probabilities $\mathbf{p}$ has some dimensionality issue since the matrix $\mathcal{Q}_{\mathbf{p}}$ can be too big for numerical evaluation. On the other hand, the state space used in computation can be drastically reduced for many cases by taking advantage of some special structures of the $MMFF$ processes.

Our main contributions in this chapter are two-fold. First, we review and refine the existing theory on multi-layer $MMFF$ processes. We consider all possible transitions (i.e., crossing, reflecting and entering) on borders. Second, we develop an efficient algorithm for computing the joint stationary distributions. Although many existing algorithms have been studied in the literature, we improve some computational steps (e.g., Lemma 3.9 simplifies the computational steps for those coefficients in [66]) and provide a clear and easy to implement algorithm.

In the following applications to queueing models, we utilize a special type of $MMFF$ processes, which is called *canonical fluid flow process*. The fluid changing rate of this special type of $MMFF$ processes can only be 1 or -1. Therefore, there is no need to consider $\mathcal{S}_0^{(n)}$ in computations since they are empty, and there is no need to construct and use $C_+^{(n)}$ and $C_-^{(n)}$ in computation (only multiplication involved) since they are identity matrices. In addition, some other special structures and technical computational issues will be discussed and Algorithm 1 will be modified to make the algorithm numerically more efficient in the following chapters.

Chapter 4

The $MAP/PH/K{+}GI$ Queue

In this chapter, we study the $MAP/PH/K + GI$ queueing model by the multi-layer $MMFF$ processes developed in Chapter 3. Because MAP can approximate any arrival process, phase-type random variables can approximate any non-negative random variables, and the abandonment time is a random variable with the general discrete distribution, the model is a very general queueing system. We develop an efficient algorithm for computing the steady state waiting times distributions, abandonment probabilities and queue lengths. Some of the quantities are difficult to obtain by other methods.

This chapter is organized as follows. In Section 4.1, we first introduce the queueing model explicitly. In Section 4.2, we introduce a Markov process associated with the age of the customer at the head of the waiting queue, to be called *the age process*. Based on the age process, we introduce a multi-layer $MMFF$ process and present an algorithm for the stationary distribution of the age process in Section 4.3. In Section 4.4, computational procedures are developed for a number of queueing quantities. Numerical examples are presented in Section 4.5.

4.1 Definitions

In this section, we define the multi-server queueing model with random customer abandonment time. Upon arrivals, all customers join a single queue with the first-come-first-serve discipline. There are K identical servers. When the waiting time of a customer reaches (random) time τ, the customer leaves the system without service.

i) The arrival process of customers follows a continuous time Markovian arrival process (MAP) (D_0, D_1), where D_0 and D_1 are square matrices of order m_a. Intuitively, D_0 contains the transition rates without arrival and D_1 contains the transition rates with one arrival. The underlying Markov chain of the arrival process $\{I_a(t), t \geq 0\}$ has an irreducible infinitesimal generator $D = D_0 + D_1$. The stationary distribution $\boldsymbol{\theta}_a$ of the underlying Markov chain satisfies $\boldsymbol{\theta}_a D = 0$ and $\boldsymbol{\theta}_a \mathbf{e} = 1$. The (average) customer arrival rate is given by $\lambda = \boldsymbol{\theta}_a D_1 \mathbf{e}$.

ii) All customers join a single queue waiting for service based on the first-come-first-serve discipline. If a customer's waiting time reaches random time τ, the customer leaves the system immediately without service. The abandonment time τ has a discrete distribution: $P\{\tau = l_n\} = \eta_n$, for $n = 1, 2, ..., N$, where $l_1 = 0 < l_2 < ... < l_{N-1} < l_N = \infty$.

iii) There are K identical servers. When a server becomes available, the customer at the head of the queue (if there is any) enters the server for service. If an arriving customer finds an available server, the customer enters the server directly upon arrival.

iv) The service time of each customer has an identical phase-type distribution with PH-representation $(\boldsymbol{\beta},\ T)$ of order m_s. We assume that $\boldsymbol{\beta}\mathbf{e} = 1$, i.e., the service time of a customer is always greater than 0. The mean service time is given by $-\boldsymbol{\beta} T^{-1}\mathbf{e}$. Let $\mu_s = 1/(-\boldsymbol{\beta} T^{-1}\mathbf{e})$, which is the service rate.

v) Define $\rho = \lambda/(K\mu_s)$. We assume $\eta_N \rho < 1$ to ensure the stability of the queueing system. Since $\eta_N \lambda$ is the arrival rate of customers with infinite abandonment time, and $K\mu_s$ is the total service rate of the system, $\eta_N \rho < 1$ ensures that all customers

are either served or abandon the system in finite time. Consequently, the system is stable.

4.2 The Age Process and a Multi-Layer $MMFF$ Process

In order to analyze the queueing model by $MMFF$ processes, we introduce a Markov process associated with the age of the customer at the head of the queue. The *age* of a customer is defined as the time elapsed since the customer enters the system. We assume the customers arrive according to an MAP and service times are of phase-type. Then tracking the age of the customer at the head of the queue, phase of the arrival process, and phases of the service processes of individual servers, provides enough information to describe the dynamics of the queueing system. Define

- $a(t)$: the age of the customer at the head of the queue at time t, if the (waiting) queue is not empty; otherwise, $a(t) = 0$ (See Figure 4.1(a)). If $l_n < a(t) < l_{n+1}$, for $n = 1, 2, ..., N-1$, $a(t)$ increases linearly at rate one if there is no service completion, otherwise, $a(t+0) = \max\{0, a(t) - u\}$, where u is the interarrival time between the customer at the head of the queue and the customer who is currently behind it. If $a(t) = l_n$, for $n = 2, 3, ..., N-1$, $a(t)$ continues to increase linearly at rate one with probability $1 - \eta_n/(\eta_n + ... + \eta_N)$; Otherwise, $a(t+0) = \max\{0, l_n - u\}$, where u is the interarrival time between the departing customer (since its waiting time reaches l_n) and the customer who is currently behind it. By this definition, if $a(t) = 0$, there is no customer waiting for service.

- $I_{(a)}(t)$: If $a(t) > 0$, $I_{(a)}(t)$ is the phase of the customer arrival process when the customer now at the head of the queue first entered the queue for service; and if $a(t) = 0$, $I_{(a)}(t)$ is the phase of the customer arrival process at time t (i.e., $I_{(a)}(t) = I_a(t)$.) By this definition, $I_{(a)}(t)$ is piece-wise constant and its value changes only when $a(t)$ drops down.

- $n_i(t)$: the number of servers whose service phase is i at time t, for $i = 1, 2, \ldots, m_s$.

The process $\{(a(t), I_{(a)}(t), n_1(t), \ldots, n_{m_s}(t)),\ t \geq 0\}$ is a continuous time Markov chain because both arrival and service are controlled by an underlying Markov chain and $a(t)$ only depends on the arrival and service processes. According to the total number of working servers, the state space of $(n_1(t), \ldots, n_{m_s}(t))$ can be organized as $\Omega(0) \cup \Omega(1) \cup \ldots \cup \Omega(K)$, where, for $k = 0, 1, 2, \ldots, K$,

$$\Omega(k) = \left\{ \mathbf{n} = (n_1, ..., n_{m_s}) :\ n_i \geq 0, n_i \text{ integer},\ i = 1,\ ...,\ m_s,\ \sum_{i=1}^{m_s} n_i = k \right\}. \qquad (4.2.1)$$

The set $\Omega(k)$ consists of all states such that there are exactly k customers in service (or k working servers), for $k = 0, 1, \ldots, K$. The number of states in $\Omega(k)$ is given by $(k + m_s - 1)!/(k!(m_s - 1)!)$. Then the state space of the Markov process can be written as

$$\left\{ \{0\} \times \{1, ..., m_a\} \times \{\cup_{k=0}^{K} \Omega(k)\} \right\} \cup \{(0, \infty) \times \{1, ..., m_a\} \times \Omega(K)\}. \qquad (4.2.2)$$

Note: We use the $CSFP$ method to track the service process in this chapter. The number of states required by the underlying Markov chain for this approach is $O\left(\begin{pmatrix} K + m_s - 1 \\ m_s - 1 \end{pmatrix}\right)$, which is significant smaller than $O(m_s^K)$, the number of required states by the $TPFS$ method.

Because of the independence between arrival process and service time distribution, we have, if $a(t) > 0$, the phase of the customer arrival process is frozen (i.e., constant) except for down-jump epochs. On the other hand, the phases of the service processes are changing according to rate matrices $Q(K, m_s)$ for no service completion and $Q^-(K, m_s)P^+(K-1, m_s)$ for service completion, and be frozen for down-jump epochs. If $a(t) = 0$, the arrival phases

and service phases are all changing according to rate matrix

$$Q_{bb}^{(1)} = \begin{pmatrix} A_{0,0} & A_{0,1} & & & & \\ A_{1,0} & A_{1,1} & A_{1,2} & & & \\ & \ddots & \ddots & \ddots & & \\ & & A_{K-1,K-2} & A_{K-1,K-1} & A_{K-1,K} & \\ & & & A_{K,K-1} & (D_0 + \eta_1 D_1) \otimes I + I \otimes Q(K, m_s) \end{pmatrix}, \tag{4.2.3}$$

where $A_{0,0} = D_0$, $A_{0,1} = D_1 \otimes P^+(0, m_s)$, $A_{k,k-1} = I \otimes Q^-(k, m_s)$, $A_{k,k} = D_0 \otimes Q^(k, m_s)$ and $A_{k,k+1} = D_1 \otimes P^+(k, m_s)$. Matrices $Q(k, m_s)$, $Q^-(k, m_s)$ and $P^+(k, m_s)$ are defined in [64]. The construction of these matrices is quite complicated. Please refer to [64] for the algorithm details.

Remark: For notational convenience, in the rest of this chapter, we let

$$\begin{aligned} Q_0(K) &= Q(K, m_s); \\ Q_1(K) &= Q^-(K, m_s) P^+(K-1, m_s). \end{aligned} \tag{4.2.4}$$

Next, we define a multi-layer $MMFF$ process $\{(X(t), \phi(t)), t \geq 0\}$ based on the age process introduced above. The basic idea is to change the down jumps of the age process into periods of decreasing fluid, keep the increasing periods of the age process for the periods of increasing fluid, and keep the periods with $a(t) = 0$ for the periods with zero fluid (See (a) and (b) in Figure 4.1). More specifically, we define

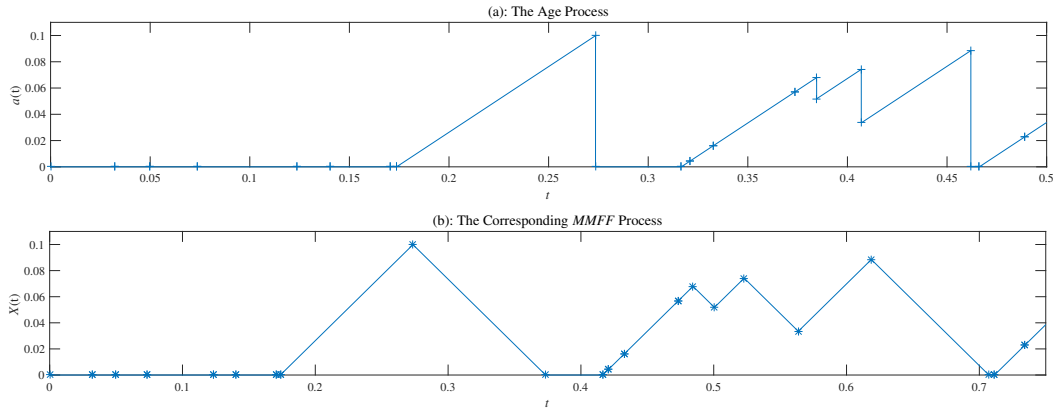

Figure 4.1: Sample paths of the age process and its corresponding $MMFF$ process

1. There are N layers with borders l_n, for $n = 1, 2, ..., N$. Layer 1 is empty (i.e., $\mathcal{S}^{(1)} = \emptyset$).

2. For layer $n \geq 2$, the state space for $\phi(t)$ is:

$$\mathcal{S}_+^{(n)} = \{+\} \times \{1, ..., m_a\} \times \Omega(K), \quad \mathcal{S}_-^{(n)} = \{-\} \times \{1, ..., m_a\} \times \Omega(K), \quad \text{and} \quad \mathcal{S}_0^{(n)} = \emptyset. \tag{4.2.5}$$

   The Q-matrix $Q^{(n)}$ of the underlying Markov chain is:

$$Q^{(n)} = \begin{matrix} \mathcal{S}_+^{(n)} \\ \mathcal{S}_-^{(n)} \end{matrix} \begin{pmatrix} I \otimes Q_0(K) & I \otimes Q_1(K) \\ (\eta_n + ... + \eta_N) D_1 \otimes I & (\eta_1 + ... + \eta_{n-1}) D_1 \otimes I + D_0 \otimes I \end{pmatrix}. \tag{4.2.6}$$

   The fluid flow rates are all 1 or -1, i.e., $C_+^{(n)} = C_-^{(n)} = I$.

3. Within border 1 (i.e., $l_1 = 0$), the transition rates of the underlying Markov chain are given by Equation (4.2.3) for $Q_{bb}^{(1)}$ and

$$Q_{b+}^{(1)} = \begin{pmatrix} 0 \\ (\eta_2 + ... + \eta_N) D_1 \otimes I \end{pmatrix}. \tag{4.2.7}$$

4. The transition probabilities of approaching border 1 are given by (Note: There is no

layer 1, thus can only enter border 1 from above)

$$P_{-b+}^{(1)} = 0; \quad P_{-bb}^{(1)} = (0, ..., 0, I). \tag{4.2.8}$$

When entering from layer 2 to border 1, the underlying process $\phi(t)$ enters the set $\{0\} \times \{1, ..., m_a\} \times \Omega(K)$.

5. All other borders ($n > 1$) have no state. The probabilities of approaching border n, for $2 \leq n \leq N-1$, from below are given by

$$P_{+b-}^{(n)} = \frac{\eta_n}{\eta_n + ... + \eta_N} I; \quad P_{+b+}^{(n)} = \frac{\eta_{n+1} + ... + \eta_N}{\eta_n + \eta_{n+1} + ... + \eta_N} I. \tag{4.2.9}$$

The probabilities of approaching border n, for $2 \leq n \leq N-1$, from above are given by

$$P_{-b-}^{(n)} = I; \quad P_{-b+}^{(n)} = 0. \tag{4.2.10}$$

The joint stationary distribution of the multi-layer $MMFF$ process can be obtained by Algorithm 1.

4.3 Joint Stationary Distribution of the Age Process

Similar to the age process, if $\phi(t) \in \cup_{n=2}^{N} \mathcal{S}_{+}^{(n)}$ (i.e., increase periods), the service process evolves and the state of the arrival process is frozen in the multi-layer $MMFF$ process, and, if $\phi(t) \in \cup_{n=2}^{N} \mathcal{S}_{-}^{(n)}$ (i.e., decrease periods), the states of the service processes are frozen and the arrival process evolves. Therefore, it is easy to see that the age process can be obtained by censoring out states in $\cup_{n=2}^{N} \mathcal{S}_{-}^{(n)}$. Computations can be done by implementing Algorithm 1. However, the state space required for Algorithm 1 is too large to handle large systems if K is big.

The bottleneck of the complexity of this algorithm is the state space of the transition matrix $Q^{(n)}$. By $CSFP$ method, the number of states of servers is $O\left(\begin{pmatrix} K + m_s - 1 \\ m_s - 1 \end{pmatrix}\right)$. Details about the complexity of the algorithm require more exploration. Using certain

special structure of the $MMFF$ process, we can improve Algorithm 1 and reduce the required state space for its implementation as follows:

i) **Border Probabilities**: Since all borders, except Border 1, are empty. We only have to compute $\mathbf{p}^{(1)}$, which satisfies $\mathbf{p}^{(1)}\mathcal{Q}_{\mathbf{p}}^{(1)} = 0$, where

$$\mathcal{Q}_{\mathbf{p}}^{(1)} = Q_{bb}^{(1)} + Q_{b+}^{(1)} T_{+}^{(1)} P_{-bb}^{(1)}, \tag{4.3.1}$$

where $T_{+}^{(1)}$ contains the first passage probabilities from the set above Border 1 (up) to return to Border 1 (from above), which can be computed recursively as follows. We define $T_{+}^{(n)}$ the state transition probabilities that the process goes up leaving Border n and returns to Border n (from above) for the first time (i.e., starting in $\mathcal{S}_{+}^{(n+1)}$ and ending in $\mathcal{S}_{-}^{(n+1)}$). Immediately, we have $T_{+}^{(N-1)} = \Psi^{(N)}$, and, for $n = 2, 3, ..., N-1$,

$$\begin{aligned} T_{+}^{(n-1)} &= \Psi_{+-}^{(l_n - l_{n-1})} + \Lambda_{++}^{(l_n - l_{n-1})}(P_{+b-}^{(n)} + P_{+b+}^{(n)} T_{+}^{(n)}) \\ &\quad \times \left(I - \widehat{\Psi}_{-+}^{(l_n - l_{n-1})}(P_{+b-}^{(n)} + P_{+b+}^{(n)} T_{+}^{(n)})\right)^{-1} \widehat{\Lambda}_{--}^{(l_n - l_{n-1})}. \end{aligned} \tag{4.3.2}$$

ii) **Vector $\mathbf{p}^{(1)}$**: Due to the special structure of $Q_{b+}^{(1)}$ and $P_{-bb}^{(1)}$, we obtain

$$\mathcal{Q}_{\mathbf{p}}^{(1)} = \begin{pmatrix} A_{0,0} & A_{0,1} & & & & \\ A_{1,0} & A_{1,1} & A_{1,2} & & & \\ & \ddots & \ddots & \ddots & & \\ & & A_{K-1,K-2} & A_{K-1,K-1} & A_{K-1,K} \\ & & & A_{K,K-1} & \tilde{A}_{K,K} \end{pmatrix}, \tag{4.3.3}$$

where $\tilde{A}_{K,K} = (D_0 + \eta_1 D_1) \otimes I + I \otimes Q_0(K) + ((\eta_2 + ... + \eta_N) D_1 \otimes I) T_{+}^{(1)}$. We can explore the quasi-birth-and-death (QBD) structure in $\mathcal{Q}_{\mathbf{p}}^{(1)}$ to reduce the state space required for computing $\mathbf{p}^{(1)}$ as follows. Define

$$\begin{aligned} B_1 &= A_{1,0}(-A_{0,0})^{-1}; \\ B_k &= A_{k,k-1}(-A_{k-1,k-1} - B_{k-1}A_{k-2,k-1})^{-1}, \quad \text{for } k = 2, 3, ..., K. \end{aligned} \tag{4.3.4}$$

Define $\mathcal{Q}^{(1)}_{\mathbf{p},K} = D_0 \otimes I + I \otimes Q_0(K) + (D_1 \otimes I)T^{(1)}_+ + B_K A_{K-1,K}$. We also divide $\mathbf{p}^{(1)}$ according to the number of busy servers into $(\mathbf{p}^{(1)}_0, \mathbf{p}^{(1)}_1, ..., \mathbf{p}^{(1)}_K)$. Then $\mathbf{p}^{(1)}_K$ satisfies $\mathbf{p}^{(1)}_K \mathcal{Q}^{(1)}_{\mathbf{p},K} = 0$, and $\mathbf{p}^{(1)}_{k-1} = \mathbf{p}^{(1)}_k B_k$, for $k = K, K-1, ..., 1$. In the computation, we set $\mathbf{p}^{(1)}_K \mathbf{e} = 1$ and normalize the vectors later.

iii) **Coefficients**: Instead of constructing the embedded discrete Markov chain and solving the linear system in Subsection 3.3.2, we can simplify the equations as there is only one sticky border (i.e., Border 1) and some probabilities of approaching borders are 0. Let $\boldsymbol{w}(n) = (\mathbf{w}^{(n+1)}_L, \mathbf{w}^{(n)}_U)$, for $n = 1, ..., N-1$. After we obtain vector $\mathbf{p}^{(1)}$, the coefficients can be obtained directly by solving the following set of linear equations:

$$
\begin{aligned}
&\boldsymbol{w}(1) = \mathbf{p}^{(1)}(Q^{(1)}_{b+}, 0);\\
&\boldsymbol{w}(n) = \boldsymbol{w}(n) \begin{pmatrix} \Psi^{(l_{n+1}-l_n)}_{+-}(P^{(n)}_{-b+}, P^{(n)}_{-b-}) \\ \widehat{\Psi}^{(l_n-l_{n-1})}_{-+}(P^{(n)}_{+b+}, P^{(n)}_{+b-}) \end{pmatrix} + \boldsymbol{w}(n+1) \begin{pmatrix} 0 \\ \widehat{\Lambda}^{(l_{n+1}-l_n)}_{--}(P^{(n)}_{-b+}, P^{(n)}_{-b-}) \end{pmatrix}\\
&\qquad\qquad + \boldsymbol{w}(n-1) \begin{pmatrix} \Lambda^{(l_n-l_{n-1})}_{++}(P^{(n)}_{+b+}, P^{(n)}_{+b-}) \\ 0 \end{pmatrix}, \quad \text{for} \quad n = 2, ..., N-2;\\
&\boldsymbol{w}(N-1) = \boldsymbol{w}(N-1) \begin{pmatrix} \Psi^{(N)}(P^{(N-1)}_{-b+}, P^{(N-1)}_{-b-}) \\ \widehat{\Psi}^{(l_{N-1}-l_{N-2})}_{-+}(P^{(N-1)}_{+b+}, P^{(N-1)}_{+b-}) \end{pmatrix}\\
&\qquad\qquad + \boldsymbol{w}(N-2) \begin{pmatrix} \Lambda^{(l_{N-1}-l_{N-2})}_{++}(P^{(N-1)}_{+b+}, P^{(N-1)}_{+b-}) \\ 0 \end{pmatrix}.
\end{aligned}
\tag{4.3.5}
$$

Denote by $\mathbf{f}(x)$ the joint stationary density function of the age process. Let $\mathbf{f}^{(n)}(x) = \mathbf{f}(x)$, if $l_{n-1} < x < l_n$. By Theorem 3.2 and censoring out $\cup_{n=2}^{N} \mathcal{S}^{(n)}_-$, we obtain the following result.

Theorem 4.1. *([66]) We assume that $\eta_N \rho < 1$ and $(\eta_n + ... + \eta_N)\rho \neq 1$ for $n = 2, 3, ..., N-1$. Then the steady state distribution of the age process exists and its density function is given by*

$$
\begin{aligned}
P\{a(t) = 0\} &= \sum_{k=0}^{K} \hat{\mathbf{p}}^{(1)}_k \mathbf{e};\\
\mathbf{f}^{(n)}(x) &= \mathbf{v}^{(n)}_+ e^{\mathcal{K}^{(n)}(x-l_{n-1})} + \mathbf{v}^{(n)}_- e^{\widehat{\mathcal{K}}^{(n)}(l_n - x)} \widehat{\Psi}^{(n)}, \ \textit{for } l_{n-1} \le x < l_n, \ n = 2, ..., N.
\end{aligned}
\tag{4.3.6}
$$

where $\hat{\mathbf{p}}_k^{(1)} = \mathbf{p}_k^{(1)}/\hat{c}_{norm}$, $\mathbf{v}_+^{(n)} = \mathbf{u}_+^{(n)}/\hat{c}_{norm}$, $\mathbf{v}_-^{(n)} = \mathbf{u}_-^{(n)}/\hat{c}_{norm}$ *and* $\mathbf{v}_+^{(1)} = 0$ *and* $\mathbf{v}_-^{(N)} = 0$. *By the law of total probability, the normalization factor* $\hat{c}_{norm}$ *is given as*

$$\hat{c}_{norm} = \sum_{k=0}^{K} \mathbf{p}_k^{(1)}\mathbf{e} + \sum_{n=2}^{N} \int_{l_{n-1}}^{l_n} \left(\mathbf{u}_+^{(n)} e^{\mathcal{K}^{(n)}(y-l_{n-1})} + \mathbf{u}_-^{(n)} e^{\widehat{\mathcal{K}}^{(n)}(l_n-y)} \widehat{\Psi}^{(n)} \right) \mathbf{e} dy. \tag{4.3.7}$$

Proof. For the existence of the stationary of the age process, we need to show that $\eta_N \rho < 1$ if and only if $\zeta^{(N)} < 0$. To do so, we find $\boldsymbol{\theta}$ satisfying $\boldsymbol{\theta} Q^{(n)} = 0$ and $\boldsymbol{\theta}\mathbf{e} = 1$. We divide $\boldsymbol{\theta}$ into $(\boldsymbol{\theta}_+, \boldsymbol{\theta}_-)$ according to $\mathcal{S}_+^{(n)}$ and $\mathcal{S}_-^{(n)}$. By routine calculations, we obtain $\boldsymbol{\theta}_+ = (\eta_n + ... + \eta_N)(\boldsymbol{\theta}_a D_1) \otimes \tilde{\boldsymbol{\theta}}_s/(\boldsymbol{\theta}_+\mathbf{e} + \boldsymbol{\theta}_-\mathbf{e})$ and $\boldsymbol{\theta}_- = \boldsymbol{\theta}_a \otimes (\tilde{\boldsymbol{\theta}}_s Q_1(K))/(\boldsymbol{\theta}_+\mathbf{e} + \boldsymbol{\theta}_-\mathbf{e})$, where $\tilde{\boldsymbol{\theta}}_s$ satisfies $\tilde{\boldsymbol{\theta}}_s(Q_0(K) + Q_1(K)) = 0$ and $\tilde{\boldsymbol{\theta}}_s\mathbf{e} = 1$. It has been shown in [67] that $\tilde{\boldsymbol{\theta}}_s Q_1(K)\mathbf{e} = K\mu_s$ (i.e., the total service rate). Consequently, we obtain $\mu_n = \boldsymbol{\theta}_+\mathbf{e} - \boldsymbol{\theta}_-\mathbf{e} = ((\eta_n + ... + \eta_N)\lambda - K\mu_s)/(\boldsymbol{\theta}_+\mathbf{e} + \boldsymbol{\theta}_-\mathbf{e})$, which leads to the condition of the existence of the stationary distribution. Also, the relationship shows that $(\eta_n + ... + \eta_N)\lambda - K\mu_s = 0$ if and only if $\zeta^{(n)} = 0$. Thus, all assumptions in Theorem 3.2 are satisfied. The closed form solution of the density function of the age process is obtained from that of the multi-layer $MMFF$ by censoring. □

Remark: For notational convenience, we use notation with time variable t for the stationary counterparts of some quantities (e.g., $a(t)$ for the age in steady state) in this thesis.

Again, evaluation of integrals in the above equation can be done by applying Lemma B.1 in Appendix B. Next, we summarize the modified Algorithm 1 for computing the joint

stationary distribution of the age process as Algorithm 2.

Algorithm 2: The joint stationary distribution of the age process

1. Input Parameters: K, N, $\{l_1 = 0, l_2, ..., l_N = \infty\}$, $\{\eta_1, \eta_2, ..., \eta_N\}$, $\{m_a, D_0, D_1\}$, and $\{m_s, \boldsymbol{\beta}, T\}$;
2. Construct $\{Q_{bb}^{(1)}, Q_0(K), Q_1(K)\}$ by applying the algorithm in [64];
3. Construct transition blocks for the multi-layer $MMFF$ process:
   - 3.1 Borders: $\{l_0 = -\infty, l_1 = 0, ..., l_N = \infty\}$;
   - 3.2 Construct $\{Q^{(n)}, C_+^{(n)}, C_-^{(n)}, n = 1, 2, ..., N\}$ using Equation (4.2.6); (Note: $C_+^{(n)}$ and $C_-^{(n)}$ are not necessary since they are identity matrices);
   - 3.3 Construct $\{Q_{bb}^{(1)}, Q_{b+}^{(1)}, Q_{b-}^{(1)}\}$ using Equations (4.2.3) and (4.2.7);
   - 3.4 Construct $\{P_{+b+}^{(n)}, P_{+bb}^{(n)}, P_{+b-}^{(n)}, P_{-b+}^{(n)}, P_{-bb}^{(n)}, P_{-b-}^{(n)}, n = 1, 2, ..., N-1\}$ using Equations (4.2.8), (4.2.9) and (4.2.10);
4. Similar to Steps 2 and 3 in Algorithm 1, compute $\{\Psi^{(n)}, \mathcal{K}^{(n)}, \mathcal{U}^{(n)}, \widehat{\Psi}^{(n)}, \widehat{\mathcal{K}}^{(n)}, \widehat{\mathcal{U}}^{(n)}\}$ for $\{Q^{(n)}, C_+^{(n)}, C_-^{(n)}\}$; Compute $\{\Psi_{+-}^{(l_n - l_{n-1})}, \widehat{\Psi}_{-+}^{(l_n - l_{n-1})}, \Lambda_{++}^{(l_n - l_{n-1})}, \widehat{\Lambda}_{--}^{(l_n - l_{n-1})}\}$ for $\{Q^{(n)}, C_+^{(n)}, C_-^{(n)}\}$, for $n = 1, 2, ..., N-1$;
5. Compute $T_+^{(1)}$ using Equation (4.3.2); Construct $\mathcal{Q}_{\mathbf{p},K}^{(1)}$ using (4.3.4); and solve $\mathbf{p}_K^{(1)} \mathcal{Q}_{\mathbf{p},K}^{(1)} = 0$ and $\mathbf{p}_K^{(1)} \mathbf{e} = 1$, and Compute $\mathbf{p}^{(1)}$;
6. Compute $\{\boldsymbol{w}(n),\ n = 1, 2, ..., N-1\}$ by Equation (4.3.5); and $\{\mathbf{u}_+^{(n)}, \mathbf{u}_-^{(n)},\ n = 1, 2, ..., N\}$ by using Algorithm 1;
7. Compute $\hat{c}_{norm}$ by using Equation (4.3.7), and use $\hat{c}_{norm}$ to get $\{\hat{\mathbf{p}}_k^{(1)}, n = 0, 1, ..., K\}$ and $\{\mathbf{v}_+^{(n)}, \mathbf{v}_-^{(n)},\ n = 1, 2, ..., N\}$;
8. Use the $\{\hat{\mathbf{p}}_k^{(1)}, n = 0, 1, ..., K\}$ and $\{\mathbf{v}_+^{(n)}, \mathbf{v}_-^{(n)},\ n = 1, 2, ..., N\}$ and Equation (4.3.6) to compute the density function of the age process.

The summarized computation process can be further simplified. For example, there

is no need to do Step 3 since all subsequent computations can be done by directly using matrices constructed in Step 2.

4.4 Queueing Quantities

Based on the joint stationary distribution of the age process, we find three sets of queueing quantities: i) Customer abandonment/loss probabilities; ii) Waiting times; and iii) Queue lengths. We assume that conditions stated in Theorem 4.1 hold throughout this section.

4.4.1 Abandonment Probabilities

Proposition 4.1. *([66]) The probability that a customer will eventually receive service is given by*

$$p_S = \frac{1}{\lambda}\sum_{k=0}^{K-1}\hat{\mathbf{p}}_k^{(1)}(D_1\otimes I)\mathbf{e} + \frac{1}{\lambda}\sum_{n=2}^{N}\left(\mathbf{v}_+^{(n)}\mathcal{L}_{l_{n-1},l_n}^{\mathcal{K}^{(n)}} + \mathbf{v}_-^{(n)}\widetilde{\mathcal{L}}_{l_{n-1},l_n}^{\widehat{\mathcal{K}}^{(n)}}\widehat{\Psi}^{(n)}\right)(I\otimes Q_1(K))\mathbf{e}, \quad (4.4.1)$$

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 Lemma B.1. Then the customer abandonment probability is $p_L = 1 - p_S$. We decompose p_L into two parts: i) loss probability $p_{L,1}$ of customers at the head of the waiting queue; and ii) loss probability $p_{L,>1}$ of customers before reaching the head of the waiting queue. Then we obtain $p_{L,>1} = p_L - p_{L,1}$, and

$$p_{L,1} = \frac{\hat{\mathbf{p}}_k^{(1)}((\eta_1 D_1)\otimes I)\mathbf{e}}{\lambda} + \frac{1}{\lambda}\sum_{n=2}^{N-1}\left(\mathbf{v}_+^{(n)}e^{\mathcal{K}^{(n)}(l_n-l_{n-1})}\mathbf{e} + \mathbf{v}_-^{(n)}\widehat{\Psi}^{(n)}\mathbf{e}\right)\frac{\eta_n}{\sum_{m=n}^{N}\eta_m}. \quad (4.4.2)$$

Proof. By definitions, we have

$$p_S = \frac{1}{\lambda}\left(\sum_{k=0}^{K-1}\hat{\mathbf{p}}_k^{(1)}(D_1\otimes I)\mathbf{e} + \int_0^{\infty}\mathbf{f}(x)(I\otimes Q_1(K))\mathbf{e}\mathrm{d}x\right). \quad (4.4.3)$$

First note that the numerator in Equation (4.4.3) is the sum of transition rates that a customer enters a server for service, and the denominator in Equation (4.4.3) is the arrival

rate. Then the ratio is the percentage of customers who received service, which is also the probability that a customer will eventually receive service. The desired expression is obtained by combining Equation (4.4.3) and Lemma B.1.

The probability that a customer sees exactly K customers in service and no waiting queue, and abandons the queue is $\frac{\hat{\mathbf{p}}_K^{(1)}((\eta_1 D_1)\otimes I)\mathbf{e}}{\lambda}$. For a customer at the head of the queue to abandon the queue, its age must reach l_n for some $n = 2, 3, ..., N-1$. If its age reaches l_n, its age must be greater than l_{n-1}, which occurs with probability $\eta_n + ... + \eta_N$. Then the probability that it abandons the queue is $\eta_n/(\eta_n + ... + \eta_N)$. Combining with the transition rate for the age to reach l_n, which is $\mathbf{f}(l_n)\mathbf{e}$, we obtain

$$p_{L,1} = \frac{\hat{\mathbf{p}}_k^{(1)}((\eta_1 D_1)\otimes I)\mathbf{e}}{\lambda} + \frac{1}{\lambda}\sum_{n=2}^{N-1}\mathbf{f}(l_n)\mathbf{e}\frac{\eta_n}{\sum_{m=n}^{N}\eta_m}, \tag{4.4.4}$$

which leads to the desired result. □

4.4.2 Waiting Times

Proposition 4.2. *([66]) The distribution of waiting time W_S of customers received service is*

$$\begin{aligned} P\{W_S = 0\} &= \frac{1}{p_S\lambda}\sum_{k=0}^{K-1}\hat{\mathbf{p}}_k^{(1)}(D_1\otimes I)\mathbf{e}; \\ \frac{\mathrm{d}P\{W_S < x\}}{\mathrm{d}x} &= \frac{1}{p_S\lambda}\left(\mathbf{v}_+^{(n)}e^{\mathcal{K}^{(n)}(x-l_{n-1})} + \mathbf{v}_-^{(n)}e^{\widehat{\mathcal{K}}^{(n)}(l_n-x)}\widehat{\Psi}^{(n)}\right)(I\otimes Q_1(K))\mathbf{e}, \\ &\qquad\qquad \textit{for}\ \ l_{n-1}\le x < l_n,\ \ n = 2, 3, ..., N. \end{aligned} \tag{4.4.5}$$

The distribution of waiting time $W_{L,1}$ of customers lost at the head of the waiting queue is given by

$$\begin{aligned} P\{W_{L,1} = l_n\} &= \frac{\hat{\mathbf{p}}_k^{(1)}((\eta_1 D_1)\otimes I)\mathbf{e}}{p_{L,1}\lambda}, \qquad \textit{for } n = 1; \\ P\{W_{L,1} = l_n\} &= \left(\frac{\eta_n}{\eta_n + ... + \eta_N}\right)\frac{\mathbf{f}^{(n)}(l_n)\mathbf{e}}{p_{L,1}\lambda}, \quad \textit{for } n = 2, 3, ..., N-1. \end{aligned} \tag{4.4.6}$$

The abandonment time $W_{L,>1}$ of a customer that abandons the queue before reaching the head of the queue, we have, for $k = 1, 2, 3, ..., N-1$,

$$P\{W_{L,>1} = l_k\} = \left(\sum_{n=k+1}^{N} \left(\mathbf{v}_+^{(n)} \mathcal{L}_{l_{n-1},l_n}^{\mathcal{K}^{(n)}} \Psi^{(n)} + \mathbf{v}_-^{(n)} \widetilde{\mathcal{L}}_{l_{n-1},l_n}^{\widehat{\mathcal{K}}^{(n)}} \right) (D_1 \otimes I)\mathbf{e} \right) \frac{\eta_k}{p_{L,>1}\lambda}. \tag{4.4.7}$$

Proof. First note that $W_S = 0$ occurs if a server is available when a customer arrives, which leads to the expression for $P\{W_S = 0\}$. When the age is $x > 0$ and there is an service completion, the waiting time of the customer at the head of the queue is exactly x, so the rate that the waiting time is x of customer received service is given by $\mathbf{f}(x)(I \otimes Q_1(K))\mathbf{e}$, then the rate ration given below gives the probability density function,

$$\frac{\mathrm{d}P\{W_S < x\}}{\mathrm{d}x} = \frac{1}{p_S\lambda}\mathbf{f}(x)(I \otimes Q_1(K))\mathbf{e}, \quad \text{for } x > 0, \tag{4.4.8}$$

which leads to the desired result.

For $W_{L,1}$, it is clear that $W_{L,1} = l_n$ if $a(t)$ reaches l_n from below and an abandonment occurs. The probability for $W_{L,1}$ to reach l_n is $\mathbf{f}^{(n)}(l_n)\mathbf{e}/(p_{L,1}\lambda)$. The probability for the abandonment to occur is $\eta_n/(\eta_n + ... + \eta_N)$. Then expression (4.4.6) can be obtained easily.

We use the joint stationary distribution of the multi-layer $MMFF$ process to find the distribution of $W_{L,>1}$. When the multi-layer $MMFF$ process is in $S_-^{(n)}$ and there is an arrival, the arriving customer will abandon the queue in the future with probability $\eta_2 + ... + \eta_{n-1}$ if $l_{n-1} < x < l_n$. Since customer arrivals take place only when the fluid level of the $MMFF$ process is decreasing, we censor out the periods of time in which the fluid level is increasing. Using the censored process, we obtain, for $k = 2, 3, ..., N-1$,

$$P\{W_{L,>1} = l_k\} = \frac{c_{norm}}{\hat{\hat{c}}_{norm} p_{L,>1}\lambda} \left(\sum_{n=k}^{N-1} \int_{l_n}^{l_{n+1}} \boldsymbol{\pi}_-^{(n+1)}(x)\mathrm{d}x(D_1 \otimes I) \right) \mathbf{e}\eta_k, \tag{4.4.9}$$

where

$$\hat{\hat{c}}_{norm} = \sum_{k=0}^{K} \mathbf{p}_k^{(1)}\mathbf{e} + \sum_{n=2}^{N} \int_{l_{n-1}}^{l_n} \left(\mathbf{u}_+^{(n)} e^{\mathcal{K}^{(n)}(y-l_{n-1})}\Psi^{(n)} + \mathbf{u}_-^{(n)} e^{\widehat{\mathcal{K}}^{(n)}(l_n-y)} \right) \mathbf{e}\mathrm{d}y. \tag{4.4.10}$$

In the multi-layer $MMFF$ process, the fluid level increases and decreases both at rate 1. If the process is ergodic, probabilities that the process is increasing or decreasing at an arbitrary time are equal. Thus, we must have $\hat{c}_{norm} = \hat{\hat{c}}_{norm}$, which leads to the desired result in Equation (4.4.7). □

According to the law of total probability, we must have $P\{W_S < \infty\} = 1$ and $\sum_{n=2}^{N-1} P\{W_{L,1} = l_n\} = 1$, which can be used to check computation accuracy. The law of total probability $\sum_{n=2}^{N-1} P\{W_{L,>1} = l_n\} = 1$ can also be used to check computation accuracy. The mean waiting time $\mathbb{E}[W_S]$ can be calculated by:

$$\mathbb{E}[W_S] = \frac{1}{p_S \lambda} \sum_{n=2}^{N} \left(\mathbf{v}_+^{(n)} \mathcal{M}_{l_{n-1},l_n}^{\mathcal{K}^{(n)}} + \mathbf{v}_-^{(n)} \widetilde{\mathcal{M}}_{l_{n-1},l_n}^{\widehat{\mathcal{K}}^{(n)}} \widehat{\Psi}^{(n)} \right) (I \otimes Q_1(K))\mathbf{e}, \tag{4.4.11}$$

where $\mathcal{M}_{l_{n-1},l_n}^{\mathcal{K}^{(n)}}$ and $\widetilde{\mathcal{M}}_{l_{n-1},l_n}^{\widehat{\mathcal{K}}^{(n)}}$ are defined in Lemma B.1. The distribution of the waiting time W of an arbitrary customer can be found from that of W_S, $W_{L,1}$, and $W_{L,>1}$. The mean waiting time can be found by

$$\mathbb{E}[W] = p_S \mathbb{E}[W_S] + p_{L,1} \mathbb{E}[W_{L,1}] + p_{L,>1} \mathbb{E}[W_{L,>1}]. \tag{4.4.12}$$

4.4.3 Queue Lengths

Let $q_S(t)$ be the number of customers in service (or busy servers) and $q_W(t)$ the waiting queue length at an arbitrary time t. The distribution of $q_S(t)$ can be found directly from the border probability vector $\hat{\mathbf{p}}^{(1)}$. The z-transform of $q_W(t)$ can be derived based on the joint distribution of the age process. If $a(t) = x$ at an arbitrary time t, the waiting queue length consists of the customer at the head of the queue and all customers arrived after that customer (i.e., in the period $(t-x, t)$) who have not abandoned the queue yet. To identify who are still waiting in queue and who have abandoned the queue, we divide the interval $(t-x, t)$ into $(t-l_2, t)$, $(t-l_3, t-l_2)$, ..., $(t-x, t-l_{n-1})$, if $l_{n-1} < x < l_n$ (See Figure 4.2). For customers who arrived in $(t-l_2, t)$, they abandon the queue before t with probability η_1 and are still in the queue at time t with probability $1-\eta_1$. The conditional probability generating function of the number of such customers is $e^{(D_0+(\eta_1+(1-\eta_1)z)D_1)l_2}$

(see Theorem 2.3.1 in [62] or Lemma B.2 in Appendix B). For customers who arrived in $(t-l_3, t-l_2)$, they abandon the queue before t with probability η_2 and are still in the queue at time t with probability $1-\eta_2$. The conditional probability generating function is given by $e^{(D_0+(\eta_1+\eta_2+(1-\eta_1-\eta_2)z)D_1)(l_3-l_2)}$. In general, for customers arrived in $(t-l_m, t-l_{m-1})$, they abandon the queue before t with probability $1-\hat{\eta}_m$ and are still in the queue at time t with probability $\hat{\eta}_m$, where $\hat{\eta}_m = \eta_m + \eta_{m+1} + ... + \eta_N$. The conditional probability generating function is given by $e^{(D_0+(1-\hat{\eta}_m+\hat{\eta}_m z)D_1)(l_m-l_{m-1})}$.

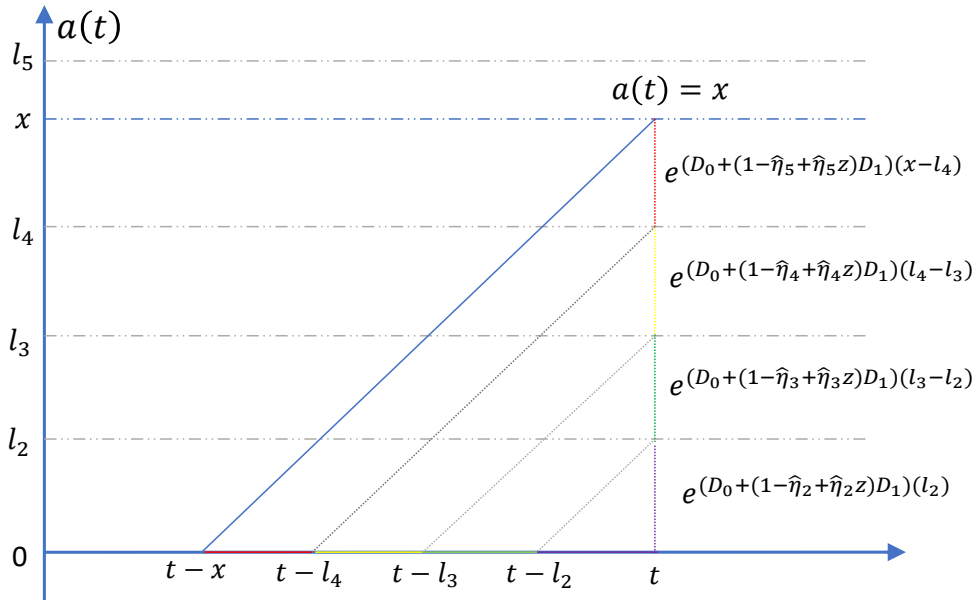

Figure 4.2: The conditional probability generating function of the number of customers in each interval

Denote by $P^*(\eta, z, x) = e^{(D_0+(1-\eta+\eta z)D_1)x} \otimes I$.

Lemma 4.1. *Conditioning on $a(t)$ at an arbitrary time t, for $z \geq 0$, the probability generating function of $q_W(t)$ can be found as follows,*

$$\mathbb{E}[z^{q_W(t)}] = \hat{\mathbf{p}}^{(1)}\mathbf{e} + z\sum_{n=2}^{N}\int_{l_{n-1}}^{l_n} \mathbf{f}^{(n)}(x)P^*(\hat{\eta}_n, z, x-l_{n-1})\prod_{m=n-1}^{2} P^*(\hat{\eta}_m, z, b_m)\mathrm{d}x\mathbf{e}. \quad (4.4.13)$$

(Remark: $b_m = l_m - l_{m-1}$, for $m = 2, 3, ..., N$.)

Proof. By the definition of probability generating function, we have

$$\mathbb{E}[z^{q_W(t)}] = \sum_{i=0}^{\infty} P\{q_W(t) = i\} z^i. \tag{4.4.14}$$

If $i = 0$, we have

$$P\{q_W(t) = 0\} z^0 = \hat{\mathbf{p}}^{(1)} \mathbf{e}. \tag{4.4.15}$$

If $i \geq 1$, there must be a customer at the head of the queue and $a(t)$ has to be positive, thus the probability generating function of the number of customers (always 1) at the head of the queue at time t is z. Conditioning on $a(t) = x$ at an arbitrary time t and $x \in (l_{n-1}, l_n)$, the probability generating function of the number of customers behind the head of the queue at time t is

$$P^*(\hat{\eta}_n, z, x - l_{n-1}) \times P^*(\hat{\eta}_{n-1}, z, l_{n-1} - l_{n-2}) \times \cdots \times P^*(\hat{\eta}_2, z, l_2), \tag{4.4.16}$$

thus we have

$$\begin{aligned}
&\sum_{i=1}^{\infty} P\{q_W(t) = i\} z^i \\
&= z \sum_{n=2}^{N} \int_{l_{n-1}}^{l_n} \mathbf{f}^{(n)}(x) P^*(\hat{\eta}_n, z, x - l_{n-1}) \times P^*(\hat{\eta}_{n-1}, z, l_{n-1} - l_{n-2}) \times \cdots \times P^*(\hat{\eta}_2, z, l_2) \mathrm{d}x \mathbf{e} \\
&= z \sum_{n=2}^{N} \int_{l_{n-1}}^{l_n} \mathbf{f}^{(n)}(x) P^*(\hat{\eta}_n, z, x - l_{n-1}) \prod_{m=n-1}^{2} P^*(\hat{\eta}_m, z, b_m) \mathrm{d}x \mathbf{e}.
\end{aligned} \tag{4.4.17}$$

The sum of Equation (4.4.15) and Equation (4.4.17) leads to the desired result. □

By Theorem 2.3.2 in [62] or Lemma B.2 in Appendix B, we have the following result.

Proposition 4.3. *([66]) The distribution of $q_S(t)$ is given by*

$$P\{q_S(t)=k\}=\begin{cases}\hat{\mathbf{p}}_k^{(1)}\mathbf{e}, & \text{if } k=0,1,...,K-1;\\ 1-\sum_{k=0}^{K-1}\hat{\mathbf{p}}_k^{(1)}\mathbf{e}, & \text{if } k=K.\end{cases} \tag{4.4.18}$$

The mean waiting queue length is given by

$$\begin{aligned}&\mathbb{E}[q_W(t)]=1-\hat{\mathbf{p}}^{(1)}\mathbf{e}\\&+\sum_{n=2}^{N}\sum_{m=2}^{n-1}\int_{l_{n-1}}^{l_n}\mathbf{f}^{(n)}(x)\left(e^{D(x-l_m)}\otimes I\right)\mathrm{d}x\left(\hat{\eta}_m\lambda b_m I+(e^{Db_m}-I)(D-\mathbf{e}\boldsymbol{\theta}_a)^{-1}\hat{\eta}_m D_1\right)\mathbf{e}\otimes\mathbf{e}\\&+\sum_{n=2}^{N}\int_{l_{n-1}}^{l_n}\mathbf{f}^{(n)}(x)\left(\hat{\eta}_n\lambda(x-l_{n-1})I+(e^{D(x-l_{n-1})}-I)(D-\mathbf{e}\boldsymbol{\theta}_a)^{-1}\hat{\eta}_n D_1\right)\mathbf{e}\otimes\mathbf{e}\mathrm{d}x.\end{aligned} \tag{4.4.19}$$

To calculate the mean queue length, we need to evaluate the integral (Lemma B.1 in Appendix B), for $2\le m<n\le N$,

$$\int_{l_{n-1}}^{l_n}\left(\mathbf{v}_+^{(n)}e^{\mathcal{K}^{(n)}(x-l_{n-1})}+\mathbf{v}_-^{(n)}e^{\widehat{\mathcal{K}}^{(n)}(l_n-x)}\widehat{\Psi}^{(n)}\right)\left(e^{D(x-l_{n-1})}\otimes I\right)\mathrm{d}x. \tag{4.4.20}$$

Combining Proposition 4.3 and Lemma B.1, we obtain

$$
\begin{aligned}
\mathbb{E}[q_W(t)] &= 1 - \hat{\mathbf{p}}^{(1)}\mathbf{e} \\
&+ \sum_{n=2}^{N} \left(\mathbf{v}_+^{(n)} \mathcal{L}_{l_{n-1},l_n}^{(\mathcal{K}^{(n)},D)} + \mathbf{v}_-^{(n)} \widetilde{\mathcal{L}}_{l_{n-1},l_n}^{(\widehat{\mathcal{K}}^{(n)},D)} \right) \\
&\quad \times \left(\sum_{m=2}^{n-1} e^{D(l_{n-1}-l_m)} \left(\hat{\eta}_m \lambda b_m I + (e^{Db_m} - I)(D - \mathbf{e}\boldsymbol{\theta}_a)^{-1} \hat{\eta}_m D_1 \right) \otimes I \right) \mathbf{e} \\
&+ \sum_{n=2}^{N} \hat{\eta}_n \lambda \left(\mathbf{v}_+^{(n)} \left(\mathcal{M}_{l_{n-1},l_n}^{\mathcal{K}^{(n)}} - l_{n-1} \mathcal{L}_{l_{n-1},l_n}^{\mathcal{K}^{(n)}} \right) + \mathbf{v}_-^{(n)} \left(\widetilde{\mathcal{M}}_{l_{n-1},l_n}^{\widehat{\mathcal{K}}^{(n)}} - l_{n-1} \widetilde{\mathcal{L}}_{l_{n-1},l_n}^{\widehat{\mathcal{K}}^{(n)}} \right) \widehat{\Psi}^{(n)} \right) \mathbf{e} \\
&+ \sum_{n=2}^{N} \left(\mathbf{v}_+^{(n)} \left(\mathcal{L}_{l_{n-1},l_n}^{(\mathcal{K}^{(n)},D)} - \mathcal{L}_{l_{n-1},l_n}^{\mathcal{K}^{(n)}} \right) + \mathbf{v}_-^{(n)} \left(\widetilde{\mathcal{L}}_{l_{n-1},l_n}^{(\widehat{\mathcal{K}}^{(n)},D)} - \widetilde{\mathcal{L}}_{l_{n-1},l_n}^{\widehat{\mathcal{K}}^{(n)}} \widehat{\Psi}^{(n)} \right) \right) \\
&\quad \times \left((D - \mathbf{e}\boldsymbol{\theta}_a)^{-1} \hat{\eta}_n D_1 \otimes I \right) \mathbf{e}.
\end{aligned}
\tag{4.4.21}
$$

Note that Lemma B.1 is used in the above expression.

Let $q_{tot}(t)$ be the total number of customers in the queueing system at an arbitrary time t. Then the probability generating function and the mean of $q_{tot}(t)$ can be found as

$$
\begin{aligned}
\mathbb{E}[z^{q_{tot}(t)}] &= \sum_{k=0}^{K} z^k \hat{\mathbf{p}}_k^{(1)} \mathbf{e} + z^K \mathbb{E}[z^{q_W(t)}], \\
\mathbb{E}[q_{tot}(t)] &= \sum_{k=0}^{K} k \hat{\mathbf{p}}_k^{(1)} \mathbf{e} + K(1 - \hat{\mathbf{p}}^{(1)}\mathbf{e}) + \mathbb{E}[q_W(t)].
\end{aligned}
\tag{4.4.22}
$$

The queueing quantities are connected to each other by the well-known Little's law: i) $\mathbb{E}[q_W(t)] = \lambda \mathbb{E}[W]$ for the number of waiting customers and the actual waiting times of customers; ii) $\mathbb{E}[q_S(t)] = \lambda p_S \boldsymbol{\beta}(-T)^{-1}\mathbf{e}$ for the number of customers in service and service times; and iii) $\mathbb{E}[q_{tot}(t)] = \lambda \mathbb{E}[W] + \lambda p_S \boldsymbol{\beta}(-T)^{-1}\mathbf{e}$ for the total of number of customers in the queueing system and the sojourn times of customers. The relationships can be used for checking computation accuracy.

4.4.4 Summary of Queueing Quantities

We summarize all important queueing quantities in Table 4.1 to help readers quickly find the meaning and equations of these quantities.

Notations	Quantities	Equations
$\mathbf{f}^{(n)}(x)$	Density of the age process	(4.3.6)
P_S, P_L, $P_{L,1}$, $P_{L,>1}$	Abandonment probabilities	(4.4.1), (4.4.2)
W_S, $\mathbb{E}[W_S]$	Waiting time of served customers	(4.4.5), (4.4.11)
$W_{L,1}$, $W_{L,>1}$	Waiting time of abandoned customers	(4.4.6), (4.4.7)
$\mathbb{E}[W]$	Mean waiting time	(4.4.12)
$q_S(t)$, $\mathbb{E}[q_W(t)]$, $\mathbb{E}[q_{tot}(t)]$	Queue lengths	(4.4.18), (4.4.21), (4.4.22)

Table 4.1: Summary of queueing quantities in Chapter 4

4.5 Numerical Examples

Example 4.1. We consider an $MAP/PH/K{+}GI$ queue with $K = 3$, $N = 6$, $(l_1, l_2, l_3, l_4, l_5, l_6) = (0, 1, 2, 3, 4, \infty)$, $\boldsymbol{\eta} = (0, 0.1, 0.3, 0.3, 0.2, 0.1)$,

$$D_0 = \begin{pmatrix} -14 & 0 \\ 4.5 & -5.5 \end{pmatrix}, \quad D_1 = \begin{pmatrix} 12 & 2 \\ 0.5 & 0.5 \end{pmatrix}; \quad \boldsymbol{\beta} = (0.5, 0.5), \quad T = \begin{pmatrix} -5.5 & 4.5 \\ 5 & -5.8 \end{pmatrix}. \tag{4.5.1}$$

Applying Algorithm 2, a number of queueing quantities can be obtained. First, we plot the stationary density functions of the age of the customer at the head of the queue and the waiting time of an arbitrary served customer in Figure 4.3. It seems that most of the customers have to wait in the queue for service. It is interesting to see that i) The density of the waiting time of the served customer is closed to the density of age of the customer at the head of the queue; ii) The density of waiting time concentrated around $l_4 = 3$ and $l_5 = 4$ even though this is for served customers. Second, we present the (conditional) distributions of the waiting times of customers who abandoned the queue in Table 4.2. While the possibility of customers abandoning the queue varies significantly before they reach the head of the queue. Lastly, we summarize other queueing quantities in Table 4.3.

Again, the mean age of the customer at the head of the queue is closed to the mean waiting time of the arbitrarily served customer. The mean number of working servers is 3, which means they are always busy. The mean total queue length is 30.68558, which is far greater than 3. Next example, we increase the number of servers K to see the changes of these queueing quantities.

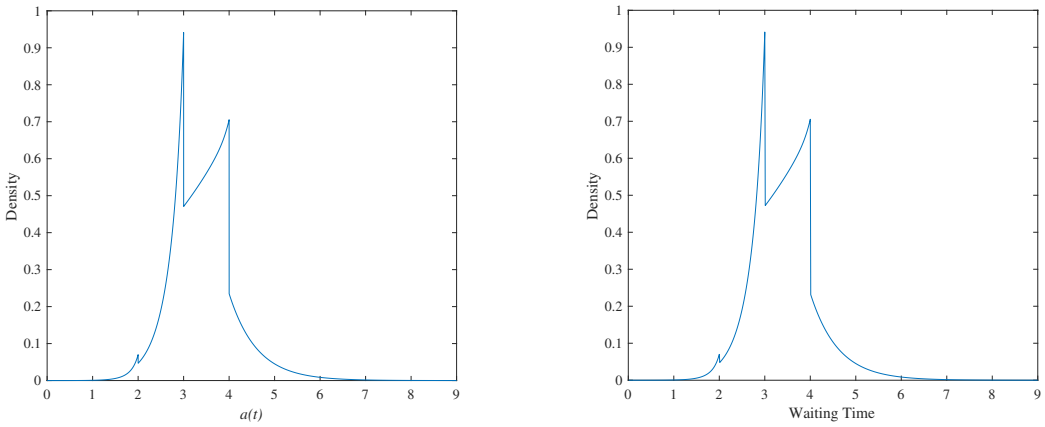

Figure 4.3: The stationary density functions of $a(t)$ and W_S for Example 4.1

	l_1	l_2	l_3	l_4	l_5	l_6
$P\{W_{L,1} = l_n\}$	0	0.0	0.025	0.488	0.487	0
$P\{W_{L,>1} = l_n\}$	0	0.157	0.461	0.337	0.045	0
$P\{W_L = l_n\}$	0	0.136	0.406	0.356	0.102	0

Table 4.2: Conditional distributions of waiting times of customers abandoned the queue

$\mathbb{E}[a(t)]$	p_S	p_L	$p_{L,1}$	$p_{L,>1}$	$p_{q,0}$	$\mathbb{E}[W_S]$
3.4375	0.2636	0.7364	0.0938	0.6426	0.0	3.4376
$\mathbb{E}[W_{L,1}]$	$\mathbb{E}[W_{L,>1}]$	$\mathbb{E}[W_L]$	$\mathbb{E}[W]$	$\mathbb{E}[q_S]$	$\mathbb{E}[q_W]$	$\mathbb{E}[q_{tot}]$
3.4633	2.2731	2.4246	2.6917	3.0000	27.6858	30.6858

Table 4.3: Summary of queueing quantities for Example 4.1

Example 4.2. (Example 4.1 continued) For Example 4.1, we change the number of servers

from $K = 2$ to $K = 50$, and compute queueing quantities for those queueing systems. The results are divided into three groups $\{p_L, p_{L,1}, p_{L,>1}\}$, $\{\mathbb{E}[W_S], \mathbb{E}[W_{L,1}], \mathbb{E}[W_{L,>1}], \mathbb{E}[W_{tot}]\}$, and $\{\mathbb{E}[q_S], \mathbb{E}[q_W], \mathbb{E}[q_{tot}]\}$. The results are plotted in Figure 4.4.

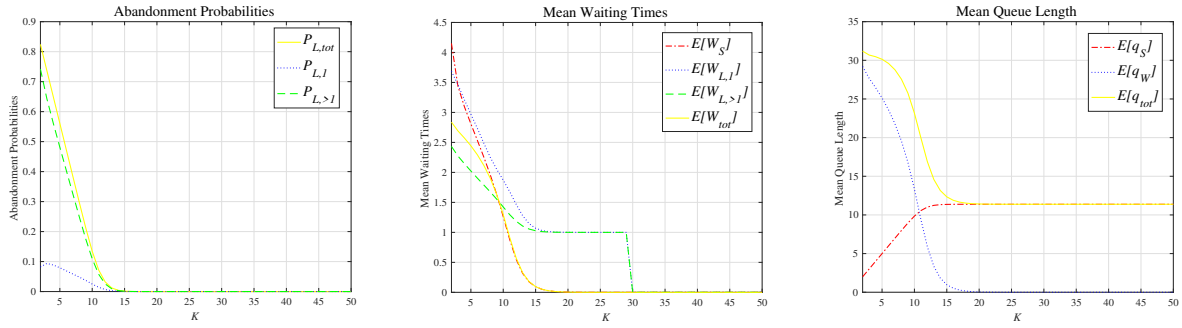

Figure 4.4: Summary of queueing quantities for Example 4.2

From Figure 4.4, it is interesting to see that i) The abandonment probability $p_{L,1}$ is not monotone as K increases; ii) The mean waiting times are all decreasing (which is intuitive); and iii) The abandonment probabilities and mean waiting times go to 0 when the number of servers is large.

We also plot the density function of the waiting time of served customers for W_S for $K = 2$, $K = 6$, $K = 10$, $K = 14$,$K = 18$,and $K = 22$ in Figure 4.5. It is interesting to see how the waiting time distribution shifts as K changes. One thing particularly interesting is the impact of the abandonment epochs on the waiting time distribution, which becomes less significant as K increases. Intuitively, it is due to fewer customers are forced to make abandonment decisions as more servers become available.

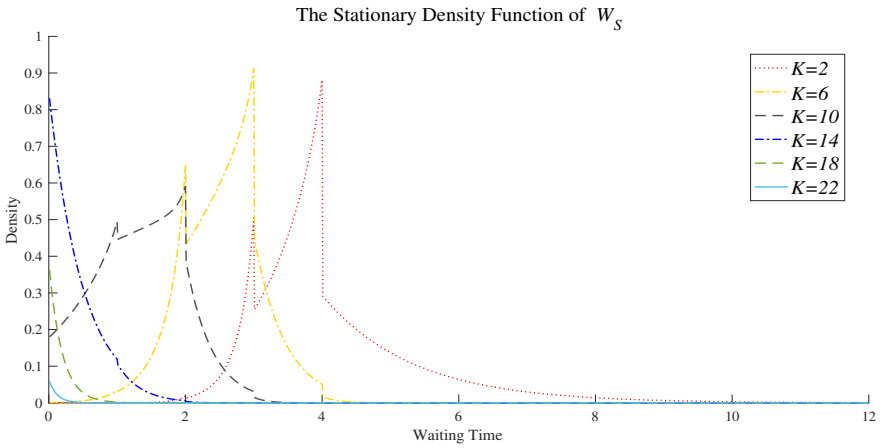

Figure 4.5: The stationary density functions of W_S for $K = 2$, $K = 6$, $K = 10$, $K = 14$, $K = 18$, and $K = 22$

Example 4.3. In this example, we consider a queueing system with a bursty arrival process and service times with a big variation. We assume $N = 5$, $l_1 = 0$, $l_2 = 1$, $l_3 = 2$, $l_4 = 3$, $l_5 = \infty$, $\boldsymbol{\eta} = (0, 0.2, 0.3, 0.4, 0.1)$,

$$m_a = 4, \quad D_0 = \begin{pmatrix} -15 & 0 & 2 & 2 \\ 20 & -45 & 2 & 2 \\ 1 & 2 & -25 & 5 \\ 1 & 0 & 2 & -15 \end{pmatrix}, \quad D_1 = \begin{pmatrix} 5 & 5 & 1 & 0 \\ 10 & 5 & 1 & 5 \\ 1 & 6 & 5 & 5 \\ 5 & 1 & 1 & 5 \end{pmatrix}; \tag{4.5.2}$$

$$m_s = 3, \quad \boldsymbol{\beta} = (0.1, 0.0, 0.9), \quad T = \begin{pmatrix} -17 & 0 & 10 \\ 0 & -2 & 0 \\ 0 & 2 & -2 \end{pmatrix}.$$

This example is special since the arrival process is bursty and the service times have a special distribution as shown in Figure 4.6.

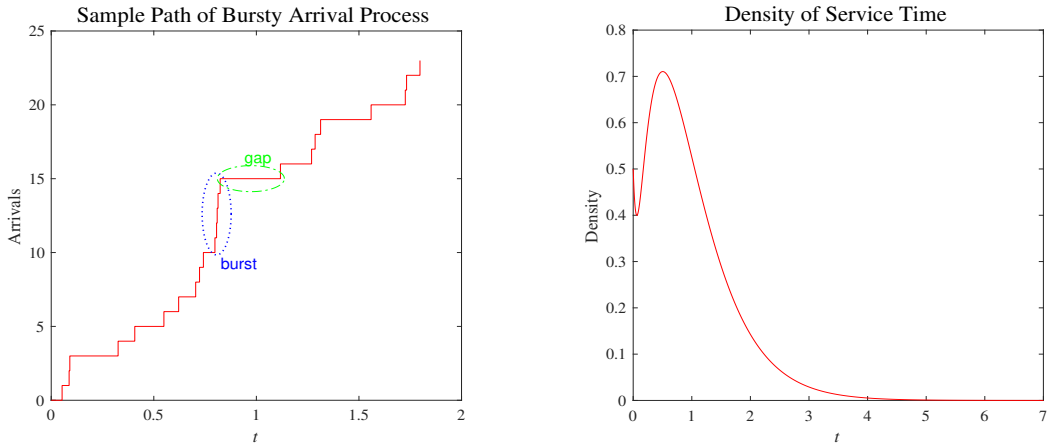

Figure 4.6: Burstiness of the arrival process and density function of the service times

Let K go from 2 to 16. We compute queueing quantities for Example 4.3. Results related to customer abandonment, waiting times and queue lengths are plotted in Figure 4.7.

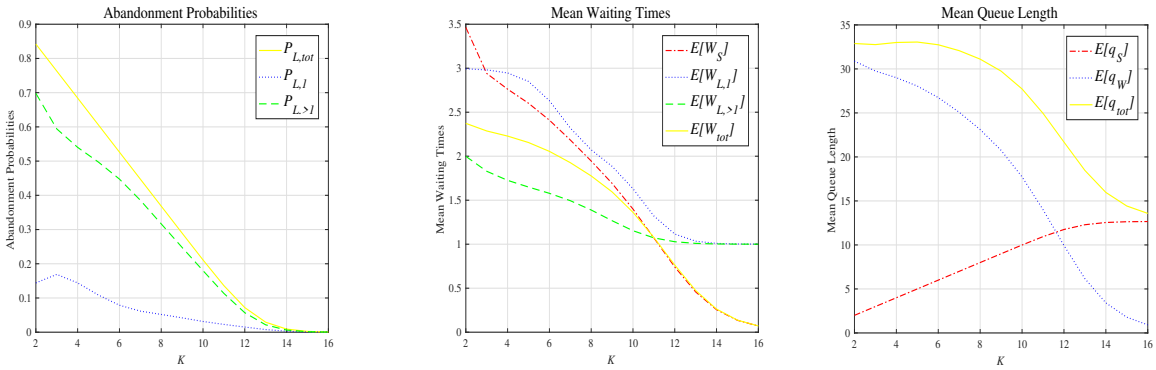

Figure 4.7: Summary of queueing quantities for Example 4.3

One issue related to the analysis of complicated stochastic systems is state space explosion. Specifically, for our $MAP/PH/K+GI$ queue, the number of states in $\Omega(K)$ can be very big. For Examples 4.2 and 4.3, the number of states for each layer is given by $m_a \begin{pmatrix} K+m_s-1 \\ m_s-1 \end{pmatrix}$. We present the number of states as a function of K in Table 4.4.

K	1	5	8	10	12	14	15	50	100
Example 4.2	4	12	18	22	26	30	32	102	202
Example 4.3	12	84	180	264	364	480	544	5304	20604
$m_a = 4$, $m_s = 4$	16	224	660	1144	1820	2720	3264	93704	707404

Table 4.4: Number of states in $\mathcal{S}_+^{(n)} \cup \mathcal{S}_-^{(n)}$ for Examples 4.2 and 4.3

It is shown that, if m_a and m_s are small, Algorithm 2 can be applied for computing queueing quantities for K up to over 100. Since one can generate all kinds of arrival processes and service times even for small m_a and m_s (e.g., Examples 4.2 and 4.3), the method can be useful for researchers and practitioners.

Next, we use our algorithm to address the performance insensitivity to abandonment time distributions, an issue examined in [48].

Example 4.4.([66]) We use the example in Section 6 in [48]. We consider an $M/M/100+GI$ queue with Poisson arrival process $\{D_0 = -105, D_1 = 105\}$ and exponential service time $\{\boldsymbol{\beta} = 1, T = -1\}$. The distribution of the abandonment time τ can be i) an exponential distribution with parameter α, denoted as exp, ii) a uniform distribution on $[0, 1/\alpha]$, denoted as $Unif$, or iii) a phase-type distribution with $\{\boldsymbol{\beta}_\tau = (0.7, 0.3)$ and $T_\tau = \begin{pmatrix} -0.3\alpha & 0 \\ 0 & -79\alpha/30 \end{pmatrix}\}$, denoted as H_2, which is the well-known Hyperexponential distribution, where α is a positive constant.

To use Algorithm 2, we discretize the above three abandonment distributions with $N = 1000$, which gives satisfactory approximation results to the continuous case (as compared to results in [48]). Specifically, for abandonment time τ with an exponential or H_2 distribution, the interval $[0, 3\mathbb{E}[\tau]]$ is divided into $N-1$ identical intervals of length $\delta = 3\mathbb{E}[\tau]/(N-1)$. Then we define $\eta_1 = 0$, $\eta_n = P\{(n-1)\delta \le \tau < n\delta\}$, for $n = 2, 3, ..., N-1$, and $\eta_N = P\{\tau \ge N\delta\}$. For τ with an uniform distribution, the interval $[0, 2\mathbb{E}[\tau]]$ is divided into $N-1$ identical intervals of length $\delta = 2\mathbb{E}[\tau]/(N-1)$. Then we define $\eta_1 = 0$, $\eta_n = 1/(N-1)$, for $n = 2, 3, ..., N-1$, and $\eta_N = 0$.

Paper [48] observes that the performance of the queue is insensitive to abandonment time distributions. Specifically, through simulation, they have observed that the queue

with those three abandonment time distributions perform similarly, even though, for given α, the three abandonment times have different means and variances. Results presented in Table 4.5 indicates that queueing performance, with respect to more queueing quantities than those in [48], is insensitive to abandonment time distributions, which is consistent with the conclusion in [48].

	$\mathbb{E}[a(t)]$			p_L			$p_{L,1}$			$p_{L,>1}$		
α	Exp	$Unif$	H_2	Exp	$Unif$	H_2	Exp	$Unif$	H_2	Exp	$Unif$	H_2
0.1	0.5176	0.5010	0.5562	0.0496	0.0497	0.0493	0.0009	0.0010	0.0009	0.0487	0.0487	0.0484
0.5	0.1216	0.1154	0.1319	0.0601	0.0605	0.0593	0.0037	0.0040	0.0034	0.0564	0.0565	0.0559
1	0.0660	0.0614	0.0728	0.0668	0.0674	0.0658	0.0063	0.0069	0.0057	0.0605	0.0605	0.0601
2	0.0354	0.0319	0.0402	0.0738	0.0747	0.0726	0.0103	0.0116	0.0091	0.0635	0.0631	0.0635
10	0.0074	0.0056	0.0099	0.0886	0.0901	0.0868	0.0276	0.0340	0.0225	0.0609	0.0561	0.0643
	$\mathbb{E}[W_S]$			$\mathbb{E}[W_{L,1}]$			$\mathbb{E}[W_{L,>1}]$			$\mathbb{E}[W]$		
α	Exp	$Unif$	H_2	Exp	$Unif$	H_2	Exp	$Unif$	H_2	Exp	$Unif$	H_2
0.1	0.5187	0.5021	0.5572	0.5390	0.5306	0.5701	0.3460	0.3414	0.3731	0.5103	0.4943	0.5483
0.5	0.1232	0.1170	0.1336	0.1566	0.1556	0.1621	0.1113	0.1100	0.1172	0.1227	0.1167	0.1327
1	0.0673	0.0627	0.0742	0.0995	0.0994	0.1019	0.0720	0.0710	0.0752	0.0678	0.0635	0.0744
2	0.0364	0.0329	0.0413	0.0651	0.0654	0.0659	0.0474	0.0465	0.0492	0.0373	0.0341	0.0420
10	0.0078	0.0059	0.0104	0.0257	0.0264	0.0255	0.0182	0.0169	0.0192	0.0089	0.0072	0.0113
	$p_{q,0}$			$\mathbb{E}[q_S]$			$\mathbb{E}[q_W]$			$\mathbb{E}[q_{tot}]$		
α	Exp	$Unif$	H_2	Exp	$Unif$	H_2	Exp	$Unif$	H_2	Exp	$Unif$	H_2
0.1	0.0340	0.0355	0.0287	99.794	99.784	99.826	53.582	51.904	57.57	153.38	151.69	157.39
0.5	0.2165	0.2238	0.2027	96.686	98.642	98.770	12.880	12.250	13.94	111.57	110.90	112.71
1	0.3316	0.3425	0.3144	97.988	97.922	98.092	7.122	6.667	7.816	105.11	104.59	105.91
2	0.4532	0.4684	0.4323	97.250	97.158	97.377	3.921	3.581	4.410	101.17	100.74	101.79
10	0.7089	0.7356	0.6774	95.699	95.537	95.890	0.936	0.758	1.185	96.63	96.30	97.07

Table 4.5: Summary of queueing quantities for Example 4.4: Part I

The observation seems to hold for queueing systems with a Poisson arrival process and exponential service times. However, it may not hold, even approximately, for queueing systems with a non-Poisson arrival process. Now, we change the customer arrival process from Poisson to MAP with

$$D_0 = \begin{pmatrix} -1 & 0.2 \\ 1 & -310 \end{pmatrix}, \quad D_1 = \begin{pmatrix} 0.1 & 0.7 \\ 1 & 308 \end{pmatrix}. \tag{4.5.3}$$

The average arrival rate is 96.4483. The arrival process is bursty since the arrival rates

in the two states of the underlying Markov chain are drastically different. Quantities in Table 4.5 are reproduced and presented in Table 4.6. Table 4.6 demonstrates that some quantities can be significantly different for the three abandonment times (e.g., $p_{L,1}$ and $\mathbb{E}[q_W]$ for $\alpha \geq 2$), which indicates that the queueing performance is no longer insensitive to the abandonment time distributions.

	$\mathbb{E}[a(t)]$			p_L			$p_{L,1}$			$p_{L,>1}$		
α	Exp	$Unif$	H_2	Exp	$Unif$	H_2	Exp	$Unif$	H_2	Exp	$Unif$	H_2
0.1	1.2385	1.1090	1.4095	0.1470	0.1550	0.1376	0.0007	0.0008	0.0006	0.1463	0.1542	0.1371
0.5	0.3686	0.2842	0.4968	0.2525	0.2719	0.2302	0.0026	0.0038	0.0018	0.2500	0.2681	0.2284
1	0.2020	0.1432	0.3027	0.2994	0.3219	0.2715	0.0043	0.0071	0.0028	0.2951	0.3148	0.2687
2	0.1062	0.0694	0.1783	0.3413	0.3629	0.3105	0.0072	0.0137	0.0042	0.3341	0.3493	0.3063
10	0.0213	0.0117	0.0449	0.4031	0.4134	0.3821	0.0247	0.0617	0.0111	0.3785	0.3517	0.3710
	$\mathbb{E}[W_S]$			$\mathbb{E}[W_{L,1}]$			$\mathbb{E}[W_{L,>1}]$			$\mathbb{E}[W]$		
α	Exp	$Unif$	H_2	Exp	$Unif$	H_2	Exp	$Unif$	H_2	Exp	$Unif$	H_2
0.1	1.5054	1.3608	1.6947	1.8257	1.8502	1.8126	1.3504	1.3316	1.3851	1.4829	1.3567	1.6523
0.5	0.5113	0.4048	0.6691	0.7401	0.7240	0.7535	0.4933	0.4590	0.5296	0.5074	0.4205	0.6374
1	0.2990	0.2190	0.4307	0.4842	0.4518	0.5151	0.3015	0.2672	0.3401	0.3005	0.2358	0.4066
2	0.1671	0.1130	0.2681	0.3077	0.2666	0.3530	0.1763	0.1472	0.2139	0.1712	0.1270	0.2518
10	0.0370	0.0207	0.0753	0.0893	0.0630	0.1383	0.0427	0.0306	0.0640	0.0404	0.0268	0.0718
	$p_{q,0}$			$\mathbb{E}[q_S]$			$\mathbb{E}[q_W]$			$\mathbb{E}[q_{tot}]$		
α	Exp	$Unif$	H_2	Exp	$Unif$	H_2	Exp	$Unif$	H_2	Exp	$Unif$	H_2
0.1	0.3182	0.3321	0.3020	82.269	81.498	83.172	143.03	130.85	159.36	225.30	212.35	242.53
0.5	0.5009	0.5344	0.4622	72.091	70.226	74.247	48.94	40.56	61.47	121.03	110.78	135.72
1	0.5821	0.6210	0.5337	67.567	65.400	70.266	28.99	22.74	39.22	96.55	88.14	109.48
2	0.6546	0.6921	0.6013	63.530	61.443	66.501	16.51	12.25	24.29	80.04	73.69	90.79
10	0.7618	0.7796	0.7254	57.567	56.580	59.597	3.90	2.58	6.93	61.46	59.16	66.52

Table 4.6: Summary of queueing quantities for Example 4.4: Part II

To end this section, we analyze the $M/E_2/100 + E_2$ queue and compare our results to that in [115].

Example 4.5.([66]) We consider the example in Section 2 in [115]. Instead of limiting the waiting spaces to 200 in the original example (i.e., $M/E_2/100/200+E_2$ with 200 extra waiting spaces), we assume that the queue has unlimited waiting space (i.e., $M/E_2/100+E_2$). The arrival process and service time follow a Poisson arrival process $\{D_0 = -102, D_1 = 102\}$ and Erlang-E_2 service time distribution $\{\boldsymbol{\beta} = [1, 0], T = \begin{pmatrix} -2 & 2 \\ 0 & -2 \end{pmatrix}\}$ respec-

tively. The abandonment time τ has a Erlang distribution with phase-type representation $\{\boldsymbol{\beta}_\tau = [1,0], T_\tau = \begin{pmatrix} -2 & 2 \\ 0 & -2 \end{pmatrix}\}$. Similar to Example 4.4, we discretize the above Erlang distribution with $N = 1000$.

For the queueing model, the customer arrival rate is $\lambda = 102$ and the service rate of a server is $\mu_s = 1$. Then $\rho = 1.02$. Since η_N is almost zero, $\eta_N \rho$ is nearly zero and the queueing system is stable. Due to customer abandonments, the (waiting) queue length rarely reaches 200. Thus, the performance of the $M/E_2/100/200 + E_2$ queue and the $M/E_2/100 + E_2(discretized)$ queue is very close. Results are presented in Table 4.7.

Perfoemance Measure	Simulation (Whitt)	Approximation (Whitt)	$MMFF$
$P\{W = 0\}$	0.217±0.0021	0.250	0.2153
p_L	0.0351±0.00029	0.0381	0.0350
$\mathbb{E}[q_W]$	11.52±0.075	11.41	11.620
$\mathbb{E}[q_{tot}]$	109.9±0.092	109.5	110.05
$\mathbb{E}[W_S]$	0.1115±0.00071	0.1102	0.1125
$\mathbb{E}[W_L]$	0.1508±0.00042	0.1521	0.1524

Table 4.7: Summary of queueing quantities for Example 4.5

We note that the half-widths of 95% confidence intervals are shown in the column for simulation results. Table 4.7 shows that our numerical results are fairly close to simulation results. Some of our results are not in the 95% intervals of corresponding quantities since their model has finite waiting space while our model has infinite waiting space. In addition, the following two reasons may contribute to the difference in the numerical results: i) There is always a chance that the actual quantity is outside of the confidence interval; and ii) The abandonment time distributions are different for our and their models.

4.6 Summary

In this chapter, we apply the theory on multi-layer $MMFF$ processes to the $MAP/PH/K+GI$ queue and develop computational methods for queueing quantities such as the customer abandonment probabilities, distributions of waiting times, and the mean queue lengths.

Our main contributions in this chapter are i) combining the $MMFF$ method and the $CSFP$ method to analyse the $MAP/PH/K+GI$ queueing model with a moderately large number of servers; ii) finding queueing quantities related to customers abandoning the queue at the head of the queue and customers abandoning the queue before reaching the head of the queue and queue length distributions, which are difficult to derive by other methods and can be useful for both practitioners and researchers.

There are still some unsolved problems in this chapter for future research including: i) the variance and the distribution of the queue lengths for the $MAP/PH/K+GI$ queue; ii) the $MMAP[K]/PH[K]/N/G[K]$ queue in which there are multiple types of customers; iii) the $MMAP/PH/K+GI$ queue with customer priorities; iv) applying this queueing model to analyze left-without-being-seen (LWBS) phenomenon in the emergency department.

Chapter 5

Double-sided Queues with $MMAP$ and Abandonment

A double-sided queueing model with marked Markovian arrival processes ($MMAP$) and finite discrete abandonment times is investigated in this chapter. Various types of passengers arrive at the system at random times to match any type of taxis that also arrive at random times. Although the structure of our model is simple and similar to classical double-sided queues, the model's generality in terms of arrival processes (from a single type to multiple types) and abandonment times appeals both researchers and practitioners.

To study the queueing model, we use the theory of multi-layer $MMFF$ processes. For the queueing system, we first define three age processes and convert them into a multi-layer $MMFF$ process. Then we analyze the multi-layer $MMFF$ process to find queueing performance measures related to the age processes, matching rates/probabilities, waiting times, and queue lengths for both sides of the queueing system. We obtain a number of aggregate quantities as well as quantities for individual types of inputs, which can be useful for the analysis and design of some stochastic systems, such as passenger-taxi service systems and organ transplantation systems.

This chapter is organized as follows. In Section 5.1, we define the double-sided queueing model. Section 5.2 introduces three age processes for the queueing system and constructs a multi-layer $MMFF$ process. In Section 5.3, we develop a computation method for

computing the joint stationary distribution of the age process. All queueing quantities are obtained in Section 5.4. In Section 5.5, we present several numerical examples. Section 5.6 concludes this chapter.

5.1 Definitions

We define a double-sided queueing model for the stochastic system described above in this section. The structure of the system is depicted in Figure 5.1. We assume that the matching discipline for passengers and taxis is first-come-first-matched, and does not depend on the types of passengers and taxis. Next, components of the queueing model are defined explicitly, including i) Passenger arrival process; ii) Passenger's abandonment time; iii) Taxi arrival process; and iv) Taxi's abandonment time.

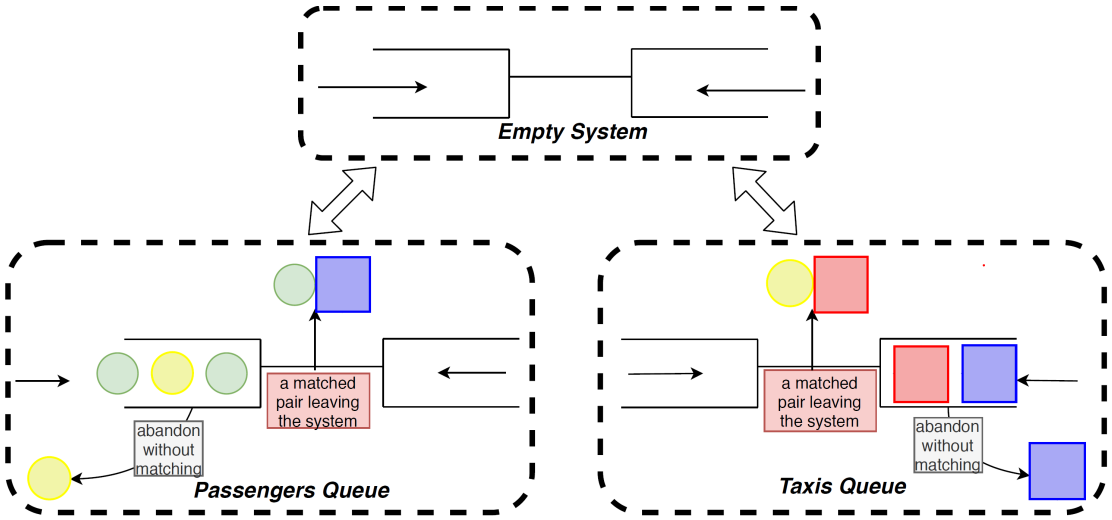

Figure 5.1: A diagram for the double-sided queue with multiple types of inputs

i) Passengers arrive to the queueing system according to a continuous time marked Markovian arrival process ($MMAP$), which is defined by a set of square matrices $(D_0, D_1, ..., D_K)$ of order m_a. Intuitively, D_0 contains the transition rates without an arrival, D_k contains the transition rates with the arrival of a type k passenger,

where $k = 1, ..., K$. The underlying Markov chain of the arrival process $\{I_a(t), t \geq 0\}$ has an irreducible infinitesimal generator $D = D_0 + D_1 + ... + D_K$. The stationary distribution $\boldsymbol{\theta}_a$ of the underlying Markov chain satisfies $\boldsymbol{\theta}_a D = 0$ and $\boldsymbol{\theta}_a \mathbf{e} = 1$. The (average) type k passenger arrival rate is given by $\lambda_k = \boldsymbol{\theta}_a D_k \mathbf{e}$, where $k = 1, ..., K$. Define $\lambda = \sum_{k=1}^{K} \lambda_k$.

ii) The abandonment time τ_k for type k passengers has a discrete distribution: $P\{\tau_k = \tilde{l}_n\} = \eta_{k,n}$, for $k = 1, ..., K$ and $n = 0, 1, ..., N$, where $\tilde{l}_0 = 0 < \tilde{l}_1 < ... < \tilde{l}_{N-1} < \tilde{l}_N = \infty$.

iii) Taxis arrive to the queueing system according to an $MMAP$ with a set of square matrices $(B_0, B_1, ..., B_H)$ of order m_b. Similarly, B_0 contains the transition rates without an arrival, B_h contains the transition rates with the arrival of a type h taxi, where $h = 1, ..., H$. The underlying Markov chain of the arrival process $\{I_b(t), t \geq 0\}$ has an irreducible infinitesimal generator $B = B_0 + B_1 + ... + B_H$. The stationary distribution $\boldsymbol{\theta}_b$ of the underlying Markov chain satisfies $\boldsymbol{\theta}_b B = 0$ and $\boldsymbol{\theta}_b \mathbf{e} = 1$. The (average) type h taxis arrival rate is given by $\mu_h = \boldsymbol{\theta}_b B_h \mathbf{e}$, where $h = 1, ..., H$. Define $\mu = \sum_{h=1}^{H} \mu_h$.

iv) The abandonment time $\hat{\tau}_h$ for type h taxis has a discrete distribution: $P\{\hat{\tau}_h = \hat{l}_m\} = \hat{\eta}_{h,m}$, for $h = 1, ..., H$ and $m = 0, 1, ..., M$, where $\hat{l}_0 = 0 < \hat{l}_1 < ... < \hat{l}_{M-1} < \hat{l}_M = \infty$.

In the rest of this chapter, we make the following assumptions.

- We assume that the arrival processes and the abandonment times are independent.
- We assume $\sum_{k=1}^{K} \eta_{k,N} \lambda_k < \mu$ and $\sum_{h=1}^{H} \hat{\eta}_{h,M} \mu_h < \lambda$ to ensure the stability of the queueing system. We note that $\eta_{k,N}$ can be interpreted as the proportion of type k passengers who stay in the queue forever until being served. Then $\sum_{k=1}^{K} \eta_{k,N} \lambda_k$ is the total number of arrivals per unit time of passengers who have to be served. The condition $\sum_{k=1}^{K} \eta_{k,N} \lambda_k < \mu$ means that there are enough taxis to serve all those passengers. Consequently, the passenger queue can reach a steady state. The condition $\sum_{h=1}^{H} \hat{\eta}_{h,M} \mu_h < \lambda$ can be interpreted similarly.

- We assume the matching time is negligible, so the queue in the system can be a passenger queue or a taxi queue, which never co-exist.

The two sides of the double-sided queueing system are structurally symmetric, which implies that if we can obtain the queueing quantities of one side, we can easily obtain the queueing quantities of the other side by exchanging the parameters of the two sides. This property can also be used to verify the accuracy of the results and to explore the relationship between the quantities.

5.2 Age Processes and a Multi-Layer $MMFF$ Process

In this section, we first define the ages of the passengers and taxis in the double-sided queueing model. Based on the age processes, we introduce a multi-layer $MMFF$ process.

The *age* of a passenger (taxi) is defined as the amount of time that has passed since the passenger (taxi) entered the system. Because the passenger queue and the taxi queue cannot coexist, the ages of the passengers and the ages of taxis can never co-exist either. Let $a_P(t)$ be the age of the passenger at the head of the passenger queue at time t, if the passenger queue is not empty; otherwise, $a_P(t) = 0$. Similarly, let $a_T(t)$ be the age of the taxi at the head of the taxi queue at time t, if the taxi queue is not empty; otherwise, $a_T(t) = 0$. It is obvious that at most one of $a_P(t)$ and $a_T(t)$ can be positive. If both $a_P(t)$ and $a_T(t)$ are zero, then the system is empty at time t. We can reduce the two-dimensional age process $\{(a_P(t), a_T(t)), t \geq 0\}$ to a one-dimensional stochastic process by flipping the age of taxis over the horizontal axis (i.e., the time axis) (see Figure 5.2). Based on this observation, we define a stochastic process $\{a(t), t \geq 0\}$, to be called *the age process*, as $a(t) = a_P(t)$, if $a_P(t) > 0$; $a(t) = -a_T(t)$, if $a_T(t) > 0$; and $a(t) = 0$, otherwise.

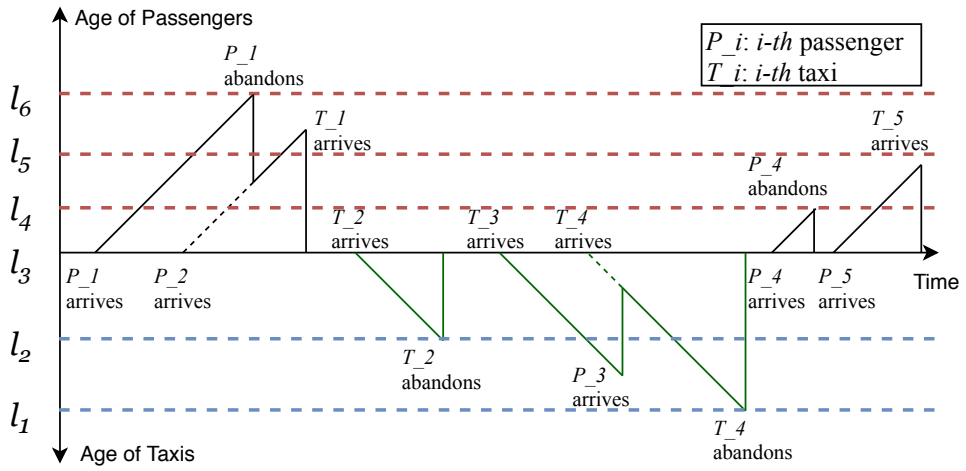

Figure 5.2: The sample path of the age process in a double-sided queue with abandonment ($M = 3$ and $N = 4$)

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 $a(t)$ can be described as follows.

- If $0 \leq l_{M+n} < a(t) < l_{M+n+1}$, for $n = 0, 1, ..., N - 1$, $a(t)$ equals the age of passenger and increases linearly at rate one if there is no taxi arrival; otherwise, a taxi arrives at time t, and $a(t + 0) = \max\{0, a(t) - u\}$, where u is the interarrival time between the departing passenger (due to matching) and the passenger who is currently behind it. If $a(t) = l_{M+n} > 0$ and the type of the passenger at the head of the queue is k, for $n = 1, 2, ..., N-1$ and $k = 1, ..., K$, $a(t)$ continues to increase linearly at rate one with probability $1 - \eta_{k,n}/(\eta_{k,n} + ... + \eta_{k,N})$; otherwise, $a(t+0) = \max\{0, l_{M+n} - u\}$, where u is the interarrival time between the departing passenger (due to abandonment) and the passenger who is currently behind it. We note that some passengers who arrived after the departing passenger may have left the queueing system due to abandonment.

- If $l_{n-1} < a(t) < l_n \leq 0$, for $n = 1, ..., M$, $a(t)$ equals the flipped age of taxi and decreases linearly at rate one if there is no passenger arrival; otherwise, $a(t + 0) = \min\{0, a(t) + u\}$, where u is the interarrival time between the departing taxi (due to

matching) and the taxi that is currently behind it. If $a(t) = l_n < 0$ and the type of taxi at the head of the queue is h, for $n = 1, 2..., M-1$ and $h = 1, ..., H$, $a(t)$ continues to decrease linearly at rate one with probability $1 - \hat{\eta}_{h,M-n}/(\hat{\eta}_{h,M-n} + ... + \hat{\eta}_{h,M})$; otherwise, $a(t+0) = \min\{0, l_n + u\}$, where u is the interarrival time between the departing taxi (due to abandonment) and the taxi that is currently behind it.

- If $a(t) = 0$, $a(t)$ remains zero until either a passenger or a taxi with positive abandonment time arrives. If a passenger arrives first, $a(t)$ would equal the age of that passenger. If a taxi arrives first, $a(t)$ would equal the flipped age of the taxi.

To analyze the age process $\{a(t), t \geq 0\}$, we introduce three supplementary variables $s(t)$, $I_{(a)}(t)$, and $I_{(b)}(t)$, which are related to the type of the passenger or taxi at the head of the queue and the phases $I_a(t)$ and $I_b(t)$ of the two arrival processes.

- $s(t)$: If the system is empty at time t, $s(t) = 0$; If there is a passenger queue, then $s(t) = k$ is the type of passenger at the head of the queue, for $k = 1, ..., K$; and if there is a taxi queue, then $s(t) = h$ is the type of the taxi at the head of the queue, for $h = 1, ..., H$.

- $I_{(a)}(t)$: If $a(t) > 0$, $I_{(a)}(t)$ equals the phase of the passenger arrival process $I_a(t)$ right after the arrival of the passenger at the head of the queue; and if $a(t) \leq 0$, $I_{(a)}(t)$ equals the phase of the passenger arrival process at time t (i.e., $I_{(a)}(t) = I_a(t)$ if $a(t) \leq 0$.) By this definition, the value of $I_{(a)}(t)$ changes when $a(t)$ is going down or $a(t) = 0$.

- $I_{(b)}(t)$: If $a(t) < 0$, $I_{(b)}(t)$ is the phase of the taxi arrival process $I_b(t)$ right after the arrival of the taxi at the head of the queue; and if $a(t) \geq 0$, the phase of the taxi arrival process at time t (i.e., $I_{(b)}(t) = I_b(t)$ if $a(t) \geq 0$.) By this definition, the value of $I_{(b)}(t)$ changes when $a(t)$ is going up or $a(t) = 0$.

It turns out that $\{(a(t), s(t), I_{(a)}(t), I_{(b)}(t)),\ t \geq 0\}$ is a continuous time Markov process with state space

$$\{\{(-\infty, 0) \times \{1, ..., H\}\} \cup \{0\} \cup \{(0, \infty) \times \{1, ..., K\}\}\} \times \{1, ..., m_a\} \times \{1, ..., m_b\}. \quad (5.2.1)$$

We shall recycle the name *age process* again and call the continuous time Markov process $\{(a(t), s(t), I_{(a)}(t), I_{(b)}(t)), t \geq 0\}$ an age process.

Next, based on the age process, we introduce a multi-layer $MMFF$ process $\{(X(t), \phi(t)), t \geq 0\}$. The idea is that convert the down jumps of the age process into periods of decreasing fluid at slope -1 when the age $a(t)$ is above 0, and maintain the increasing periods of the age process for the periods of increasing fluid; change the up jumps of the age process into periods of increasing fluid at slope $+1$ when the age $a(t)$ is below 0, and maintain the decreasing periods of the age process for the periods of decreasing fluid; and keep the periods with $a(t) = 0$ for the periods with zero fluid. As a result, we add fictitious time periods whose lengths equal the heights of the up or down jumps of $a(t)$ to construct the multi-layer $MMFF$ process (see Figure 5.3). More specifically, we define

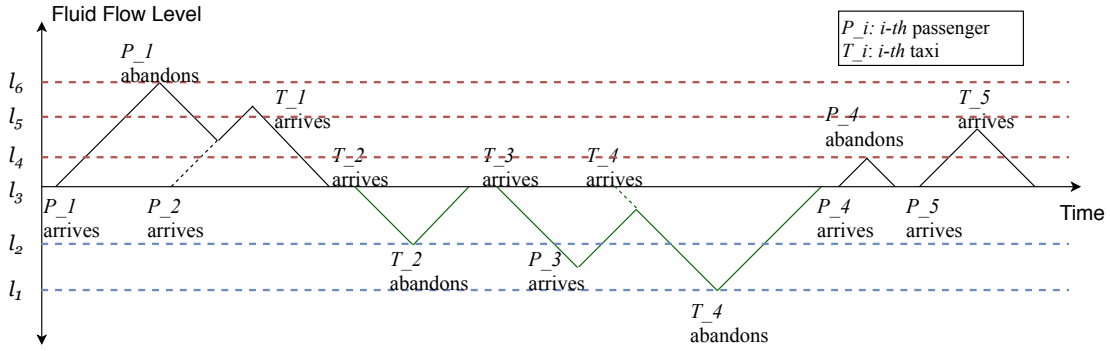

Figure 5.3: The corresponding $MMFF$ process for the age process in Figure 5.2

- The fluid level $X(t)$ equals $a(t)$ during the real time periods. In the fictitious time periods, $X(t)$ is determined by the linear interpolation of the fluid levels at the start and end of the period.

- $\phi(t) = (s(t), I_{(a)}(t), I_{(b)}(t))$ for $t \geq 0$, which is defined as follows. In any time period in which $X(t)$ is increasing, $I_{(a)}(t)$ is fixed at the phase of the passenger arrival process at the beginning of the period, and $I_{(b)}(t) = I_b(t)$, the phase of the taxi arrival process, which evolves in this period of time. In any time period in which $X(t)$ is decreasing, $I_{(b)}(t)$ is fixed at the phase of the taxi arrival process at the beginning

of the period, and $I_{(a)}(t) = I_a(t)$, the phase of the passenger arrival process, which evolves in this period of time. If $X(t) > 0$ and $X(t)$ is decreasing, $s(t) = 0$ since this is the fictitious time period in which the next head of the queue passenger is coming. If $X(t) < 0$ and $X(t)$ is increasing, $s(t) = 0$.

By the above definitions, $I_a(t)$ is frozen and $I_b(t)$ is evolving when $X(t)$ is in increasing periods, so $I_{(a)}(t)$ is fixed at the phase of the passenger arrival process at the beginning of the period and $I_{(b)}(t) = I_b(t)$. Similarly, $I_b(t)$ is frozen and $I_a(t)$ is evolving when $X(t)$ is in decreasing periods, so $I_{(b)}(t)$ is fixed at the phase of the taxi arrival process at the beginning of the period and $I_{(a)}(t) = I_a(t)$. By this definition, if $a(t) = 0$, then $s(t) = 0$ and neither passenger nor taxi is waiting in the system.

The state space and transition matrices of $X(t)$ and $\phi(t)$ 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 $l_{M+N} = \infty$. (Note that we use two constants M and N to define the Borders and Layers for notational convenience.)

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 $n = M + 1, M + 2, ..., M + N$,

$$\begin{array}{rcl} \mathcal{S}_+^{(n)} & = & \{+\} \times \{1, ..., K\} \times \{1, ..., m_a\} \times \{1, ..., m_b\}\,; \\ \mathcal{S}_-^{(n)} & = & \{-\} \times \{1, ..., m_a\} \times \{1, ..., m_b\}\,, \end{array} \tag{5.2.2}$$

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

$$\begin{array}{rcl} \mathcal{S}_+^{(n)} & = & \{+\} \times \{1, ..., m_a\} \times \{1, ..., m_b\}\,; \\ \mathcal{S}_-^{(n)} & = & \{-\} \times \{1, ..., H\} \times \{1, ..., m_a\} \times \{1, ..., m_b\}\,. \end{array} \tag{5.2.3}$$

The Q-matrix $Q^{(n)}$ of the underlying Markov chain is, for $n = M+1, M+2, ..., M+N$,

$$Q^{(n)} = \begin{array}{c} \mathcal{S}_+^{(n)} \\ \mathcal{S}_-^{(n)} \end{array} \begin{pmatrix} I \otimes I \otimes B_0 & \mathbf{e} \otimes I \otimes (B_1 + ... + B_H) \\ (\mathcal{D}_{1,n} \otimes I, ..., \mathcal{D}_{K,n} \otimes I) & \hat{\mathcal{D}}_n \otimes I \end{pmatrix}, \tag{5.2.4}$$

where $\mathcal{D}_{k,n} = D_k(\eta_{k,n-M} + ... + \eta_{k,N})$, $\hat{\mathcal{D}}_n = D - \sum_{k=1}^{K} D_k(\eta_{k,n-M} + ... + \eta_{k,N})$; and, for $n = 1, 2, ..., M$,

$$Q^{(n)} = \begin{array}{c} \mathcal{S}_+^{(n)} \\ \mathcal{S}_-^{(n)} \end{array} \begin{pmatrix} I \otimes \hat{\mathcal{B}}_n & (I \otimes \mathcal{B}_{1,n}, ..., I \otimes \mathcal{B}_{H,n}) \\ \mathbf{e} \otimes (D_1 + ... + D_K) \otimes I & I \otimes D_0 \otimes I \end{pmatrix}, \tag{5.2.5}$$

where $\mathcal{B}_{h,n} = B_h(\hat{\eta}_{h,M-n+1} + ... + \hat{\eta}_{h,M})$, $\hat{\mathcal{B}}_n = B - \sum_{h=1}^{H} B_h(\hat{\eta}_{h,M-n+1} + ... + \hat{\eta}_{h,M})$. We note that $(\hat{\mathcal{D}}_n, \mathcal{D}_{1,n}, ..., \mathcal{D}_{K,n})$ defines a new $MMAP$ of passengers in Layer n, in which $\hat{\mathcal{D}}_n$ can be interpreted as the transition rates without an arrival of passengers, and $\mathcal{D}_{k,n}$ contains the transition rates of an arrival of type k passenger with abandonment time greater than l_n. Similarly, $(\hat{\mathcal{B}}_n, \mathcal{B}_{1,n}, ..., \mathcal{B}_{H,n})$ can be interpreted as a new $MMAP$ of taxi for Layer n.

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

$$\begin{array}{rcl} Q_{bb}^{(M)} & = & \left(D_0 + \sum_{k=1}^{K} \eta_{k,0} D_k\right) \otimes I + I \otimes \left(B_0 + \sum_{h=1}^{H} \hat{\eta}_{h,0} B_h\right); \\ Q_{b+}^{(M)} & = & ((1 - \eta_{1,0}) D_1 \otimes I,\ ...,\ (1 - \eta_{K,0}) D_K \otimes I); \\ Q_{b-}^{(M)} & = & (I \otimes (1 - \hat{\eta}_{1,0}) B_1,\ ...,\ I \otimes (1 - \hat{\eta}_{H,0}) B_H), \end{array} \tag{5.2.6}$$

where $\left((D_0 + \sum_{k=1}^{K} \eta_{k,0} D_k), (1 - \eta_{1,0}) D_1, ..., (1 - \eta_{K,0}) D_K\right)$ defines a new $MMAP$ of passengers for Border M, and $\left((B_0 + \sum_{h=1}^{H} \hat{\eta}_{h,0} B_h), (1 - \hat{\eta}_{1,0}) B_1, ..., (1 - \hat{\eta}_{H,0}) B_H\right)$ defines a new $MMAP$ of taxis for Border M.

4. The transition probabilities entering Border M are given by

$$P_{-b+}^{(M)} = 0; \quad P_{-b-}^{(M)} = 0; \quad P_{-bb}^{(M)} = I; \quad P_{+b+}^{(M)} = 0; \quad P_{+b-}^{(M)} = 0; \quad P_{+bb}^{(M)} = I. \tag{5.2.7}$$

Note that there is no passing or reflection for Border M.

5. All other borders have no state. The probabilities of approaching Border n, for

$1 \leq n \leq N-1$, are $P_{-b+}^{(n+M)} = P_{-bb}^{(n+M)} = 0$, $P_{-b-}^{(n+M)} = I$, $P_{+bb}^{(n+M)} = 0$,

$$P_{+b-}^{(n+M)} = \begin{pmatrix} \frac{\eta_{1,n}}{\eta_{1,n}+\ldots+\eta_{1,N}} I \\ \vdots \\ \frac{\eta_{K,n}}{\eta_{K,n}+\ldots+\eta_{K,N}} I \end{pmatrix};$$

$$P_{+b+}^{(n+M)} = \begin{pmatrix} \frac{\eta_{1,n+1}+\ldots+\eta_{1,N}}{\eta_{1,n}+\eta_{1,n+1}+\ldots+\eta_{1,N}} I & 0 & \ldots & 0 \\ 0 & \frac{\eta_{2,n+1}+\ldots+\eta_{2,N}}{\eta_{2,n}+\eta_{2,n+1}+\ldots+\eta_{2,N}} I & \ldots & 0 \\ \ddots & \ddots & \ddots & \ddots \\ 0 & \ldots & 0 & \frac{\eta_{K,n+1}+\ldots+\eta_{K,N}}{\eta_{K,n}+\eta_{K,n+1}+\ldots+\eta_{K,N}} I \end{pmatrix}. \tag{5.2.8}$$

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

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

$$P_{-b+}^{(n)} = \begin{pmatrix} \frac{\hat{\eta}_{1,M-n}}{\hat{\eta}_{1,M-n}+\ldots+\hat{\eta}_{1,M}} I \\ \vdots \\ \frac{\hat{\eta}_{H,M-n}}{\hat{\eta}_{H,M-n}+\ldots+\hat{\eta}_{H,M}} I \end{pmatrix}. \tag{5.2.9}$$

If the system satisfies the following three conditions,

$$\begin{array}{lll} \sum_{k=1}^{K} \eta_{k,N}\lambda_k/\mu < 1 & \text{and} & \sum_{h=1}^{H} \hat{\eta}_{h,M}\mu_h/\lambda < 1; \\ \sum_{k=1}^{K}(\sum_{l=n}^{N} \eta_{k,l})\lambda_k/\mu \neq 1 & \text{for} & n = 1,2,\ldots,N-1; \\ \sum_{h=1}^{H}(\sum_{l=n}^{M} \hat{\eta}_{h,l})\mu_h/\lambda \neq 1 & \text{for} & n = 1,2,\ldots,M-1. \end{array} \tag{5.2.10}$$

The joint stationary distribution of the multi-layer $MMFF$ process $\{(X(t), \phi(t)), t \geq 0\}$ can be found by Algorithm 1 in Chapter 3. We use similar notations for the border probabilities in Border M as $\mathbf{p}^{(M)}$ and the coefficients $\{\mathbf{u}_+^{(n)}, \mathbf{u}_-^{(n)}, n = 1,2,\ldots,M+N\}$.

Our next two steps for analyzing the queueing model are: i) We find the joint stationary distributions of age processes $\{(a(t), s(t), I_{(a)}(t), I_{(b)}(t)), t \geq 0\}$, $\{(a_P(t), s(t), I_{(a)}(t), I_{(b)}(t)), t \geq 0\}$, and $\{(a_T(t), s(t), I_{(a)}(t), I_{(b)}(t)), t \geq 0\}$ in Section 5.3; ii) We find other queueing quantities in Section 5.4.

5.3 Joint Stationary Distribution of Age Processes

Although Algorithm 1 has been developed for computing the joint stationary distributions for multi-layer $MMFF$ processes, we simplify the algorithm and make it more efficient by taking advantage of the structure of the double-sided queues.

i) **Border Probabilities**: Since the multi-layer $MMFF$ process has only one sticky border (i.e., Border M, which has a non-empty state space), we construct a censored continuous time Markov process $Q_{\mathbf{p}}^{(M)}$ for the border probabilities $\mathbf{p}^{(M)}$ similar to the one in Chapter 4, which satisfies $\mathbf{p}^{(M)} Q_{\mathbf{p}}^{(M)} = 0$ and $\mathbf{p}^{(M)} \mathbf{e} = 1$. But the fluid level can go up and down from the Border M, thus we need to consider the first passage probabilities to return to the (only) sticky Border M from both above and below (Recall we consider only the probabilities from above in Chapter 4 equation 4.3.1), we have

$$Q_{\mathbf{p}}^{(M)} = Q_{bb}^{(M)} + Q_{b+}^{(M)} T_{+}^{(M)} P_{-bb}^{(M)} + Q_{b-}^{(M)} T_{-}^{(M)} P_{+bb}^{(M)}, \tag{5.3.1}$$

where $T_{+}^{(M)}$ 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 $n = M+1, M+2, ..., M+N-1$,

$$\begin{aligned} T_+^{(n-1)} &= \Psi_{+-}^{(l_n - l_{n-1})} + \Lambda_{++}^{(l_n - l_{n-1})}(P_{+b-}^{(n)} + P_{+b+}^{(n)} T_+^{(n)}) \\ &\times \left(I - \widehat{\Psi}_{-+}^{(l_n - l_{n-1})}(P_{+b-}^{(n)} + P_{+b+}^{(n)} T_+^{(n)})\right)^{-1} \widehat{\Lambda}_{--}^{(l_n - l_{n-1})}. \end{aligned} \tag{5.3.2}$$

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

$$\begin{aligned} T_-^{(n+1)} &= \widehat{\Psi}_{-+}^{(l_{n+1} - l_n)} + \widehat{\Lambda}_{--}^{(l_{n+1} - l_n)}(P_{-b+}^{(n)} + P_{-b-}^{(n)} T_-^{(n)}) \\ &\times \left(I - \Psi_{+-}^{(l_{n+1} - l_n)}(P_{-b+}^{(n)} + P_{-b-}^{(n)} T_-^{(n)})\right)^{-1} \Lambda_{++}^{(l_{n+1} - l_n)}. \end{aligned} \tag{5.3.3}$$

ii) **Coefficients**: Let $\boldsymbol{w}(n) = (\mathbf{w}_L^{(n+1)}, \mathbf{w}_U^{(n)})$, for $n = 1, ..., M+N-1$. After we obtain vector $\mathbf{p}^{(M)}$, the coefficients can be obtained by solving the following set of linear equations:

$$\begin{aligned} &\boldsymbol{w}(M) = \mathbf{p}^{(M)}(Q_{b+}^{(M)}, Q_{b-}^{(M)}); \\ &\boldsymbol{w}(1) = \boldsymbol{w}(1) \begin{pmatrix} \Psi_{+-}^{(l_2 - l_1)}(P_{-b+}^{(1)}, P_{-b-}^{(1)}) \\ \widehat{\Psi}^{(1)}(P_{+b+}^{(1)}, P_{+b-}^{(1)}) \end{pmatrix} + \boldsymbol{w}(2) \begin{pmatrix} 0 \\ \widehat{\Lambda}_{--}^{(l_2 - l_1)}(P_{-b+}^{(1)}, P_{-b-}^{(1)}) \end{pmatrix}; \\ &\boldsymbol{w}(n) = \boldsymbol{w}(n) \begin{pmatrix} \Psi_{+-}^{(l_{n+1} - l_n)}(P_{-b+}^{(n)}, P_{-b-}^{(n)}) \\ \widehat{\Psi}_{-+}^{(l_n - l_{n-1})}(P_{+b+}^{(n)}, P_{+b-}^{(n)}) \end{pmatrix} + \boldsymbol{w}(n+1) \begin{pmatrix} 0 \\ \widehat{\Lambda}_{--}^{(l_{n+1} - l_n)}(P_{-b+}^{(n)}, P_{-b-}^{(n)}) \end{pmatrix} \\ &\qquad + \boldsymbol{w}(n-1) \begin{pmatrix} \Lambda_{++}^{(l_n - l_{n-1})}(P_{+b+}^{(n)}, P_{+b-}^{(n)}) \\ 0 \end{pmatrix}, \\ &\qquad \text{for } \; n = 2, ..., M-1, M+1, ..., M+N-2; \\ &\boldsymbol{w}(M+N-1) = \boldsymbol{w}(M+N-1) \begin{pmatrix} \Psi^{(M+N)}(P_{-b+}^{(M+N-1)}, P_{-b-}^{(M+N-1)}) \\ \widehat{\Psi}_{-+}^{(l_{M+N-1} - l_{M+N-2})}(P_{+b+}^{(M+N-1)}, P_{+b-}^{(M+N-1)}) \end{pmatrix} \\ &\qquad + \boldsymbol{w}(M+N-2) \begin{pmatrix} \Lambda_{++}^{(l_{M+N-1} - l_{M+N-2})}(P_{+b+}^{(M+N-1)}, P_{+b-}^{(M+N-1)}) \\ 0 \end{pmatrix}. \end{aligned} \tag{5.3.4}$$

Let $\mathbf{f}(x)$ be the joint stationary density function of the age process $\{(a(t), s(t), I_{(a)}(t), I_{(b)}(t)), t \geq 0\}$. Note that the state space of $\{(a(t), s(t), I_{(a)}(t), I_{(b)}(t)), t \geq 0\}$ is given in (5.2.1). Let $\mathbf{f}^{(n)}(x) = \mathbf{f}(x)$, if $l_{n-1} < x < l_n$, for $n = 1, 2, ..., M+N$.

Theorem 5.1. *([116]) Under the conditions given in Equation (5.2.10), the joint stationary distribution of the age process* $\{(a(t), s(t), I_{(a)}(t),\ I_{(b)}(t)), t \geq 0\}$ *exists and its density function is given by*

$$\begin{aligned} P\{a(t)=0\} &= \hat{\mathbf{p}}^{(M)}\mathbf{e}; \\ \mathbf{f}^{(n)}(x) &= \mathbf{v}_{+}^{(n)} e^{\mathcal{K}^{(n)}(x-l_{n-1})}\Psi^{(n)} + \mathbf{v}_{-}^{(n)} e^{\widehat{\mathcal{K}}^{(n)}(l_n-x)},\ \textit{for } l_{n-1} < x \leq l_n,\ n=1,...,M; \\ \mathbf{f}^{(n)}(x) &= \mathbf{v}_{+}^{(n)} e^{\mathcal{K}^{(n)}(x-l_{n-1})} + \mathbf{v}_{-}^{(n)} e^{\widehat{\mathcal{K}}^{(n)}(l_n-x)}\widehat{\Psi}^{(n)},\ \textit{for } l_{n-1} \leq x < l_n,\ n=M+1,...,M+N. \end{aligned} \tag{5.3.5}$$

where $\mathbf{v}_{+}^{(1)} = 0$, $\mathbf{v}_{-}^{(M+N)} = 0$, $\mathbf{v}_{+}^{(n)} = \mathbf{u}_{+}^{(n)}/\hat{c}_{norm}$, $\mathbf{v}_{-}^{(n)} = \mathbf{u}_{-}^{(n)}/\hat{c}_{norm}$ *and* $\hat{\mathbf{p}} = \mathbf{p}/\hat{c}_{norm}$. *According to* $P\{-\infty < a(t) < \infty\} = 1$ *(i.e., the law of total probability), the normalization factor is*

$$\hat{c}_{norm} = \mathbf{p}^{(M)}\mathbf{e} + \sum_{n=1}^{M}\left(\mathbf{u}_{+}^{(n)}\mathcal{L}_{l_{n-1},l_n}^{\mathcal{K}^{(n)}}\Psi^{(n)} + \mathbf{u}_{-}^{(n)}\widetilde{\mathcal{L}}_{l_{n-1},l_n}^{\widehat{\mathcal{K}}^{(n)}}\right)\mathbf{e} + \sum_{n=M+1}^{M+N}\left(\mathbf{u}_{+}^{(n)}\mathcal{L}_{l_{n-1},l_n}^{\mathcal{K}^{(n)}} + \mathbf{u}_{-}^{(n)}\widetilde{\mathcal{L}}_{l_{n-1},l_n}^{\widehat{\mathcal{K}}^{(n)}}\widehat{\Psi}^{(n)}\right)\mathbf{e}. \tag{5.3.6}$$

Proof. The results are obtained by using Theorem 3.2 and censoring out the decreasing periods when $a(t) > 0$ and increasing periods when $a(t) < 0$ (i.e., $\{(x,j) : 0 < x < \infty, j \in \mathcal{S}_-\} \cup \{(x,j), -\infty < x < 0, j \in \mathcal{S}_+\}$). The evaluation of the integrals $\mathcal{L}_{l_{n-1},l_n}^{\mathcal{K}^{(n)}}$ and $\widetilde{\mathcal{L}}_{l_{n-1},l_n}^{\widehat{\mathcal{K}}^{(n)}}$ is given in Lemma B.1. □

Based on the above joint stationary distribution, the joint distributions related to the passenger age and the taxi age can be obtained immediately. Let $\mathbf{f}_P^{(n)}(x)$, for $x > 0$ and $n = M+1,...,M+N$, and $\mathbf{f}_T^{(n)}(x)$, for $x < 0$ and $n = 1,...,M$, be the joint stationary density functions of the age processes of passengers and taxis, respectively.

Corollary 5.1.1. *([116]) The joint stationary distribution for the age process for passengers is*

$$\begin{aligned} P\{a_P(t)=0\} &= \hat{\mathbf{p}}^{(M)}\mathbf{e} + \sum_{n=1}^{M}\left(\mathbf{v}_{+}^{(n)}\mathcal{L}_{l_{n-1},l_n}^{\mathcal{K}^{(n)}}\Psi^{(n)} + \mathbf{v}_{-}^{(n)}\widetilde{\mathcal{L}}_{l_{n-1},l_n}^{\widehat{\mathcal{K}}^{(n)}}\right)\mathbf{e}; \\ \mathbf{f}_P^{(n)}(x) &= \mathbf{v}_{+}^{(n)} e^{\mathcal{K}^{(n)}(x-l_{n-1})} + \mathbf{v}_{-}^{(n)} e^{\widehat{\mathcal{K}}^{(n)}(l_n-x)}\widehat{\Psi}^{(n)}, \\ &\quad \textit{for}\quad l_{n-1} \leq x < l_n,\ n=M+1,...,M+N. \end{aligned} \tag{5.3.7}$$

The joint stationary distribution for the age process for taxis is

$$\begin{aligned} P\{a_T(t)=0\} &= \hat{\mathbf{p}}^{(M)}\mathbf{e}+\sum_{n=M+1}^{M+N}\left(\mathbf{v}_+^{(n)}\mathcal{L}_{l_{n-1},l_n}^{\mathcal{K}^{(n)}}+\mathbf{v}_-^{(n)}\widetilde{\mathcal{L}}_{l_{n-1},l_n}^{\widehat{\mathcal{K}}^{(n)}}\widehat{\Psi}^{(n)}\right)\mathbf{e}; \\ \mathbf{f}_T^{(n)}(x) &= \mathbf{v}_+^{(n)}e^{\mathcal{K}^{(n)}(-x-l_{n-1})}\Psi^{(n)}+\mathbf{v}_-^{(n)}e^{\widehat{\mathcal{K}}^{(n)}(l_n+x)}, \\ & \quad for \quad -l_{n-1}>x\geq -l_n,\ n=1,...,M. \end{aligned} \tag{5.3.8}$$

Immediately, the probabilities related to the type of queue (i.e., passenger queue or taxi queue) can be obtained: 1) Taxi queue: Empty passenger queue and non-empty taxi queue ($P\{a_P(t)=0,a_T(t)>0\}$); 2) Passenger queue: Empty taxi queue and non-empty passenger queue ($P\{a_P(t)>0,a_T(t)=0\}$); and 3) Empty system ($P\{a_P(t)=0,a_T(t)=0\}$).

$$\begin{aligned} p_T =: P\{a_P(t)=0,a_T(t)>0\} &= \sum_{n=1}^{M}\left(\mathbf{v}_+^{(n)}\mathcal{L}_{l_{n-1},l_n}^{\mathcal{K}^{(n)}}\Psi^{(n)}+\mathbf{v}_-^{(n)}\widetilde{\mathcal{L}}_{l_{n-1},l_n}^{\widehat{\mathcal{K}}^{(n)}}\right)\mathbf{e}; \\ p_P =: P\{a_P(t)>0,a_T(t)=0\} &= \sum_{n=M+1}^{M+N}\left(\mathbf{v}_+^{(n)}\mathcal{L}_{l_{n-1},l_n}^{\mathcal{K}^{(n)}}+\mathbf{v}_-^{(n)}\widetilde{\mathcal{L}}_{l_{n-1},l_n}^{\widehat{\mathcal{K}}^{(n)}}\widehat{\Psi}^{(n)}\right)\mathbf{e}; \\ P\{a_P(t)=0,a_T(t)=0\} &= \hat{\mathbf{p}}^{(M)}\mathbf{e}. \end{aligned} \tag{5.3.9}$$

Similarly, a number of queueing quantities for individual types of passengers and taxis can also be obtained by using the joint density function of the multi-layer $MMFF$ process. The idea is to utilize the underlying states associated with individual types of passengers and taxis (i.e., states for fixed $s(t)$).

Let $\mathbf{f}_P^{(n)}(k,x)$ be the joint density function of $(a(t),s(t)=k,I_{(a)}(t),I_{(b)}(t))$, for $k=1,2,...,K$, $l_{n-1}<x<l_n$, and $n=M+1,...,M+N$. Let $p_P(k)$ be the probability that a type k passenger is at the head of the queue at an arbitrary time. Recall that the state space of $\{(a(t),k,I_{(a)}(t),I_{(b)}(t)),t\geq 0\}$, for $k=1,...,K$, is $(0,\infty)\times\{k\}\times\{1,...,m_a\}\times\{1,...,m_b\}$.

Then we have, for $k = 1, ..., K$, and $l_{n-1} < x < l_n, n = M+1, ..., M+N$,

$$\begin{aligned} \mathbf{f}_P^{(n)}(k,x) &= \mathbf{f}_P^{(n)}(x)(\mathbf{e}(k) \otimes I) = \left(\mathbf{v}_+^{(n)} e^{\mathcal{K}^{(n)}(x-l_{n-1})} + \mathbf{v}_-^{(n)} e^{\widehat{\mathcal{K}}^{(n)}(l_n - x)} \widehat{\Psi}^{(n)} \right) (\mathbf{e}(k) \otimes I), ; \\ p_P(k) &= \sum_{n=M+1}^{M+M} \left(\mathbf{v}_+^{(n)} \mathcal{L}_{l_{n-1},l_n}^{\mathcal{K}^{(n)}} + \mathbf{v}_-^{(n)} \widetilde{\mathcal{L}}_{l_{n-1},l_n}^{\widehat{\mathcal{K}}^{(n)}} \widehat{\Psi}^{(n)} \right) (\mathbf{e}(k) \otimes \mathbf{e}), \end{aligned} \tag{5.3.10}$$

where $\mathbf{e}(k)$ is a column vector of size K whose k-th element is one and all others zero. It is easy to see that $p_P = p_P(1) + p_P(2) + ... + p_P(K)$.

The computational steps for computing the joint stationary distribution are presented

at the end of this section.

Algorithm 3: The joint stationary distribution of the age process

1. Input Parameters: M, N, K, H, $\{\tilde{l}_0 = 0, ..., \tilde{l}_N\}$, $\{\hat{l}_0 = 0, ..., \hat{l}_M\}$, $\{\eta_{k,0}, ..., \eta_{k,N}\}$, for $k = 1, 2, ..., K$, $\{\hat{\eta}_{h,0}, ..., \hat{\eta}_{h,M}\}$, for $h = 1, 2, ..., H$, $\{m_a, D_0, ..., D_K\}$, and $\{m_b, B_0, ..., B_H\}$;

2. Construct transition blocks for the multi-layer $MMFF$ process:

   2.1 Borders: $\{l_0 = -\infty, ..., l_M, ..., l_{M+N} = \infty\}$ 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 $l_{M+N} = \infty$;

   2.2 Construct $\{Q^{(n)}, n = 1, 2, ..., M+N\}$ using Equations (5.2.4) and (5.2.5);

   2.3 Construct $\{Q_{bb}^{(M)}, Q_{b+}^{(M)}, Q_{b-}^{(M)}\}$ using Equation (5.2.6);

   2.4 Construct $\{P_{+b+}^{(n)}, P_{+b0}^{(n)}, P_{+b-}^{(n)}, P_{-b+}^{(n)}, P_{-b0}^{(n)}, P_{-b-}^{(n)}, n = 1, 2, ..., M+N-1\}$ using Equations (5.2.7), (5.2.8) and (5.2.9);

3. Use Algorithm 1 to compute $\{\Psi^{(n)}, \mathcal{K}^{(n)}, \mathcal{U}^{(n)}, \widehat{\Psi}^{(n)}, \widehat{\mathcal{K}}^{(n)}, \widehat{\mathcal{U}}^{(n)}\}$ for $n = 1, ..., M+N$ and $\{\Psi_{+-}^{(l_n - l_{n-1})}, \widehat{\Psi}_{-+}^{(l_n - l_{n-1})}, \Lambda_{++}^{(l_n - l_{n-1})}, \widehat{\Lambda}_{--}^{(l_n - l_{n-1})}\}$ for $n = 2, 3, ..., M+N-1$;

4. Compute $T_+^{(M)}$ and $T_-^{(M)}$ using Equations (5.3.2) and (5.3.3); Construct $Q_{\mathbf{p}}^{(M)}$ using Equation (5.3.1); and solve $\mathbf{p}^{(M)} Q_{\mathbf{p}}^{(M)} = 0$ and $\mathbf{p}^{(M)} \mathbf{e} = 1$ to get border probabilities;

5. Compute the coefficients $\{\boldsymbol{w}(n),\ n = 1, 2, ..., M+N-1\}$ by solving the set of linear equations (5.3.4); and compute the joint density function of the multi-layer $MMFF$ process;

6. Compute $\hat{c}_{norm}$ by using Equation (5.3.6), and use $\hat{c}_{norm}$ to get $\hat{\mathbf{p}}^{(M)}$ and $\{\mathbf{v}_+^{(n)}, \mathbf{v}_-^{(n)},\ n = 1, 2, ..., M+N\}$;

7. Use $\hat{\mathbf{p}}^{(M)}$, $\{\mathbf{v}_+^{(n)}, \mathbf{v}_-^{(n)},\ n = 1, 2, ..., M+N\}$ and Equation (5.3.5) to compute the density function of the age process;

8. Compute the density function of the age process for passengers using Equation (5.3.7) and for taxis using Equation (5.3.8).

5.4 Queueing Quantities

In this section, we use the result for the age process to find queueing quantities related to: i) Matching rate and abandonment probabilities; ii) Waiting times; and iii) Queue lengths. Due to the symmetry between the passenger queue and the taxi queue, in the rest of this section, we shall mainly focus on quantities related to the passenger queue. Formulas for the quantities related to the taxi queue are similar and most of them will be omitted.

5.4.1 Matching Rate and Abandonment Probabilities

Let ω be the number of pairs of matched passengers and taxis per unit time, to be called *the matching rate*.

Proposition 5.1. *([116]) The matching rate of the system is*

$$\begin{aligned}\omega &= \sum_{n=1}^{M}\left(\mathbf{v}_+^{(n)}\mathcal{L}^{\mathcal{K}^{(n)}}_{l_{n-1},l_n}\Psi^{(n)}+\mathbf{v}_-^{(n)}\widetilde{\mathcal{L}}^{\widehat{\mathcal{K}}^{(n)}}_{l_{n-1},l_n}\right)(I\otimes(D-D_0)\otimes I)\mathbf{e}\\ &+ \sum_{n=M+1}^{M+N}\left(\mathbf{v}_+^{(n)}\mathcal{L}^{\mathcal{K}^{(n)}}_{l_{n-1},l_n}+\mathbf{v}_-^{(n)}\widetilde{\mathcal{L}}^{\widehat{\mathcal{K}}^{(n)}}_{l_{n-1},l_n}\widehat{\Psi}^{(n)}\right)(I\otimes I\otimes(B-B_0))\mathbf{e}.\end{aligned} \tag{5.4.1}$$

Proof. According to Theorem 5.1 and Equation (5.3.9), the joint probability that there is a taxi queue can be found by $\sum_{n=1}^{M}\left(\mathbf{v}_+^{(n)}\mathcal{L}^{\mathcal{K}^{(n)}}_{l_{n-1},l_n}\Psi^{(n)}+\mathbf{v}_-^{(n)}\widetilde{\mathcal{L}}^{\widehat{\mathcal{K}}^{(n)}}_{l_{n-1},l_n}\right)$, and the total arrival rate of passengers of all types can be found by $I\otimes(D-D_0)\otimes I$. Therefore, the first summation in Equation (5.4.1) is the matching rate when there is a taxi queue. Similarly, the second summation is the matching rate when there is a passenger queue. Then the sum of those two summations gives the total matching rate. □

Let $\omega_P(k)$ be the matching rate (service rate) of type k passengers (with any type of

taxis). By fixing passenger type at k, for $k = 1, ..., K$, we obtain

$$\begin{aligned}\omega_P(k) &= \sum_{n=1}^{M} \left(\mathbf{v}_+^{(n)} \mathcal{L}_{l_{n-1},l_n}^{\mathcal{K}^{(n)}} \Psi^{(n)} + \mathbf{v}_-^{(n)} \widetilde{\mathcal{L}}_{l_{n-1},l_n}^{\widehat{\mathcal{K}}^{(n)}}\right) (I \otimes D_k \otimes I)\mathbf{e} \\ &\quad + \sum_{n=M+1}^{M+N} \left(\mathbf{v}_+^{(n)} \mathcal{L}_{l_{n-1},l_n}^{\mathcal{K}^{(n)}} + \mathbf{v}_-^{(n)} \widetilde{\mathcal{L}}_{l_{n-1},l_n}^{\widehat{\mathcal{K}}^{(n)}} \widehat{\Psi}^{(n)}\right) (\mathbf{e}(k) \otimes \mathbf{e} \otimes ((B - B_0)\mathbf{e})).\end{aligned} \tag{5.4.2}$$

Using the same argument, we can actually find the matching rate for any type of passengers with any type of taxis, denoted as $\omega(k, h)$ for type k passengers with type h taxis, as

$$\begin{aligned}\omega(k,h) &= \sum_{n=1}^{M} \left(\mathbf{v}_+^{(n)} \mathcal{L}_{l_{n-1},l_n}^{\mathcal{K}^{(n)}} \Psi^{(n)} + \mathbf{v}_-^{(n)} \widetilde{\mathcal{L}}_{l_{n-1},l_n}^{\widehat{\mathcal{K}}^{(n)}}\right) (\mathbf{e}(h) \otimes (D_k\mathbf{e}) \otimes \mathbf{e}) \\ &\quad + \sum_{n=M+1}^{M+N} \left(\mathbf{v}_+^{(n)} \mathcal{L}_{l_{n-1},l_n}^{\mathcal{K}^{(n)}} + \mathbf{v}_-^{(n)} \widetilde{\mathcal{L}}_{l_{n-1},l_n}^{\widehat{\mathcal{K}}^{(n)}} \widehat{\Psi}^{(n)}\right) (\mathbf{e}(k) \otimes \mathbf{e} \otimes (B_h\mathbf{e})).\end{aligned} \tag{5.4.3}$$

Then we find the probability that an arbitrary type k passenger will take a type h taxi as $\omega(k, h)/\lambda_k$ and the probability that an arbitrary type h taxi will be taken by a type k passenger as $\omega(k, h)/\mu_h$, for $k = 1, ..., K$ and $h = 1, ..., H$.

We note that the first summation on the right-hand side of Equation (5.4.3) can be interpreted as the arrival rate of type k passengers times the probability that a type h taxi is at the head of the taxis queue, and the second summation is the arrival rate of type h taxis times the probability that a type k passenger is at the head of the passenger's queue. We also note that in computation, we first compute $\omega(k, h)$ by Equation (5.4.3), then sum up $\omega(k, h)$ over h to obtain $\omega_P(k)$ (i.e., $\sum_{h=1}^{H} \omega(k, h) = \omega_P(k)$), and finally sum up $\omega_p(k)$ over k to find ω (i.e., $\sum_{k=1}^{K} \omega_P(k) = \omega$).

Let $p_{P,S}$ be the probability that an arbitrary passenger is matched by a taxi, and $p_{P,L}$ be the probability that an arbitrary passenger abandons the queue.

Corollary 5.1.2. *([116])*

$$p_{P,S} = \frac{\omega}{\lambda} \tag{5.4.4}$$

and the abandonment probability for passengers is given by $p_{P,L} = 1 - p_{P,S}$.

Similarly, we can find the matching probability of taxis as $p_{T,S} = \omega/\mu$.

Corollary 5.1.3. *([116]) The ratio of the service probabilities of two sides equals to the reciprocal of the ratio of arrival rates of two sides, i.e.,*

$$\frac{p_{P,S}}{p_{T,S}} = \frac{\mu}{\lambda}. \tag{5.4.5}$$

Proof. The result is obtained by Proposition 5.1.2 and the symmetry between the two queues. □

Corollary 5.1.3 also implies that the ratio of the service probabilities of two sides is independent of the abandonment distributions and arrival processes, which is intuitive since the total number of matched passengers equals the total number of matched taxis.

For individual types of passengers, the service probability of type k passengers is given by

$$p_{P,S}(k) = \frac{\omega_P(k)}{\lambda_k} \tag{5.4.6}$$

and the abandonment probability for type k passengers is $P_{P,L}(k) = 1 - P_{P,S}(k)$.

Next, we decompose $p_{P,L}$ into two parts: i) abandonment probability $p_{PL,1}$ of passengers at the head of the queue; and ii) abandonment probability $p_{PL,>1}$ of passengers before reaching the head of the queue. Then we have the following result.

Proposition 5.2. *([116])*

$$\begin{aligned} p_{PL,1} = \frac{1}{\lambda} \sum_{n=M+1}^{M+N-1} \left(\mathbf{v}_+^{(n)} e^{\mathcal{K}^{(n)}(l_n - l_{n-1})} + \mathbf{v}_-^{(n)} \widehat{\Psi}^{(n)} \right) \left(\left(\sum_{k=1}^{K} \frac{\eta_{k,n-M}}{\sum_{m=n}^{M+N} \eta_{k,m-M}} \mathbf{e}(k) \right) \otimes \mathbf{e} \right) \\ + \frac{1}{\lambda} \hat{\mathbf{p}}^{(M)} \left(\left(\sum_{k=1}^{K} \eta_{k,0} D_k \mathbf{e} \right) \otimes \mathbf{e} \right), \end{aligned} \tag{5.4.7}$$

and $p_{PL,>1} = p_{P,L} - p_{PL,1}$.

Proof. For a passenger at the head of the queue to abandon the queue, its age must reach l_n for some $n = M+1, M+2, ..., M+N-1$. If its age reaches l_n, its age must be

greater than l_{n-1}, which occurs with probability $\eta_{k,n-M} + ... + \eta_{k,N}$, if the passenger is of type k. Then the conditional probability that it abandons the queue at $t = l_{M+n}$ is $\eta_{k,n-M}/(\eta_{k,n-M} + ... + \eta_{k,N})$. Combining with the transition rate for the age to reach l_n, which is $\mathbf{f}^{(n)}(l_n)(\mathbf{e}(k) \otimes \mathbf{e})$, if the passenger is of type k, we obtain

$$p_{PL,1} = \frac{1}{\lambda} \sum_{n=M+1}^{M+N-1} \mathbf{f}_P^{(n)}(l_n) \left(\left(\sum_{k=1}^{K} \frac{\eta_{k,n-M}}{\sum_{m=n}^{M+N} \eta_{k,m-M}} \mathbf{e}(k) \right) \otimes \mathbf{e} \right) + \frac{1}{\lambda} \hat{\mathbf{p}}^{(M)} \left(\left(\sum_{k=1}^{K} \eta_{k,0} D_k \mathbf{e} \right) \otimes \mathbf{e} \right). \tag{5.4.8}$$

By Equation (5.3.7), we obtain the desired result. □

Remark: From Proposition 5.2, we can see that by tracking the age process, we can get some quantities for passengers at the head of the queue and before reaching the head of the queue. These quantities can be useful in practice. On the other hand, we can generalize our model by introducing different abandonment time distributions to the passengers at the head of the queue and before reaching the head of the queue. We will discuss this generalization in the next queueing model in Chapter 6.

For type k passengers, the probability that they abandon the system at the head of the queue is

$$\begin{aligned} p_{PL,1}(k) &= \frac{1}{\lambda_k} \sum_{n=M+1}^{M+N-1} \left(\mathbf{v}_+^{(n)} e^{\mathcal{K}^{(n)}(l_n - l_{n-1})} + \mathbf{v}_-^{(n)} \widehat{\Psi}^{(n)} \right) (\mathbf{e}(k) \times \mathbf{e}) \frac{\eta_{k,n-M}}{\sum_{m=n}^{M+N} \eta_{k,m-M}} \\ &\quad + \frac{1}{\lambda_k} \hat{\mathbf{p}}^{(M)} \left((\eta_{k,0} D_k \mathbf{e}) \otimes \mathbf{e} \right). \end{aligned} \tag{5.4.9}$$

Consequently, the probability that a type k passenger abandons the queue before reaching the head of the queue is given by $p_{PL,>1}(k) = p_{P,L}(k) - p_{PL,1}(k)$.

5.4.2 Waiting Times

In this subsection, we consider the waiting times (i.e., the sojourn time) of passengers. We shall consider four types of waiting times. Namely, the waiting times of passengers i) Who are matched with a taxi; ii) Who abandon the system; iii) Who abandon the system when

the passenger is at the head of the passenger queue; and iv) Who abandon the system before reaching the head of the passenger queue.

Proposition 5.3. *([116]) The distribution of waiting time $W_{P,S}$ of passengers received service is*

$$\begin{aligned} P\{W_{P,S}=0\} &= \frac{1}{p_{P,S}\lambda}\sum_{n=1}^{M}\left(\mathbf{v}_{+}^{(n)}\mathcal{L}_{l_{n-1},l_n}^{\mathcal{K}^{(n)}}\Psi^{(n)}+\mathbf{v}_{-}^{(n)}\widetilde{\mathcal{L}}_{l_{n-1},l_n}^{\widehat{\mathcal{K}}^{(n)}}\right)(I\otimes(D-D_0)\otimes I)\mathbf{e};\\ \frac{\mathrm{d}P\{W_{P,S}<x\}}{\mathrm{d}x} &= \frac{1}{p_{P,S}\lambda}\left(\mathbf{v}_{+}^{(n)}e^{\mathcal{K}^{(n)}(x-l_{n-1})}+\mathbf{v}_{-}^{(n)}e^{\widehat{\mathcal{K}}^{(n)}(l_n-x)}\widehat{\Psi}^{(n)}\right)(I\otimes I\otimes(B-B_0))\mathbf{e},\\ &\qquad \text{for } l_{n-1}\le x<l_n,\ n=M+1,M+2,...,M+N. \end{aligned} \tag{5.4.10}$$

The distribution of waiting time $W_{PL,1}$ of passengers abandoning the system at the head of the queue is given by, for $n=M$ (i.e., $l_M=0$),

$$P\{W_{PL,1}=0\}=\frac{1}{p_{PL,1}\lambda}\hat{\mathbf{p}}^{(M)}\left(\left(\sum_{k=1}^{K}\eta_{k,0}D_k\mathbf{e}\right)\otimes\mathbf{e}\right), \tag{5.4.11}$$

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

$$P\{W_{PL,1}=l_n\}=\frac{\left(\mathbf{v}_{+}^{(n)}e^{\mathcal{K}^{(n)}(l_n-l_{n-1})}+\mathbf{v}_{-}^{(n)}\widehat{\Psi}^{(n)}\right)}{p_{PL,1}\lambda}\left(\left(\sum_{k=1}^{K}\frac{\eta_{k,n-M}}{\eta_{k,n-M}+...+\eta_{k,N}}\mathbf{e}(k)\right)\otimes\mathbf{e}\right). \tag{5.4.12}$$

The waiting time $W_{PL,>1}$ of a passenger that abandons the queue before reaching the head of the queue, we have, for $n=M,M+1,...,M+N-1$,

$$\begin{aligned} P\{W_{PL,>1}=l_n\} &= \frac{1}{p_{PL,>1}\lambda}\left(\sum_{m=n+1}^{M+N}\left(\mathbf{v}_{+}^{(m)}\mathcal{L}_{l_{m-1},l_m}^{\mathcal{K}^{(m)}}\Psi^{(m)}+\mathbf{v}_{-}^{(m)}\widetilde{\mathcal{L}}_{l_{m-1},l_m}^{\widehat{\mathcal{K}}^{(m)}}\right)\right.\\ &\quad\left.\times\left(\left(\sum_{k=1}^{K}\eta_{k,n-M}D_k\mathbf{e}\right)\otimes\mathbf{e}\right)\right). \end{aligned} \tag{5.4.13}$$

Proof. That $W_{P,S}=0$ occurs if there is a taxi queue when a passenger arrives, the ratio of the probability that passengers take taxi without waiting and $P_{P,S}$ leads to the expression for $P\{W_{P,S}=0\}$. Similarly, we use the transition rate ratio to find, for $l_{n-1}\le x<l_n,\ n=$

$M+1, M+2, ..., M+N$,

$$\frac{\mathrm{d}P\{W_{P,S} < x\}}{\mathrm{d}x} = \frac{1}{p_{P,S}\lambda}\mathbf{f}_P^{(n)}(x)(I \otimes (B - B_0))\mathbf{e}, \tag{5.4.14}$$

which leads to the desired result.

For $W_{PL,1}$, the probability $P\{W_{PL,1} = 0\}$ is obtained from the proof of Proposition 5.2. For $n \geq M+1$, it is clear that $W_{PL,1} = l_n$ if $a(t)$ reaches l_n from below and abandonment occurs. The probability for $W_{PL,1}$ to reach l_n is $\mathbf{f}_P^{(n)}(l_n)\mathbf{e}/(p_{PL,1}\lambda)$. The probability for the abandonment to occur is $\eta_{k,n-M}/(\eta_{k,n-M} + ... + \eta_{k,N})$, if the passenger at the head of the queue is of type k. Then expression (5.4.12) can be obtained easily.

To find the distribution of $W_{PL,>1}$, We need to go back to the joint stationary distribution of the multi-layer $MMFF$ process $\{(X(t), \phi(t), t \geq 0\}$. When the multi-layer $MMFF$ process is in $S_-^{(n)}$ and there is a passenger arrival, which may take place if $X(t) \geq 0$, the arriving passenger 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 type of the passenger at the head of the queue is k. Since passenger arrivals take place only when the fluid level of the multi-layer $MMFF$ process is decreasing, we censor out the periods of time in which the fluid level is increasing when $X(t) > 0$, and the periods of time in which the fluid level is decreasing when $X(t) < 0$. This censored process has the same normalization factor as the age process. Then the desired results are obtained by the density function of the $MMFF$ process. □

Remark: According to the law of total probability, we must have $P\{W_{P,S} < \infty\} = 1$ and $\sum_{n=M}^{M+N-1} P\{W_{PL,1} = l_n\} = 1$, which can be used to check computation accuracy. The law of total probability $\sum_{n=M}^{M+N-1} P\{W_{PL,>1} = l_n\} = 1$ can also be used to check computation accuracy.

The mean waiting time for served passenger $\mathbb{E}[W_{P,S}]$ can be calculated by:

$$\mathbb{E}[W_{P,S}] = \frac{1}{p_{P,S}\lambda}\sum_{n=M+1}^{M+N}\left(\mathbf{v}_+^{(n)}\mathcal{M}_{l_{n-1},l_n}^{\mathcal{K}^{(n)}} + \mathbf{v}_-^{(n)}\widetilde{\mathcal{M}}_{l_{n-1},l_n}^{\widehat{\mathcal{K}}^{(n)}}\widehat{\Psi}^{(n)}\right)(I \otimes I \otimes (B - B_0))\mathbf{e}, \tag{5.4.15}$$

where closed form expressions of $\mathcal{M}_{l_{n-1},l_n}^{\mathcal{K}^{(n)}}$ and $\widetilde{\mathcal{M}}_{l_{n-1},l_n}^{\widehat{\mathcal{K}}^{(n)}}$ can be found in Lemma B.1. The

distribution of the waiting time W_P of an arbitrary passenger can be found from that of $W_{P,S}$, $W_{PL,1}$, and $W_{PL,>1}$. The mean waiting time can be found by

$$\mathbb{E}[W_P] = p_{P,S}\mathbb{E}[W_{P,S}] + p_{PL,1}\mathbb{E}[W_{PL,1}] + p_{PL,>1}\mathbb{E}[W_{PL,>1}]. \tag{5.4.16}$$

In Equations (5.4.10), (5.4.12), and (5.4.13), if only the components of the vectors associated with $s(t) = k$ are included in the summations, the joint distributions of $(W_{P,S}, s(t) = k)$, $(W_{PL,1}, s(t) = k)$, and $(W_{PL,>1}, s(t) = k)$ can be obtained. Consequently, after proper normalization, the waiting times of type k passengers can be obtained. The joint stationary distribution of waiting time $W_{P,S}$ of type k passengers received service is

$$\begin{aligned}
&P\{W_{P,S} = 0|s(t) = k\} = \frac{1}{p_{P,S}(k)\lambda_k}\sum_{n=1}^{M}\left(\mathbf{v}_+^{(n)}\mathcal{L}_{l_{n-1},l_n}^{\mathcal{K}^{(n)}}\Psi^{(n)} + \mathbf{v}_-^{(n)}\widetilde{\mathcal{L}}_{l_{n-1},l_n}^{\widehat{\mathcal{K}}^{(n)}}\right)(\mathbf{e}\otimes(D_k\mathbf{e})\otimes\mathbf{e});\\
&\frac{\mathrm{d}P\{W_{P,S} < x|s(t) = k\}}{\mathrm{d}x}\\
&= \frac{1}{p_{P,S}(k)\lambda_k}\left(\mathbf{v}_+^{(n)}e^{\mathcal{K}^{(n)}(x-l_{n-1})} + \mathbf{v}_-^{(n)}e^{\widehat{\mathcal{K}}^{(n)}(l_n-x)}\widehat{\Psi}^{(n)}\right)(\mathbf{e}(k)\otimes\mathbf{e}\otimes(-B_0\mathbf{e})),\\
&\qquad\qquad \text{for } l_{n-1} \le x < l_n,\ n = M+1, M+2, ..., M+N.
\end{aligned} \tag{5.4.17}$$

The distribution of waiting time $W_{PL,1}$ of type k passengers abandoning the system at the head of the queue is given by, for $k = 1, ..., K$

$$\begin{aligned}
&P\{W_{PL,1} = 0|s(t) = k\} = \frac{1}{p_{PL,1}(k)\lambda_k}\hat{\mathbf{p}}^{(M)}\left((\eta_{k,0}D_k\mathbf{e})\otimes\mathbf{e}\right),\\
&P\{W_{PL,1} = l_n|s(t) = k\} = \frac{\left(\mathbf{v}_+^{(n)}e^{\mathcal{K}^{(n)}(l_n-l_{n-1})} + \mathbf{v}_-^{(n)}\widehat{\Psi}^{(n)}\right)}{p_{PL,1}(k)\lambda_k}\left(\left(\frac{\eta_{k,n-M}}{\eta_{k,n-M}+...+\eta_{k,N}}\mathbf{e}(k)\right)\otimes\mathbf{e}\right).\\
&\qquad\qquad \text{for } \ n = M+1, M+2, ..., M+N-1
\end{aligned} \tag{5.4.18}$$

The waiting time $W_{PL,>1}$ of type k passengers abandoning the queue before reaching the

head of the queue, we have, for $k = 1, ..., K$ and $n = M, M+1, ..., M+N-1$,

$$P\{W_{PL,>1} = l_n | s(t) = k\} = \frac{1}{p_{PL,>1}(k)\lambda_k} \left(\sum_{m=n+1}^{M+N} \left(\mathbf{v}_+^{(m)} \mathcal{L}_{l_{m-1},l_m}^{\mathcal{K}^{(m)}} \Psi^{(m)} + \mathbf{v}_-^{(m)} \widetilde{\mathcal{L}}_{l_{m-1},l_m}^{\widehat{\mathcal{K}}^{(m)}} \right) \times ((\eta_{k,n-M} D_k \mathbf{e}) \otimes \mathbf{e}) \right). \tag{5.4.19}$$

5.4.3 Queue Lengths

Let $q_P(t)$ be the queue length of passengers at an arbitrary time t. Similar to Subsection 4.4.3, the z-transform of $q_P(t)$ can be derived based on the joint distribution of the age process. We divide the interval $(t-x, t)$ into $(t-l_{M+1}, t)$, $(t-l_{M+2}, t-l_{M+1})$, ..., $(t-x, t-l_{n-1})$, if $l_{n-1} < x < l_n$. For a type k passenger arrived in $(t-l_{M+1}, t)$, it is still in the system at time t with probability $1-\eta_{k,0}$. The conditional probability generating function of the number of such passengers is $\exp\left\{(D_0 + \sum_{k=1}^K (\eta_{k,0} + (1-\eta_{k,0})z) D_k) l_{M+1}\right\}$ (see Lemma B.2 or Theorem 2.5.1 in [62]). For passengers arrived in $(t-l_{M+2}, t-l_{M+1})$, they abandon the queue before t with probability $\eta_{k,0}+\eta_{k,1}$ and are still in the queue at time t with probability $1-\eta_{k,0}-\eta_{k,1}$, if the passenger is of type k. The conditional probability generating function is given by $\exp\left\{\left(D_0 + \sum_{k=1}^K (\eta_{k,0} + \eta_{k,1} + (1-\eta_{k,0}-\eta_{k,1})z) D_k\right)(l_{M+2} - l_{M+1})\right\}$. In general, for passengers arrived in $(t-l_n, t-l_{n-1})$, they abandon the queue before t with probability $1-\xi_{k,n}$ and are still in the queue at time t with probability $\xi_{k,n}$, where $\xi_{k,n} = \sum_{i=n-M}^N \eta_{k,i}$, for $n = M, M+1, ..., M+N$. The probability generating function is given by $\exp\left\{\left(D_0 + \sum_{k=1}^K (1-\xi_{k,n} + \xi_{k,n} z) D_k\right)(l_n - l_{n-1})\right\}$. In general, define

$$P^*(n, z, y) = \exp\left\{ \left(D_0 + \sum_{k=1}^K (1 - \xi_{k,n} + \xi_{k,n} z) D_k \right) y \right\}. \tag{5.4.20}$$

Conditioning on $a(t)$ at an arbitrary time t, the probability generating function of $q_P(t)$ can be found as

$$\begin{aligned}\mathbb{E}[z^{q_P(t)}] &= \hat{\mathbf{p}}^{(M)}\mathbf{e} + \sum_{n=1}^{M}\left(\mathbf{v}_+^{(n)}\mathcal{L}^{\mathcal{K}^{(n)}}_{l_{n-1},l_n}\Psi^{(n)} + \mathbf{v}_-^{(n)}\widetilde{\mathcal{L}}^{\widehat{\mathcal{K}}^{(n)}}_{l_{n-1},l_n}\right)\mathbf{e} \\ &\quad + z\sum_{n=M+1}^{M+N}\int_{l_{n-1}}^{l_n}\mathbf{f}_P^{(n)}(x)\left(I\otimes\left(P^*(n,z,x-l_{n-1})\prod_{m=n-1}^{M+1}P^*(m,z,b_m)\right)\otimes I\right)\mathrm{d}x\mathbf{e},\end{aligned} \tag{5.4.21}$$

where $b_m = l_m - l_{m-1}$, for $m = M+1, M+2, ..., M+N$. By Theorem 2.3.2 in [62] or Lemma B.2, we have

$$\left.\frac{\partial P^*(n,z,x)\mathbf{e}}{\partial z}\right|_{z=1} = \left(\sum_{k=1}^{K}\xi_{k,n}\lambda_k\right)x\mathbf{e} + (e^{Dx}-I)(D-\mathbf{e}\boldsymbol{\theta}_a)^{-1}\left(\sum_{k=1}^{K}\xi_{k,n}D_k\right)\mathbf{e}. \tag{5.4.22}$$

Recall that $\lambda_k = \boldsymbol{\theta}_a D_k \mathbf{e}$ and $D = D_0 + D_1 + ... + D_K$. Consequently, we obtain

Proposition 5.4. *([116]) The mean queue length is given by*

$$\begin{aligned}\mathbb{E}[q_P(t)] &= \left.\frac{\partial\mathbb{E}[z^{q_P(t)}]}{\partial z}\right|_{z=1} = \sum_{n=M+1}^{M+N}\left(\mathbf{v}_+^{(n)}\mathcal{L}^{\mathcal{K}^{(n)}}_{l_{n-1},l_n} + \mathbf{v}_-^{(n)}\widetilde{\mathcal{L}}^{\widehat{\mathcal{K}}^{(n)}}_{l_{n-1},l_n}\widehat{\Psi}^{(n)}\right)\mathbf{e} \\ &\quad + \sum_{n=M+1}^{M+N}\sum_{m=M+1}^{n-1}\int_{l_{n-1}}^{l_n}\mathbf{f}_P^{(n)}(x)\left(I\otimes e^{D(x-l_m)}\otimes I\right)\mathrm{d}x \\ &\quad \times\left(I\otimes\left(\left(\textstyle\sum_{k=1}^{K}\xi_{k,m}\lambda_k\right)b_m I + (e^{Db_m}-I)(D-\mathbf{e}\boldsymbol{\theta}_a)^{-1}\left(\textstyle\sum_{k=1}^{K}\xi_{k,m}D_k\right)\right)\otimes I\right)\mathbf{e} \\ &\quad + \sum_{n=M+1}^{M+N}\int_{l_{n-1}}^{l_n}\mathbf{f}_P^{(n)}(x)\left(I\otimes\left(\left(\sum_{k=1}^{K}\xi_{k,n}\lambda_k\right)(x-l_{n-1})I\right.\right. \\ &\qquad\qquad \left.\left. + (e^{D(x-l_{n-1})}-I)(D-\mathbf{e}\boldsymbol{\theta}_a)^{-1}\left(\sum_{k=1}^{K}\xi_{k,n}D_k\right)\right)\otimes I\right)\mathbf{e}\mathrm{d}x.\end{aligned} \tag{5.4.23}$$

The evaluation of the integrals can be found in Lemma B.1.

The mean queue length and mean waiting time satisfy the well-known Little's law: $\mathbb{E}[q_P(t)] = \lambda\mathbb{E}[W_P]$, which is useful for checking computation accuracy.

Let $q_P(k,t)$ be the queue length of type k passengers in the queue at time t. Let $\mathbf{z} = (z_1, ..., z_K)$. Define

$$\hat{P}^*(n, \mathbf{z}, y) = \exp\left\{ \left(D_0 + \sum_{k=1}^{K} (1 - \xi_{k,n} + \xi_{k,n} z_k) D_k \right) y \right\}, \tag{5.4.24}$$

which is the joint probability generating function of the numbers of the K types of passengers arrived in $(0, y)$ and are still in the queue at time y. By taking into consideration of the passenger at the head of the passenger queue, the joint conditional probability generating function of the numbers of the K types of passengers is obtained as follows.

Proposition 5.5. *([116])*

$$\begin{aligned}\mathbb{E}\left[\Pi_{k=1}^{K} z_k^{q_P(k.t)}\right] &= \hat{\mathbf{p}}^{(M)}\mathbf{e} + \sum_{n=1}^{M} \left(\mathbf{v}_+^{(n)} \mathcal{L}_{l_{n-1},l_n}^{\mathcal{K}^{(n)}} \Psi^{(n)} + \mathbf{v}_-^{(n)} \widetilde{\mathcal{L}}_{l_{n-1},l_n}^{\widehat{\mathcal{K}}^{(n)}} \right) \mathbf{e} \\ &+ \sum_{n=M+1}^{M+N} \int_{l_{n-1}}^{l_n} \mathbf{f}_P^{(n)}(x) \left(I(\mathbf{z}) \otimes \left(\hat{P}^*(n, \mathbf{z}, x - l_{n-1}) \prod_{m=n-1}^{M+1} \hat{P}^*(m, \mathbf{z}, b_m) \right) \otimes I \right) \mathrm{d}x\mathbf{e},\end{aligned} \tag{5.4.25}$$

where $I(\mathbf{z}) = diag(\mathbf{z})$ (i.e., the matrix with $\{z_1, ..., z_K\}$ on its diagonal and all other elements being zero).

The probability generating functions of the numbers of passengers of individual passenger types in steady state can be obtained accordingly. For instance, the probability generating function of $q_P(k,t)$ is obtained by setting $\mathbf{z} = (1, ..., 1, z_k, 1, ..., 1)$ in Equation (5.4.25). By routine calculations, the mean number of type k passengers in the system can

be found accordingly:

$$\begin{aligned}\mathbb{E}[q_P(k,t)] = &\sum_{n=M+1}^{M+N}\left(\mathbf{v}_+^{(n)}\mathcal{L}_{l_{n-1},l_n}^{\mathcal{K}^{(n)}} + \mathbf{v}_-^{(n)}\widetilde{\mathcal{L}}_{l_{n-1},l_n}^{\widehat{\mathcal{K}}^{(n)}}\widehat{\Psi}^{(n)}\right)(\mathbf{e}(k)\otimes\mathbf{e})\\ &+\sum_{n=M+1}^{M+N}\sum_{m=M+1}^{n-1}\int_{l_{n-1}}^{l_n}\mathbf{f}_P^{(n)}(x)\left(I\otimes e^{D(x-l_m)}\otimes I\right)\mathrm{d}x\\ &\times\left(I\otimes\left((\xi_{k,m}\lambda_k b_m)I + (e^{Db_m}-I)(D-\mathbf{e}\boldsymbol{\theta}_a)^{-1}\xi_{k,m}D_k\right)\otimes I\right)\mathbf{e}\\ &+\sum_{n=M+1}^{M+N}\int_{l_{n-1}}^{l_n}\mathbf{f}_P^{(n)}(x)\left(I\otimes(\xi_{k,n}\lambda_k(x-l_{n-1})I\right.\\ &\left.+\ (e^{D(x-l_{n-1})}-I)(D-\mathbf{e}\boldsymbol{\theta}_a)^{-1}\xi_{k,n}D_k\right)\otimes I\Big)\mathbf{e}\mathrm{d}x.\end{aligned} \tag{5.4.26}$$

The mean queue length and mean waiting time of type k passengers satisfy the well-known Little's law: $\mathbb{E}[q_P(k,t)] = \lambda_k\mathbb{E}[W_P|s(t)=k]$, which is useful for checking computation accuracy.

5.4.4 Summary of Queueing Quantities

Since we derived many queueing quantities in this section, we summarize all important queueing quantities in Table 5.1 to enable readers to quickly locate the meaning and equations of these quantities.

Notations	Quantities	Equations	Individual type k
$\mathbf{f}^{(n)}(x)$	Density of the age process	(5.3.5)	$\mathbf{f}_P^{(n)}(k,x)$ (5.3.10)
p_T & p_P	Probability of the type of the queue	(5.3.9)	$p_P(k)$ (5.3.10)
ω	Matching rate	(5.4.1)	$\omega_P(k)$ (5.4.2) & $\omega(k,h)$ (5.4.3)
$P_{P,S}$, $P_{P,L}$, $P_{PL,1}$ & $P_{PL,>1}$	Abandonment probabilities	(5.4.4), (5.4.7)	$P_{P,S}(k)$, $P_{P,L}(k)$ (5.4.6), $P_{PL,1}(k)$ & $P_{PL,>1}(k)$ (5.4.9)
$W_{P,S}$ & $\mathbb{E}[W_{P,S}]$	Waiting time of served passengers	(5.4.10), (5.4.15)	$(W_{P,S}, s(t)=k)$ (5.4.17)
$W_{PL,1}$ & $W_{PL,>1}$	Waiting time of abandoned passengers	(5.4.12), (5.4.13)	$(W_{PL,1}, s(t)=k)$ (5.4.18) & $(W_{PL,>1}, s(t)=k)$ (5.4.19)
$\mathbb{E}[W_P]$	Mean waiting time	(5.4.16)	
$q_P(t)$ & $\mathbb{E}[q_P(t)]$	Queue lengths	(5.4.21), (5.4.23)	$q_P(k,t)$ (5.4.25) (5.4.26)

Table 5.1: Summary of queueing quantities in Chapter 5

5.5 Numerical Examples

In this section, we show several numerical examples to gain insight into double-sided queues. We also compare our approach to the diffusion approximation methods in [82] (see Example 5.2).

Example 5.1. We consider a double-sided queue with $M = 6$, $N = 6$, and abandonment distributions given in Table 5.2.

n	1	2	3	4	5	6
$\tilde{l}_n$	1	2	3	4	5	∞
η_n	0.1	0.1	0.2	0.3	0.2	0.1
$\hat{l}_n$	1	2	3	4	5	∞
$\hat{\eta}_n$	0.1	0.1	0.2	0.3	0.2	0.1

Table 5.2: Distributions of abandonment times for Example 5.1

We assume that there is one type of input for each side. For the arrival processes, we first consider two Poisson processes for the two sides with $\lambda = 3$ and $\mu = 4.5$, respectively. Then we use two Markovian arrival processes with parameters

$$D_0 = \begin{pmatrix} -3 & 1 \\ 1 & -5 \end{pmatrix}, \quad D_1 = \begin{pmatrix} 1 & 1 \\ 1 & 3 \end{pmatrix}; \quad B_0 = \begin{pmatrix} -5 & 2 \\ 1 & -7 \end{pmatrix}, \quad B_1 = \begin{pmatrix} 2 & 1 \\ 2 & 4 \end{pmatrix}, \quad (5.5.1)$$

which have the same average arrival rates as the Poisson processes. Our objective is to compare queueing quantities for such queueing systems. The distributions of the waiting times of matched passengers and taxis for the two cases are plotted in Figure 5.4, and all other queueing quantities are collected in Table 5.3.

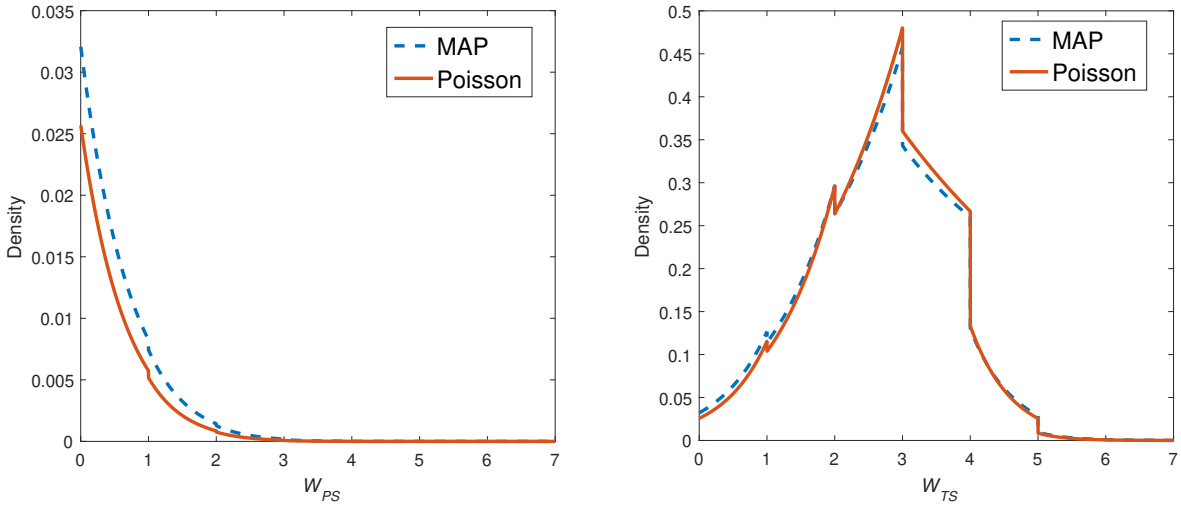

Figure 5.4: Comparison of the stationary density functions of W_{PS} and W_{TS} for Example 5.1

Model	$p_{P,S}$	$p_{P,L}$	$\mathbb{E}[W_{P,S}]$	$\mathbb{E}[W_{P,L}]$	$\mathbb{E}[W_{PL.1}]$	$\mathbb{E}[W_{PL.>1}]$	$\mathbb{E}[W_P]$	$\mathbb{E}[q_P]$
MAP	0.999418	0.000582	0.013302	1.206424	1.260689	1.170121	0.013996	0.041988
Poisson	0.999637	0.000363	0.009128	1.172935	1.213901	1.142643	0.009550	0.028650
Model	$p_{T,S}$	$p_{T,L}$	$\mathbb{E}[W_{T,S}]$	$\mathbb{E}[W_{T,L}]$	$\mathbb{E}[W_{TL.1}]$	$\mathbb{E}[W_{TL.>1}]$	$\mathbb{E}[W_T]$	$\mathbb{E}[q_T]$
MAP	0.666279	0.333721	2.585029	2.389564	3.381066	2.125594	2.519798	11.339090
Poisson	0.666425	0.333575	2.632891	2.364083	3.354171	2.101333	2.543223	11.444504

Table 5.3: Queueing quantities for Example 5.1

We observe that the model with MAP has higher abandonment probabilities for both sides. Intuitively, that is caused by the higher squared coefficient of variation of the MAPs. We also observed that the waiting time distributions are not that different for the two models. On the other hand, some other quantities can be different significantly (e.g., $p_{P,L}$), especially for passengers.

Example 5.2.([116]) We use the numerical example in [82] to compare our approach to the diffusion approximation methods. Since the distributions of abandonment times in their paper are exponentially distributed for both passengers and taxis with parameters θ and γ

respectively, we discretize the abandonment distributions with $M = N = 1000$ and $M = N = 2000$. The distribution of interarrival time can be: i) Exponential with parameter α and β respectively; ii) Erlang Distribution with parameter α and β respectively. We compare the queue length, which is defined as the difference between the passenger queue and the taxi queue, and present the results in Table 5.4.

Parameters		Multi-layer $MMFF$ process		Liu et al, 2015			
$(\alpha, \beta)(1, 2)$	(θ, γ)	$M = N = 1000$	$M = N = 2000$	Simulation	Poisson	Diffusion 1	Diffusion 2
Erlang(2)	(1, 2)	-0.4396 (2.58%)	-0.4334 (1.13%)	-0.4285 (±0.0018)	-0.3858 (9.96%)	-0.4493 (4.87%)	-0.5 (16.69%)
Erlang(2)	(0.1, 0.2)	-5.0847 (2.04%)	-5.0407 (1.15%)	-4.9832 (±0.015)	-4.9719 (0.23%)	-4.9983 (0.30%)	-5 (0.34%)
Erlang(2)	(0.01, 0.02)	-50.8731 (1.57%)	-50.4535 (0.73%)	-50.089 (±0.1507)	-50 (0.18%)	-50 (0.18%)	-50 (0.18%)
Exponential	(1, 2)	-0.4007 (3.38%)	-0.3933 (1.46%)	-0.3876 (±0.002)	-0.3858 (0.46%)	-0.3178 (18%)	-0.5 (29%)
Exponential	(0.1, 0.2)	-5.0620 (1.69%)	-5.0171 (0.79%)	-4.9779 (±0.0157)	-4.9719 (0.12%)	-4.9776 (0.01%)	-5 (0.45%)
Exponential	(0.01, 0.02)	-50.8615 (1.80%)	-50.4406 (0.96%)	-49.9609 (±0.142)	-50 (0.08%)	-50 (0.08%)	-50 (0.08%)

Table 5.4: Comparison of the queue lengths between $MMFF$ processes and diffusion methods

The half widths of 90% confidence intervals are shown in the Simulation column in Table 5.4 for the simulation results. The percentage numbers in the bracket in all other columns show the error rate comparing to the simulation results. Those numbers show that our numerical results are fairly close to the simulation results no matter what are the parameters, especially for the non-Poisson arrival and light traffic cases, our results outperform all other methods (i.e., Poisson approximation and two diffusion models) in their paper. Because we use discrete distribution to approximate the exponential abandonment distribution in [82], our results can be improved by increasing the number of support points (i.e., M and N).

Example 5.3.([116]) In this example, we consider multiple types of passengers and taxis. All the input parameters are presented in Table 5.5.

Passengers	Type	Arrival	$\tilde{l}_1 = 2$	$\tilde{l}_2 = 3$	$\tilde{l}_3 = 4$	$\tilde{l}_4 = 5$	$\tilde{l}_5 = \infty$
$D_0 = \begin{pmatrix} -7, & 1 \\ 1, & -5 \end{pmatrix}$	1	$D_1 = \begin{pmatrix} 1, & 2 \\ 0, & 2 \end{pmatrix}$	$\eta_{1,1} = 0$	$\eta_{1,2} = 0$	$\eta_{1,3} = 0$	$\eta_{1,4} = 0.9$	$\eta_{1,5} = 0.1$
	2	$D_2 = \begin{pmatrix} 2, & 0 \\ 0, & 1 \end{pmatrix}$	$\eta_{2,1} = 0.4$	$\eta_{2,2} = 0.3$	$\eta_{2,3} = 0.2$	$\eta_{2,4} = 0.1$	$\eta_{2,5} = 0$
	3	$D_3 = \begin{pmatrix} 1, & 0 \\ 0, & 1 \end{pmatrix}$	$\eta_{3,1} = 0.1$	$\eta_{3,2} = 0.1$	$\eta_{3,3} = 0.1$	$\eta_{3,4} = 0.1$	$\eta_{3,5} = 0.6$
Taxis	Type	Arrival	$\hat{l}_1 = 2$	$\hat{l}_2 = 3$	$\hat{l}_3 = 4$	$\hat{l}_4 = 5$	$\hat{l}_5 = \infty$
$B_0 = \begin{pmatrix} -3, & 0 \\ 1, & -10 \end{pmatrix}$	1	$B_1 = \begin{pmatrix} 1, & 0 \\ 1, & 6 \end{pmatrix}$	$\hat{\eta}_{1,1} = 0.2$	$\hat{\eta}_{1,2} = 0.2$	$\hat{\eta}_{1,3} = 0.2$	$\hat{\eta}_{1,4} = 0.2$	$\hat{\eta}_{1,5} = 0.2$
	2	$B_2 = \begin{pmatrix} 2, & 0 \\ 0, & 2 \end{pmatrix}$	$\hat{\eta}_{2,1} = 0.9$	$\hat{\eta}_{2,2} = 0$	$\hat{\eta}_{2,3} = 0$	$\hat{\eta}_{2,4} = 0$	$\hat{\eta}_{2,5} = 0.1$

Table 5.5: Parameters of Example 5.3

We can get the matching rate of any type of passengers and any type of taxis (Table 5.6). Other quantities for each type are collected in Table 5.7, and the density function of waiting times for served passengers and taxis of each type are plot in Figure 5.5. Note that $\mathbb{E}[W_{P,S}](k)$, $\mathbb{E}[W_{P,L}](k)$, $\mathbb{E}[W_P](k)$ are for conditional mean waiting times of individual types of passengers. Quantity notations for taxis are defined similarly.

Matching Rate $\omega(k,h)$	Type 1 taxis	Type 2 taxis	$p_{PL,1}(k)\lambda_k$	$p_{PL,>1}(k)\lambda_k$	Sum
Type 1 passengers	**0.6380**	**1.2759**	0.2436	0.0925	$2.25 = \lambda_1$
Type 2 passengers	**0.1098**	**0.2196**	0.0867	0.8344	$1.25 = \lambda_2$
Type 3 passengers	**0.2521**	**0.5043**	0.0338	0.2098	$1 = \lambda_3$
$p_{TL,1}(h)\mu_h$	1.3e-05	0.0001	Note: Arrival rate and leaving rate (i.e., matched and abandoned) should be equal to each other.		
$p_{TL,>1}(h)\mu_h$	3.7e-06	3.3e-05			
Sum	$1 = \mu_1$	$2 = \mu_2$			

Table 5.6: Matching rate for any type

Passenger	$p_{P,S}(k)$	$p_{PL,1}(k)$	$p_{PL,>1}(k)$	$\mathbb{E}[W_{P,S}](k)$	$\mathbb{E}[W_{P,L}](k)$	$\mathbb{E}[W_P](k)$	$\mathbb{E}[q_P(k,t)]$
Type 1	0.8506	0.1083	0.0411	3.6704	5.0000	3.8690	8.7052
Type 2	0.2636	0.0689	0.6675	2.8797	2.6733	2.7277	3.4097
Type 3	0.7564	0.0338	0.2098	3.7870	2.9662	3.5871	3.5871
Taxi	$p_{T,S}(h)$	$p_{TL,1}(h)$	$p_{TL,>1}(h)$	$\mathbb{E}[W_{T,S}](h)$	$\mathbb{E}[W_{T,L}](h)$	$\mathbb{E}[W_T](h)$	$\mathbb{E}[q_T(k,t)]$
Type 1	0.99998	1.33e-05	3.70e-06	0.0023	2.0275	0.0023	0.0023
Type 2	0.99993	5.83e-05	1.63e-05	0.0022	2.0000	0.0023	0.0047

Table 5.7: Queueing quantities for Example 5.3

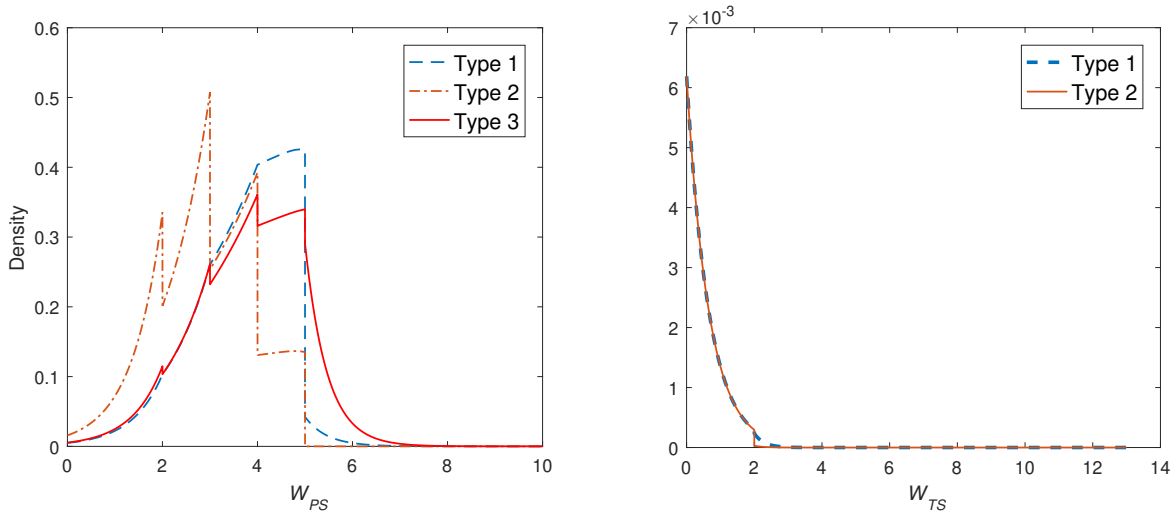

Figure 5.5: The stationary density functions of W_{PS} and W_{TS} for Example 5.3

As demonstrated in Figure 5.5 and Tables 5.6 and 5.7, the queueing performance for the individual types of passengers can be significantly different. One reason for that is the difference between the arrival patterns of different types of passengers and the abandonment time distributions. On the other hand, the performances of the two types of taxis are similar, even though their arrival pattern and their abandonment time distributions are different. Thus, this example indicates that no single element (e.g., the arrival process, abandonment time) can dominate the performance of the queueing system.

5.6 Summary

This chapter studies a double-sided queueing model with marked Markovian arrival processes and finite discrete abandonment times using multi-layer $MMFF$ processes. We develop computational methods for queueing quantities related to the age processes, the abandonment probabilities, waiting times, and queue lengths.

The contributions of this chapter are i) introducing a general double-sided queueing model with multiple types of inputs and general abandonment; ii) analyzing this queueing

model by multi-layer $MMFF$ processes and computing a variety of queueing quantities.

Our model can be further extended to a more general case in which the arrival processes depend on the age of the passenger or taxi at the head of the queue. Double-sided queues with customer priority are also worth considering. The study of such queues is more challenging, and it will be the subject of future research.

Chapter 6

Double-sided Queues with $BMAP$ and Abandonment

In this chapter, 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 to the system with a batch of orders to be matched by counterparts and the abandonment time of a customer depends on its batch size and its position in the queue. First, we obtain the joint stationary distributions of the age processes via the stationary analysis of a multi-layer $MMFF$ process. Second, using the joint stationary 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.

This chapter is organized as follows. In Section 6.1, the queueing model of interest is introduced. In Section 6.2, the age process of the buyer/seller at the head of the queue is introduced. Based on the age process, a multi-layer $MMFF$ process is constructed and analyzed. In Section 6.3, the stationary distribution of the multi-layer $MMFF$ process is utilized to obtain various queueing quantities. In Section 6.4, we apply our model to analyze a vaccine inventory system. Section 6.5 concludes this chapter.

6.1 Definitions

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 it's *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 are 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 coexist 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 research.

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 in a position behind the head of the queue or at 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 its abandonment time is sampled, conditioning on its batch size. 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. 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 the maximum batch size of a buyer is K. Buyers arrive at 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 chapter, 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),

$$\sum_{k=1}^{K} k\lambda_k \eta_{k,N} < \mu \quad \text{and} \quad \sum_{k=1}^{K} k\mu_k \hat{\eta}_{k,M} < \lambda. \tag{6.1.1}$$

In addition, 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.

It is important to note the following two points. First, 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 thesis. Second, the abandonment time distributions of a buyer (seller) are used indirectly. Instead of sampling abandonment times at arrivals or other state changing epochs in our analysis, 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.

In order to illustrate the queueing model, we present Example 6.1, in which we will draw the sample path of the age process to show the dynamics of the model in Section 6.2 and give the queueing quantities as we derive the results in Section 6.3.

Example 6.1. A double-sided queue with maximum batch size $K = 3$, and $M = N = 5$. All the input parameters are presented in Table 6.1.

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 6.1: Parameters of Example 6.1

6.2 The Age Process

In this section, we first introduce the age processes of the buyers and sellers in the double-sided queueing model in Subsection 6.2.1. Then we convert the age processes into a multi-layer $MMFF$ process for the analysis of the queueing model in Subsection 6.2.2.

6.2.1 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., the time axis) (See the green lines in Figure 6.1), 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, $s_S(t) = 0$. We flip the batch size of sellers over the horizontal axis (i.e., the time 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.

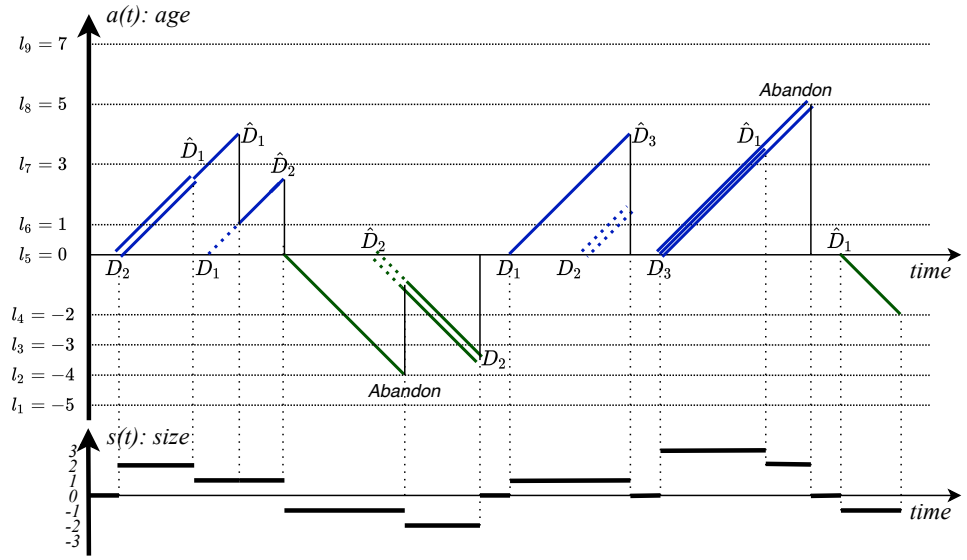

Figure 6.1: A sample path of the age process of Example 6.1

Note that in Figure 6.1, 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 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 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 and increases 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 great 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 case is symmetric to 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 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 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 great 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 case is symmetric to 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 $BMAPs$ and obtain the joint stationary distribution of the process $\{(a(t), s(t), I_b(t), I_s(t)), t \geq 0\}$. 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" them 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 $I_b(t)$ is frozen and the seller arrival process $I_s(t)$ 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 $I_b(t)$ is evolving and the seller arrival process $I_s(t)$ 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 $I_b(t)$ is evolving and the seller arrival process $I_s(t)$ 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 $I_b(t)$ is frozen and the seller arrival process $I_s(t)$ 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 $a(t)$ 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

$$\{\{(-\infty, 0) \times \{-K, ..., -1\}\} \cup \{0\} \cup \{(0, \infty) \times \{1, ..., K\}\}\} \times \{1, ..., m_b\} \times \{1, ..., m_s\}. \tag{6.2.1}$$

6.2.2 Multi-Layer *MMFF* Process

We replace jumps in the age process with linear increasing and decreasing periods and construct a multi-layer $MMFF$ process $\{(X(t), s(t), I_{(b)}(t), I_{(s)}(t)),\ t \geq 0\}$ (See Figure 6.2). Specifically, the original increasing and decreasing periods in the age process are kept in the $MMFF$ process and called *real periods*. However, the up and down jumps in the age process are replaced with *fictitious periods* of linear increase at rate 1 and decrease at rate -1, respectively, in the $MMFF$ process. The lengths of these fictitious periods equal the heights of the up or down jumps. In the real periods, $X(t) = a(t)$, and $|s(t)|$ is

the remaining batch size of the buyer (seller) at the head of the queue. In the fictitious periods, $X(t)$ is determined by the fluid level at the start of the period plus the product of 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 (buyer) (to be matched by orders of buyers (sellers) in the queue);

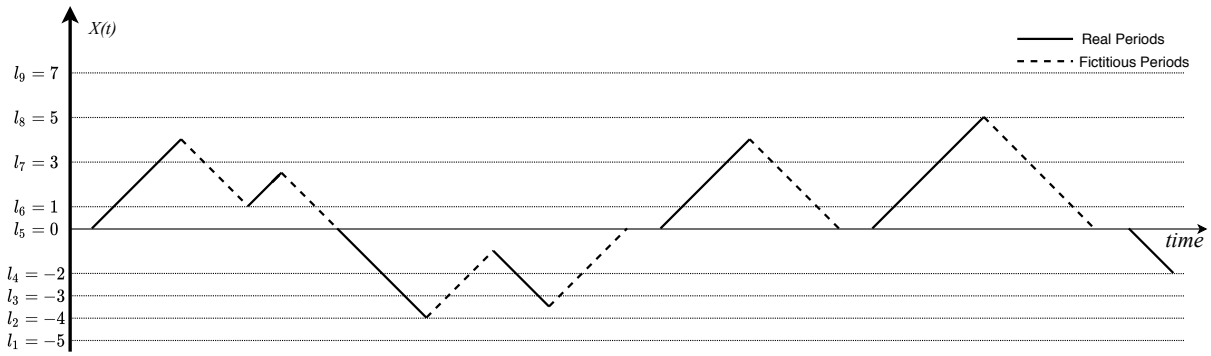

Figure 6.2: The corresponding $MMFF$ process for the age process in Figure 6.1

Let $\phi(t) = (s(t), I_{(b)}(t), I_{(s)}(t))$ for $t \geq 0$, a multi-layer $MMFF$ process $\{(X(t), \phi(t)), t \geq 0\}$ is well defined. Next, we use the items of the queueing model to construct the multi-layer $MMFF$ process $\{(X(t), \phi(t)), t \geq 0\}$. 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 $n = M + 1, M + 2, ..., M + N$,

$$\begin{array}{rcl} \mathcal{S}_+^{(n)} & = & \{1, ..., K\} \times \{1, ..., m_b\} \times \{1, ..., m_s\}\,; \\ \mathcal{S}_-^{(n)} & = & \{-K + 1, ..., 0\} \times \{1, ..., m_b\} \times \{1, ..., m_s\}\,; \end{array} \tag{6.2.2}$$

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

$$\begin{array}{rcl} \mathcal{S}_+^{(n)} & = & \{0, ..., K-1\} \times \{1, ..., m_b\} \times \{1, ..., m_s\}\,; \\ \mathcal{S}_-^{(n)} & = & \{-K, ..., -1\} \times \{1, ..., m_b\} \times \{1, ..., m_s\}\,. \end{array} \tag{6.2.3}$$

The transition rate matrix $Q^{(n)}$ of the underlying Markov chain is, for $n = M + 1, M + 2, ..., M + N$,

$$Q^{(n)} = \left(\begin{array}{cccc|ccccc} I \otimes \hat{D}_0 & I \otimes \hat{D}_1 & ... & I \otimes \hat{D}_{K-1} & I \otimes \hat{D}_K & & & & \\ & I \otimes \hat{D}_0 & ... & I \otimes \hat{D}_{K-2} & I \otimes \hat{D}_{K-1} & I \otimes \hat{D}_K & & & \\ & & \ddots & \vdots & \vdots & \vdots & \ddots & & \\ & & & I \otimes \hat{D}_0 & I \otimes \hat{D}_1 & I \otimes \hat{D}_2 & ... & I \otimes \hat{D}_K \\ \hline \mathcal{D}_{K,n} \otimes I & \mathcal{D}_{K-1,n} \otimes I & ... & \mathcal{D}_{1,n} \otimes I & \bar{\mathcal{D}}_n \otimes I & & & & \\ & \mathcal{D}_{K,n} \otimes I & ... & \mathcal{D}_{2,n} \otimes I & \mathcal{D}_{1,n} \otimes I & \bar{\mathcal{D}}_n \otimes I & & & \\ & & \ddots & \vdots & \vdots & \vdots & \ddots & & \\ & & & \mathcal{D}_{K,n} \otimes I & \mathcal{D}_{K-1,n} \otimes I & \mathcal{D}_{K-2,n} \otimes I & ... & \bar{\mathcal{D}}_n \otimes I \end{array}\right), \tag{6.2.4}$$

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

$$Q^{(n)} = \left(\begin{array}{cccc|ccccc} I \otimes \bar{\hat{\mathcal{D}}}_n & I \otimes \hat{\mathcal{D}}_{1,n} & ... & I \otimes \hat{\mathcal{D}}_{K-1,n} & I \otimes \hat{\mathcal{D}}_{K,n} & & & & \\ & I \otimes \bar{\hat{\mathcal{D}}}_n & ... & I \otimes \hat{\mathcal{D}}_{K-2,n} & I \otimes \hat{\mathcal{D}}_{K-1,n} & I \otimes \hat{\mathcal{D}}_{K,n} & & & \\ & & \ddots & \vdots & \vdots & \vdots & \ddots & & \\ & & & I \otimes \bar{\hat{\mathcal{D}}}_n & I \otimes \hat{\mathcal{D}}_{1,n} & I \otimes \hat{\mathcal{D}}_{2,n} & ... & I \otimes \hat{\mathcal{D}}_{K,n} \\ \hline D_K \otimes I & D_{K-1} \otimes I & ... & D_1 \otimes I & D_0 \otimes I & & & & \\ & D_K \otimes I & ... & D_2 \otimes I & D_1 \otimes I & D_0 \otimes I & & & \\ & & \ddots & \vdots & \vdots & \vdots & \ddots & & \\ & & & D_K \otimes I & D_{K-1} \otimes I & D_{K-2} \otimes I & ... & D_0 \otimes I \end{array}\right), \tag{6.2.5}$$

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

$$\begin{array}{rcl} Q_{bb}^{(M)} & = & \left(D_0 + \sum_{k=1}^K \dot{\eta}_{k,0} D_k\right) \otimes I + I \otimes \left(\hat{D}_0 + \sum_{k=1}^K \dot{\hat{\eta}}_{k,0} \hat{D}_k\right); \\ Q_{b+}^{(M)} & = & ((1 - \dot{\eta}_{K,0}) D_K \otimes I, \ \ldots, \ (1 - \dot{\eta}_{1,0}) D_1 \otimes I); \\ Q_{b-}^{(M)} & = & \left(I \otimes (1 - \dot{\hat{\eta}}_{1,0}) \hat{D}_1, \ \ldots, \ I \otimes (1 - \dot{\hat{\eta}}_{K,0}) \hat{D}_K\right). \end{array} \tag{6.2.6}$$

4. The transition probabilities of approaching Border M are given by, in matrix form, $P_{-b+}^{(M)} = 0$; $P_{-bb}^{(M)} = \left(I, \dot{\hat{\eta}}_{1,0}, \dot{\hat{\eta}}_{2,0}, ..., \dot{\hat{\eta}}_{K-1,0}\right)$; $P_{+b-}^{(M)} = 0$; $P_{+bb}^{(M)} = (\dot{\eta}_{K-1,0}, \dot{\eta}_{K-2,0}, ..., \dot{\eta}_{1,0}, I)$;

$$\begin{array}{rcl} P_{-b-}^{(M)} & = & \begin{pmatrix} 0 & & & & & \\ (1 - \dot{\hat{\eta}}_{1,0}) I & 0 & & & & \\ & (1 - \dot{\hat{\eta}}_{2,0}) I & 0 & & & \\ & & \ddots & \ddots & & \\ & & & (1 - \dot{\hat{\eta}}_{K-1,0}) I & 0 \end{pmatrix}; \\ P_{+b+}^{(M)} & = & \begin{pmatrix} 0 & (1 - \dot{\eta}_{K-1,0}) I & & & \\ & 0 & (1 - \dot{\eta}_{K-2,0}) I & & \\ & & 0 & \ddots & \\ & & & \ddots & (1 - \dot{\eta}_{1,0}) I \\ & & & & 0 \end{pmatrix}. \end{array} \tag{6.2.7}$$

   Note that the batch size of buyer or seller does not change when crossing Border M if the buy or seller has 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 states using the matrix $P_{+b+}^{(M)}$ for up-crossing and matrix $P_{-b-}^{(M)}$ for down-crossing.

   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 other borders have no state. The probabilities of approaching Border n, for

$1 \leq n \leq N-1$, are $P_{-b+}^{(n+M)} = 0$, $P_{-b-}^{(n+M)} = I$,

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

$$P_{+b+}^{(n+M)} = \begin{pmatrix} \frac{\dot{\eta}_{K,n+1}+\ldots+\dot{\eta}_{K,N}}{\dot{\eta}_{K,n}+\dot{\eta}_{K,n+1}+\ldots+\dot{\eta}_{K,N}} I & 0 & \ldots & 0 \\ 0 & \frac{\dot{\eta}_{K-1,n+1}+\ldots+\dot{\eta}_{K-1,N}}{\dot{\eta}_{K-1,n}+\dot{\eta}_{K-1,n+1}+\ldots+\dot{\eta}_{K-1,N}} I & \ldots & 0 \\ \ddots & \ddots & \ddots & \ddots \\ 0 & \ldots & 0 & \frac{\dot{\eta}_{1,n+1}+\ldots+\dot{\eta}_{1,N}}{\dot{\eta}_{1,n}+\dot{\eta}_{1,n+1}+\ldots+\dot{\eta}_{1,N}} I \end{pmatrix}. \tag{6.2.8}$$

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

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

$$P_{-b+}^{(n)} = \begin{pmatrix} 0 & \ldots & 0 & \frac{\dot{\hat{\eta}}_{1,M-n}}{\dot{\hat{\eta}}_{1,M-n}+\ldots+\dot{\hat{\eta}}_{1,M}} I \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \ldots & 0 & \frac{\dot{\hat{\eta}}_{K,M-n}}{\dot{\hat{\eta}}_{K,M-n}+\ldots+\dot{\hat{\eta}}_{K,M}} I \end{pmatrix}. \tag{6.2.9}$$

The joint density function of this multi-layer $MMFF$ process is important because some of the queueing quantities can be derived directly from this process instead of the age process. The joint stationary density of the process is given in the following theorem.

Theorem 6.1. *Under the conditions in Equation (6.2.10).*

$$\begin{array}{llll} \sum_{k=1}^{K} k\eta_{k,N}\lambda_k/\mu < 1 & \text{and} & \sum_{k=1}^{K} k\hat{\eta}_{k,M}\mu_k/\lambda < 1; \\ \sum_{k=1}^{K} k(\sum_{l=n}^{N}\eta_{k,l})\lambda_k/\mu \neq 1 & \text{for} & n = 1, 2, ..., N-1; \\ \sum_{k=1}^{K} k(\sum_{l=n}^{M}\hat{\eta}_{k,l})\mu_h/\lambda \neq 1 & \text{for} & n = 1, 2, ..., M-1. \end{array} \tag{6.2.10}$$

The joint density function of $\{(X(t), \phi(t)), t \geq 0\}$ *is given by, for* $x < l_1$,

$$(\boldsymbol{\pi}_+^{(1)}(x),\ \boldsymbol{\pi}_-^{(1)}(x)) = \mathbf{u}_-^{(1)} e^{\widehat{\mathcal{K}}^{(1)}(l_1 - x)}(\widehat{\Psi}^{(1)}, I); \tag{6.2.11}$$

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

$$(\boldsymbol{\pi}_+^{(n)}(x),\ \boldsymbol{\pi}_-^{(n)}(x)) = \mathbf{u}_+^{(n)} e^{\mathcal{K}^{(n)}(x - l_{n-1})}(I, \Psi^{(n)}) + \mathbf{u}_-^{(n)} e^{\widehat{\mathcal{K}}^{(n)}(l_n - x)}(\widehat{\Psi}^{(n)}, I); \tag{6.2.12}$$

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

$$(\boldsymbol{\pi}_+^{(M+N)}(x),\ \boldsymbol{\pi}_-^{(M+N)}(x)) = \mathbf{u}_+^{(M+N)} e^{\mathcal{K}^{(M+N)}(x - l_{M+N-1})}(I, \Psi^{(M+N)}). \tag{6.2.13}$$

We can compute the joint stationary density function of the process $\{(X(t), \phi(t)), t \geq 0\}$ by Algorithm 1. We can also use Algorithm 3 to compute the density function with a small change with respect to the border probabilities in Step 4 as follows.

i) **Border Probabilities**: Similar to Chapter 5, we need to 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 $\mathbf{p}^{(M)}\mathbf{e} = 1$. However, the fluid in this system can cross or enter Border M when approaching the border. (Recall that the fluid can only enter Border M when approaching it in Chapter 5). Thus we have

$$\begin{aligned} Q_{\mathbf{p}}^{(M)} &= Q_{bb}^{(M)} + \left(Q_{b+}^{(M)}, Q_{b-}^{(M)}\right) \begin{pmatrix} T_+^{(M)} & 0 \\ 0 & T_-^{(M)} \end{pmatrix} \\ &\quad \times \left(I - \begin{pmatrix} 0 & P_{-b-}^{(M)} T_-^{(M)} \\ P_{+b+}^{(M)} T_+^{(M)} & 0 \end{pmatrix} \right)^{-1} \begin{pmatrix} P_{-bb}^{(M)} \\ P_{+bb}^{(M)} \end{pmatrix}. \end{aligned} \tag{6.2.14}$$

ii) **Coefficients**: Let $\boldsymbol{w}(n) = (\mathbf{w}_L^{(n+1)}, \mathbf{w}_U^{(n)})$, for $n = 1, ..., M+N-1$. After we obtain vector $\mathbf{p}^{(M)}$, the coefficients can be obtained by solving the following set of linear equations (note that matrices $T_+^{(M)}$ and $T_-^{(M)}$ are used to find $\boldsymbol{w}(M)$):

$$
\begin{aligned}
&\boldsymbol{w}(M) = \mathbf{p}^{(M)}(Q_{b+}^{(M)}, Q_{b-}^{(M)}) \left(I - \begin{pmatrix} T_+^{(M)}(P_{-b+}^{(M)}, P_{-b-}^{(M)}) \\ T_-^{(M)}(P_{+b+}^{(M)}, P_{+b-}^{(M)}) \end{pmatrix} \right)^{-1}; \\
&\boldsymbol{w}(1) = \boldsymbol{w}(1) \begin{pmatrix} \Psi_{+-}^{(l_2-l_1)}(P_{-b+}^{(1)}, P_{-b-}^{(1)}) \\ \widehat{\Psi}^{(1)}(P_{+b+}^{(1)}, P_{+b-}^{(1)}) \end{pmatrix} + \boldsymbol{w}(2) \begin{pmatrix} 0 \\ \widehat{\Lambda}_{--}^{(l_2-l_1)}(P_{-b+}^{(1)}, P_{-b-}^{(1)}) \end{pmatrix}; \\
&\boldsymbol{w}(n) = \boldsymbol{w}(n) \begin{pmatrix} \Psi_{+-}^{(l_{n+1}-l_n)}(P_{-b+}^{(n)}, P_{-b-}^{(n)}) \\ \widehat{\Psi}_{-+}^{(l_n-l_{n-1})}(P_{+b+}^{(n)}, P_{+b-}^{(n)}) \end{pmatrix} + \boldsymbol{w}(n+1) \begin{pmatrix} 0 \\ \widehat{\Lambda}_{--}^{(l_{n+1}-l_n)}(P_{-b+}^{(n)}, P_{-b-}^{(n)}) \end{pmatrix} \\
&\qquad + \boldsymbol{w}(n-1) \begin{pmatrix} \Lambda_{++}^{(l_n-l_{n-1})}(P_{+b+}^{(n)}, P_{+b-}^{(n)}) \\ 0 \end{pmatrix}, \\
&\qquad \text{for} \quad n = 2, ..., M-1, M+1, ..., M+N-2; \\
&\boldsymbol{w}(M+N-1) = \boldsymbol{w}(M+N-1) \begin{pmatrix} \Psi^{(M+N)}(P_{-b+}^{(M+N-1)}, P_{-b-}^{(M+N-1)}) \\ \widehat{\Psi}_{-+}^{(l_{M+N-1}-l_{M+N-2})}(P_{+b+}^{(M+N-1)}, P_{+b-}^{(M+N-1)}) \end{pmatrix} \\
&\qquad + \boldsymbol{w}(M+N-2) \begin{pmatrix} \Lambda_{++}^{(l_{M+N-1}-l_{M+N-2})}(P_{+b+}^{(M+N-1)}, P_{+b-}^{(M+N-1)}) \\ 0 \end{pmatrix}.
\end{aligned}
\tag{6.2.15}
$$

The joint stationary distributions of the age processes are similar to that in Chapter 5.

Let $\mathbf{f}(x)$ be the joint stationary 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 have the following result.

Theorem 6.2. *Under the conditions in Equation (6.2.10), the joint stationary distribution*

of the age process $\{(a(t), s(t), I_b(t),\ I_s(t)), t \geq 0\}$ *exists and its density function is*

$$\begin{aligned}
&P\{a(t)=0\} = \hat{\mathbf{p}}^{(M)}\mathbf{e};\\
&\mathbf{f}^{(n)}(x) = \left(\mathbf{v}_+^{(n)} e^{\mathcal{K}^{(n)}(x-l_{n-1})}\Psi^{(n)} + \mathbf{v}_-^{(n)} e^{\widehat{\mathcal{K}}^{(n)}(l_n-x)}\right),\ \textit{for } l_{n-1} < x \leq l_n,\ n=1,...,M;\\
&\mathbf{f}^{(n)}(x) = \left(\mathbf{v}_+^{(n)} e^{\mathcal{K}^{(n)}(x-l_{n-1})} + \mathbf{v}_-^{(n)} e^{\widehat{\mathcal{K}}^{(n)}(l_n-x)}\widehat{\Psi}^{(n)}\right), \textit{for } l_{n-1} \leq x < l_n,\ n=M+1,...,M+N.
\end{aligned} \tag{6.2.16}$$

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* $\mathbf{v}_+^{(1)} = 0$ *and* $\mathbf{v}_-^{(M+N)} = 0$. *The normalization factor is*

$$\begin{aligned}
\hat{c}_{norm} &= \mathbf{p}^{(M)}\mathbf{e} + \sum_{n=1}^{M}\left(\mathbf{u}_+^{(n)}\mathcal{L}_{l_{n-1},l_n}^{\mathcal{K}^{(n)}}\Psi^{(n)} + \mathbf{u}_-^{(n)}\widetilde{\mathcal{L}}_{l_{n-1},l_n}^{\widehat{\mathcal{K}}^{(n)}}\right)\mathbf{e}\\
&\quad + \sum_{n=M+1}^{M+N}\left(\mathbf{u}_+^{(n)}\mathcal{L}_{l_{n-1},l_n}^{\mathcal{K}^{(n)}} + \mathbf{u}_-^{(n)}\widetilde{\mathcal{L}}_{l_{n-1},l_n}^{\widehat{\mathcal{K}}^{(n)}}\widehat{\Psi}^{(n)}\right)\mathbf{e}.
\end{aligned} \tag{6.2.17}$$

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

Corollary 6.2.1. *The joint stationary distribution for the age process of buyers is*

$$\begin{aligned}
&P\{a_B(t)=0\} = \hat{\mathbf{p}}^{(M)}\mathbf{e} + \sum_{n=1}^{M}\left(\mathbf{v}_+^{(n)}\mathcal{L}_{l_{n-1},l_n}^{\mathcal{K}^{(n)}}\Psi^{(n)} + \mathbf{v}_-^{(n)}\widetilde{\mathcal{L}}_{l_{n-1},l_n}^{\widehat{\mathcal{K}}^{(n)}}\right)\mathbf{e};\\
&\mathbf{f}_B^{(n)}(x) = \mathbf{f}^{(n)}(x), \quad \textit{for } l_{n-1} \leq x < l_n,\ n=M+1,...,M+N.
\end{aligned} \tag{6.2.18}$$

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

$$\mathbf{f}_B^{(n)}(k,x) = \mathbf{f}_B^{(n)}(x)(\mathbf{e}(k)\otimes I), \tag{6.2.19}$$

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 $k = 1, ..., K$,

$$p_B(k) = \sum_{n=M+1}^{M+M} \left(\mathbf{v}_+^{(n)} \mathcal{L}_{l_{n-1},l_n}^{\mathcal{K}^{(n)}} + \mathbf{v}_-^{(n)} \widetilde{\mathcal{L}}_{l_{n-1},l_n}^{\widehat{\mathcal{K}}^{(n)}} \widehat{\Psi}^{(n)} \right) (\mathbf{e}(k) \otimes \mathbf{e}). \tag{6.2.20}$$

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

$$p_B = p_B(1) + p_B(2) + ... + p_B(K). \tag{6.2.21}$$

Next, we present the algorithm for the joint stationary distribution of the age process.

Algorithm 4: The joint stationary distribution of the age process

1. Input Parameters: M, N, K, $\{\tilde{l}_0 = 0, ..., \tilde{l}_N\}$, $\{\hat{l}_0 = 0, ..., \hat{l}_M\}$, $\{\dot{\eta}_{k,0}, ..., \dot{\eta}_{k,N}\}$ and $\{\eta_{k,0}, ..., \eta_{k,N}\}$, for $k = 1, 2, ..., K$, $\{\dot{\hat{\eta}}_{k,0}, ..., \dot{\hat{\eta}}_{k,M}\}$, $\{\hat{\eta}_{k,0}, ..., \hat{\eta}_{k,M}\}$, for $k = 1, 2, ..., K$, $\{m_b, D_0, ..., D_K\}$, and $\{m_s, \hat{D}_0, ..., \hat{D}_K\}$;

2. Construct transition blocks for the multi-layer $MMFF$ process:

   2.1 Borders: $\{l_0 = -\infty, ..., l_M, ..., l_{M+N} = \infty\}$ 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 $l_{M+N} = \infty$;

   2.2 Construct $\{Q^{(n)}, n = 1, 2, ..., M+N\}$ using Equations (6.2.4) and (6.2.5);

   2.3 Construct $\{Q_{bb}^{(M)}, Q_{b+}^{(M)}, Q_{b-}^{(M)}\}$ using Equation (6.2.6);

   2.4 Construct $\{P_{+b+}^{(n)}, P_{+b0}^{(n)}, P_{+b-}^{(n)}, P_{-b+}^{(n)}, P_{-b0}^{(n)}, P_{-b-}^{(n)}, n = 1, 2, ..., M+N-1\}$ using Equations (6.2.7), (6.2.8) and (6.2.9);

3. Use Algorithm 1 to compute $\{\Psi^{(n)}, \mathcal{K}^{(n)}, \mathcal{U}^{(n)}, \widehat{\Psi}^{(n)}, \widehat{\mathcal{K}}^{(n)}, \widehat{\mathcal{U}}^{(n)}\}$ for $n = 1, ..., M+N$ and $\{\Psi_{+-}^{(l_n - l_{n-1})}, \widehat{\Psi}_{-+}^{(l_n - l_{n-1})}, \Lambda_{++}^{(l_n - l_{n-1})}, \widehat{\Lambda}_{--}^{(l_n - l_{n-1})}\}$ for $n = 2, 3, ..., M+N-1$;

4. Compute $T_+^{(M)}$ and $T_-^{(M)}$ by Equations (5.3.2) and (5.3.3); Construct $Q_{\mathbf{p}}^{(M)}$ by Equation (6.2.14); and solve $\mathbf{p}^{(M)} Q_{\mathbf{p}}^{(M)} = 0$ and $\mathbf{p}^{(M)} \mathbf{e} = 1$ to get border probabilities;

5. Solving the set of linear equations (6.2.15) to get the coefficients $\{\boldsymbol{w}(n),\ n = 1, 2, ..., M+N-1\}$; and get the joint density function of the multi-layer $MMFF$ process in Theorem 6.2.10;

6. Compute $\hat{c}_{norm}$ by using Equation (6.2.17), and use $\hat{c}_{norm}$ to get $\hat{\mathbf{p}}^{(M)}$ and $\{\mathbf{v}_+^{(n)}, \mathbf{v}_-^{(n)},\ n = 1, 2, ..., M+N\}$;

7. Use $\hat{\mathbf{p}}^{(M)}$, $\{\mathbf{v}_+^{(n)}, \mathbf{v}_-^{(n)},\ n = 1, 2, ..., M+N\}$ and Equation (6.2.16) to compute the density function of the age process;

8. Compute the density function of the age process for buyers using Equation (6.2.18).

Using the joint density functions of the age processes, in the following sections, we derive a number of queueing quantities for buyer orders (Subsection 6.3.1) and for buyers (Subsection 6.3.2). To distinguish the two sets of quantities, we use notations with superscript "o" for order level queueing quantities in Subsection 6.3.1.

6.3 Queueing Quantities

In this section, we use the results in both Theorem 6.1, Theorem 6.2 and some basic quantities of the $MMFF$ processes to find queueing quantities for the double-sided queueing model. 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.

6.3.1 Order Level Queueing Quantities

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 6.1. *The matching rate of the orders is*

$$\begin{aligned}\omega^o &= \frac{1}{\hat{c}_{norm}}\sum_{n=1}^{M}\int_{l_{n-1}}^{l_n}\left(\boldsymbol{\pi}_+^{(n)}(x)\boldsymbol{\delta}(-,+,n)+\boldsymbol{\pi}_-^{(n)}(x)\boldsymbol{\delta}(-,-)\right)\mathrm{d}x \\ &\quad +\frac{1}{\hat{c}_{norm}}\sum_{n=M+1}^{M+N}\int_{l_{n-1}}^{l_n}\left(\boldsymbol{\pi}_+^{(n)}(x)\boldsymbol{\delta}(+,+)+\boldsymbol{\pi}_-^{(n)}(x)\boldsymbol{\delta}(+,-,n)\right)\mathrm{d}x,\end{aligned} \tag{6.3.1}$$

where

$$\begin{aligned}\int_{l_{n-1}}^{l_n}\boldsymbol{\pi}_+^{(n)}(x)\mathrm{d}x &= \mathbf{u}_+^{(n)}\mathcal{L}_{l_{n-1},l_n}^{\mathcal{K}^{(n)}}+\mathbf{u}_-^{(n)}\widetilde{\mathcal{L}}_{l_{n-1},l_n}^{\widehat{\mathcal{K}}^{(n)}}\widehat{\Psi}^{(n)}, \\ \int_{l_{n-1}}^{l_n}\boldsymbol{\pi}_-^{(n)}(x)\mathrm{d}x &= \mathbf{u}_+^{(n)}\mathcal{L}_{l_{n-1},l_n}^{\mathcal{K}^{(n)}}\Psi^{(n)}+\mathbf{u}_-^{(n)}\widetilde{\mathcal{L}}_{l_{n-1},l_n}^{\widehat{\mathcal{K}}^{(n)}},\end{aligned} \tag{6.3.2}$$

and

$$\begin{aligned}
\boldsymbol{\delta}(-,-) &= \left(\left(A \otimes \mathbf{e}\mathbf{e}^T\right) \odot \left(\mathbf{e} \otimes (D_1, D_2, ..., D_K) \otimes I\right)\right) \mathbf{e}, \\
\boldsymbol{\delta}(+,-,n) &= \left(\left(\hat{A} \otimes \mathbf{e}\mathbf{e}^T\right) \odot \left(\mathbf{e} \otimes (\mathcal{D}_{1,n}, \mathcal{D}_{2,n}, ..., \mathcal{D}_{K,n}) \otimes I\right)\right) \mathbf{e}, \\
\boldsymbol{\delta}(-,+,n) &= \left(\left(\tilde{\hat{A}} \otimes \mathbf{e}\mathbf{e}^T\right) \odot \left(\mathbf{e} \otimes (I \otimes \hat{\mathcal{D}}_{1,n}, I \otimes \hat{\mathcal{D}}_{2,n}, ..., I \otimes \hat{\mathcal{D}}_{K,n})\right)\right) \mathbf{e}, \\
\boldsymbol{\delta}(+,+) &= \left(\left(\tilde{A} \otimes \mathbf{e}\mathbf{e}^T\right) \odot \left(\mathbf{e} \otimes (I \otimes \hat{D}_1, I \otimes \hat{D}_2, ..., I \otimes \hat{D}_K)\right)\right) \mathbf{e},
\end{aligned} \tag{6.3.3}$$

where $\odot$ is the Hadamard product and $\otimes$ is the Kronecker product for matrices, $A = (a_{i,j})$ 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 $\hat{A}$.

Proof. At an arbitrary time, 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

$$\delta(x, s, i_b, i_s) = \sum_{k=1}^{K} \min\{s, k\} (\hat{D}_k \mathbf{e})_{i_s}. \tag{6.3.4}$$

Let $\boldsymbol{\delta}(+,+)$ be the vector with elements $\delta(x, s, i_b, i_s)$ defined in Equation (6.3.4). 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 (6.3.2) and the last line in Equation (6.3.3), 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 that the buyer has 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 $l_{n-1} < x < l_n$,

$$\delta(x, n, s, i_b, i_s) = \sum_{k=1}^{K} \min\{-s, k\}(\mathcal{D}_{k,n}\mathbf{e})_{i_b}. \tag{6.3.5}$$

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 (6.3.2) and the second line in Equation (6.3.3), 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 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

$$\delta(x, s, i_b, i_s) = \sum_{k=1}^{K} \min\{-s, k\}(D_k\mathbf{e})_{i_b}. \tag{6.3.6}$$

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 (6.3.2) and the first line in Equation (6.3.3), 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 $l_{n-1} < x < l_n$,

$$\delta(x, n, s, i_b, i_s) = \sum_{k=1}^{K} \min\{s, k\}(\hat{\mathcal{D}}_{k,n}\mathbf{e})_{i_s}. \tag{6.3.7}$$

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 (6.3.2) and the third line in Equation (6.3.3), 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 6.2.2. *We have*

$$p^o_{B,F} = \frac{\omega^o}{\lambda}. \tag{6.3.8}$$

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 6.2.

$$p^o_{BL,1} = \frac{1}{\lambda}\sum_{n=M+1}^{M+N-1}\mathbf{f}^{(n)}(l_n)\left(\left(\sum_{k=1}^{K}\frac{k\dot{\eta}_{k,n-M}}{\sum_{m=n}^{M+N}\dot{\eta}_{k,m-M}}\mathbf{e}(K-k+1)\right)\otimes\mathbf{e}\right) + \frac{1}{\lambda}\hat{\mathbf{p}}^{(M)}\left(\left(\sum_{k=1}^{K}k\dot{\eta}_{k,0}D_k\mathbf{e}\right)\otimes\mathbf{e}\right), \tag{6.3.9}$$

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* $(K-k+1)$*-st 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 (6.3.9) (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 combine 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 (6.3.9). □

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

$$\begin{aligned} P\{W^o_{B,F}=0\} &= \frac{1}{p_{B,F}\lambda\hat{c}_{norm}}\left(\sum_{n=1}^{M}\int_{l_{n-1}}^{l_n}\boldsymbol{\pi}^{(n)}_{-}(x)\mathrm{d}x\boldsymbol{\delta}(-,-)+\int_{l_{n-1}}^{l_n}\boldsymbol{\pi}^{(n)}_{+}(x)\mathrm{d}x\boldsymbol{\delta}(-,+,n)\right);\\ \frac{\mathrm{d}P\{W^o_{B,F}<x\}}{\mathrm{d}x} &= \frac{1}{p_{B,F}\lambda\hat{c}_{norm}}\left(\boldsymbol{\pi}^{(n)}_{+}(x)\boldsymbol{\delta}(+,+)+\boldsymbol{\pi}^{(n)}_{-}(x)\boldsymbol{\delta}(+,-,n)\right),\\ &\qquad \text{for } l_{n-1}\le x<l_n,\ n=M+1,...,M+N, \end{aligned} \tag{6.3.10}$$

where the integrals are given in Equation (6.3.2). The distribution of sojourn time $W^o_{BL,1}$ of lost orders at the head of the queue is given by

$$P\{W^o_{BL,1}=0\}=\frac{1}{p^o_{BL,1}\lambda}\hat{\mathbf{p}}^{(M)}\left(\left(\sum_{k=1}^{K}k\dot{\eta}_{k,0}D_k\mathbf{e}\right)\otimes\mathbf{e}\right), \tag{6.3.11}$$

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

$$\begin{aligned} P\{W^o_{BL,1}=l_n\} &= \frac{\left(\mathbf{v}^{(n)}_{+}e^{\mathcal{K}^{(n)}(l_n-l_{n-1})}+\mathbf{v}^{(n)}_{-}\widehat{\Psi}^{(n)}\right)}{p^o_{BL,1}\lambda}\\ &\quad\times\left(\left(\sum_{k=1}^{K}\frac{k\dot{\eta}_{k,n-M}}{\dot{\eta}_{k,n-M}+...+\dot{\eta}_{k,N}}\mathbf{e}(K-k+1)\right)\otimes\mathbf{e}\right). \end{aligned} \tag{6.3.12}$$

The distribution of sojourn time $W^o_{PL,>1}$ of lost orders before reaching the head of the queue, we have, for $n=M,M+1,...,M+N-1$,

$$P\{W^o_{BL,>1}=l_n\}=\frac{1}{p^o_{BL,>1}\lambda\hat{c}_{norm}}\left(\sum_{m=n+1}^{M+N}\int_{l_{n-1}}^{l_n}\boldsymbol{\pi}^{(n)}_{-}(x)\mathrm{d}x\left(\mathbf{e}\otimes\left(\sum_{k=1}^{K}k\eta_{k,n-M}D_k\mathbf{e}\right)\otimes\mathbf{e}\right)\right). \tag{6.3.13}$$

Proof. Since matching can happen during the fictitious time periods in the multi-layer $MMFF$ processes, we need to use the joint stationary 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 6.1) 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 6.1, 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 like Proposition 6.2. The distribution of $W^o_{BL,1}$ can be obtained from the proof of Proposition 6.2 easily.

For the distribution of $W^o_{BL,>1}$, we need to use the joint stationary 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:

$$\begin{aligned}\mathbb{E}[W^o_{B,F}] \;=\; \frac{1}{p_{B,F}\lambda}\Bigg(&\sum_{n=M+1}^{M+N} \left(\mathbf{v}_+^{(n)} \mathcal{M}^{\mathcal{K}^{(n)}}_{l_{n-1},l_n} + \mathbf{v}_-^{(n)} \widetilde{\mathcal{M}}^{\widehat{\mathcal{K}}^{(n)}}_{l_{n-1},l_n} \widehat{\Psi}^{(n)}\right) \boldsymbol{\delta}(+,+) \\ &+ \left(\mathbf{v}_+^{(n)} \mathcal{M}^{\mathcal{K}^{(n)}}_{l_{n-1},l_n} \Psi^{(n)} + \mathbf{v}_-^{(n)} \widetilde{\mathcal{M}}^{\widehat{\mathcal{K}}^{(n)}}_{l_{n-1},l_n}\right) \boldsymbol{\delta}(+,-,n)\Bigg),\end{aligned} \tag{6.3.14}$$

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 Lemma B.1. The mean sojourn time of an arbitrary buyer order can be obtained by

$$\mathbb{E}[W_B^o] = p_{B,F}^o \mathbb{E}[W_{B,F}^o] + p_{BL,1}^o \mathbb{E}[W_{BL,1}^o] + p_{BL,>1}^o \mathbb{E}[W_{BL,>1}^o]. \tag{6.3.15}$$

Queue length of orders

Let $q_B^o(t)$ be the queue length of buyer orders at an arbitrary time t. 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

$$\Delta(n) := (D_0(n), D_1(n), ..., D_K(n)), \tag{6.3.16}$$

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

In general, by Theorem 2.4.1 in [62] or Lemma B.2, for buyers arrived in $(t-l_n, t-l_{n-1})$, the probability generating function for the batch size is given by

$$P^*(n,z,x) = \exp\left\{ \left(D_0(n) + \sum_{k=1}^K z^k D_k(n) \right) x \right\}. \tag{6.3.17}$$

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

$$\begin{aligned} \mathbb{E}[z^{q_B^o(t)}] &= P\{a(t) \le 0\} + \sum_{n=M+1}^{M+N} \int_{l_{n-1}}^{l_n} \mathbf{f}_B^{(n)}(x) \\ &\quad \times \left(I(z^K) \otimes \left(P^*(n,z,x-l_{n-1}) \prod_{m=n-1}^{M+1} P^*(m,z,b_m) \right) \otimes I \right) \mathrm{d}x \mathbf{e}, \end{aligned} \tag{6.3.18}$$

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 [62] or Lemma B.2, we have

$$\frac{\partial P^*(n,z,x)\mathbf{e}}{\partial z}\bigg|_{z=1} = \left(\sum_{k=1}^{K} k\xi_{k,n}\lambda_k\right) x\mathbf{e} + (e^{Dx} - I)(D - \mathbf{e}\boldsymbol{\theta}_b)^{-1}\left(\sum_{k=1}^{K} k\xi_{k,n}D_k\right)\mathbf{e}. \quad (6.3.19)$$

Consequently, we obtain

Proposition 6.4. *The mean queue length of remaining buyer orders is given by*

$$\begin{aligned}
\mathbb{E}[q_B^o(t)] = \frac{\partial \mathbb{E}[z^{q_B^o(t)}]}{\partial z}\bigg|_{z=1} &= \sum_{n=M+1}^{M+N}\left(\mathbf{v}_+^{(n)}\mathcal{L}_{l_{n-1},l_n}^{\mathcal{K}^{(n)}} + \mathbf{v}_-^{(n)}\widetilde{\mathcal{L}}_{l_{n-1},l_n}^{\widehat{\mathcal{K}}^{(n)}}\widehat{\Psi}^{(n)}\right)(I(K)\otimes I\otimes I)\mathbf{e} \\
&+ \sum_{n=M+1}^{M+N}\sum_{m=M+1}^{n-1}\int_{l_{n-1}}^{l_n}\mathbf{f}_B^{(n)}(x)\left(I\otimes e^{D(x-l_m)}\otimes I\right)\mathrm{d}x \\
&\quad\times\left(I\otimes\left(\left(\sum_{k=1}^{K}k\xi_{k,m}\lambda_k\right)b_m I(e^{Db_m}-I)(D-\mathbf{e}\boldsymbol{\theta}_b)^{-1}\left(\sum_{k=1}^{K}k\xi_{k,m}D_k\right)\right)\otimes I\right)\mathbf{e} \\
&+ \sum_{n=M+1}^{M+N}\int_{l_{n-1}}^{l_n}\mathbf{f}_B^{(n)}(x)\left(I\otimes\left(\left(\sum_{k=1}^{K}k\xi_{k,n}\lambda_k\right)(x-l_{n-1})I\right.\right. \\
&\qquad\left.\left.+\,(e^{D(x-l_{n-1})}-I)(D-\mathbf{e}\boldsymbol{\theta}_b)^{-1}\left(\sum_{k=1}^{K}k\xi_{k,n}D_k\right)\right)\otimes I\right)\mathbf{e}\mathrm{d}x,
\end{aligned} \quad (6.3.20)$$

where $I(K) = diag(K, K-1, ..., 1)$. The integrals in the above equation can be evaluated using closed form expressions in Lemma B.1.

The mean queue length (i.e., total batch size) of buyer orders can also be obtained by Little's law as $\mathbb{E}[q_B^o(t)] = \lambda\mathbb{E}[W_B^o]$, which can be used to check the computation accuracy.

6.3.2 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.

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 the partial matching.

Proposition 6.5. *The matching rate of the buyers of the system is*

$$\begin{aligned}\omega_B &= \frac{1}{\hat{c}_{norm}}\sum_{n=1}^{M}\int_{l_{n-1}}^{l_n}\left(\boldsymbol{\pi}_+^{(n)}(x)\hat{\boldsymbol{\delta}}(-,+,n)+\boldsymbol{\pi}_-^{(n)}(x)\hat{\boldsymbol{\delta}}(-,-)\right)\mathrm{d}x\\ &+\frac{1}{\hat{c}_{norm}}\sum_{n=M+1}^{M+N}\int_{l_{n-1}}^{l_n}\left(\boldsymbol{\pi}_+^{(n)}(x)\hat{\boldsymbol{\delta}}(+,+)+\boldsymbol{\pi}_-^{(n)}(x)\hat{\boldsymbol{\delta}}(+,-,n)\right)\mathrm{d}x,\end{aligned} \tag{6.3.21}$$

where

$$\begin{aligned}\hat{\boldsymbol{\delta}}(-,-) &= \left(\left(E\otimes\mathbf{e}\mathbf{e}^T\right)\odot\left(\mathbf{e}\otimes(D_1,D_2,...,D_K)\otimes I\right)\right)\mathbf{e},\\ \hat{\boldsymbol{\delta}}(+,-,n) &= \left(\left(\hat{E}\otimes\mathbf{e}\mathbf{e}^T\right)\odot\left(\mathbf{e}\otimes(\mathcal{D}_{1,n},\mathcal{D}_{2,n},...,\mathcal{D}_{K,n})\otimes I\right)\right)\mathbf{e},\\ \hat{\boldsymbol{\delta}}(-,+,n) &= \left(\left(F\otimes\mathbf{e}\mathbf{e}^T\right)\odot\left(\mathbf{e}\otimes(I\otimes\hat{\mathcal{D}}_{1,n},I\otimes\hat{\mathcal{D}}_{2,n},...,I\otimes\hat{\mathcal{D}}_{K,n})\right)\right)\mathbf{e},\\ \hat{\boldsymbol{\delta}}(+,+) &= \left(\left(\hat{F}\otimes\mathbf{e}\mathbf{e}^T\right)\odot\left(\mathbf{e}\otimes(I\otimes\hat{D}_1,I\otimes\hat{D}_2,...,I\otimes\hat{D}_K)\right)\right)\mathbf{e},\end{aligned} \tag{6.3.22}$$

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 6.1 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 6.1, 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 6.2.3.

$$p_{B,F} = \frac{\omega_B}{\sum_{k=1}^{K} \lambda_k}. \tag{6.3.23}$$

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

Proposition 6.6.

$$\begin{aligned} p_{BL,1} = & \frac{1}{\sum_{k=1}^{K} \lambda_k} \sum_{n=M+1}^{M+N-1} \left(\mathbf{v}_+^{(n)} e^{\mathcal{K}^{(n)}(l_n - l_{n-1})} + \mathbf{v}_-^{(n)} \widehat{\Psi}^{(n)} \right) \left(\left(\sum_{k=1}^{K} \frac{\dot{\eta}_{k,n-M}}{\sum_{m=n}^{M+N} \dot{\eta}_{k,m-M}} \mathbf{e}(K-k+1) \right) \otimes \mathbf{e} \right) \\ & + \frac{1}{\sum_{k=1}^{K} \lambda_k} \hat{\mathbf{p}}^{(M)} \left(\left(\sum_{k=1}^{K} \dot{\eta}_{k,0} D_k \mathbf{e} \right) \otimes \mathbf{e} \right), \end{aligned} \tag{6.3.24}$$

and $p_{BL,>1} = p_{B,L} - p_{BL,1}$.

Proof. Similar to Proposition 6.2. □

Sojourn times of buyers-sellers

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

Proposition 6.7. *The distribution of sojourn time $W_{B,F}$ of fully filled buyers is*

$$P\{W_{B,F}=0\}=\frac{1}{p_{B,F}\sum_{k=1}^{K}\lambda_k\hat{c}_{norm}}\left(\sum_{n=1}^{M}\int_{l_{n-1}}^{l_n}\boldsymbol{\pi}_{-}^{(n)}(x)\mathrm{d}x\hat{\boldsymbol{\delta}}(-,-)+\int_{l_{n-1}}^{l_n}\boldsymbol{\pi}_{+}^{(n)}(x)\mathrm{d}x\hat{\boldsymbol{\delta}}(-,+,n)\right);$$

$$\frac{\mathrm{d}P\{W_{B,F}<x\}}{\mathrm{d}x}=\frac{1}{p_{B,F}\sum_{k=1}^{K}\lambda_k\hat{c}_{norm}}\left(\boldsymbol{\pi}_{+}^{(n)}(x)\hat{\boldsymbol{\delta}}(+,+)+\boldsymbol{\pi}_{-}^{(n)}(x)\hat{\boldsymbol{\delta}}(+,-,n)\right),$$

$$\text{for } l_{n-1}\le x<l_n,\ n=M+1,...,M+N. \tag{6.3.25}$$

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

$$P\{W_{BL,1}=0\}=\frac{1}{p_{BL,1}\sum_{k=1}^{K}\lambda_k}\hat{\mathbf{p}}^{(M)}\left(\left(\sum_{k=1}^{K}\dot{\eta}_{k,0}D_k\mathbf{e}\right)\otimes\mathbf{e}\right), \tag{6.3.26}$$

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

$$P\{W_{BL,1}=l_n\}=\frac{\left(\mathbf{v}_{+}^{(n)}e^{\mathcal{K}^{(n)}(l_n-l_{n-1})}+\mathbf{v}_{-}^{(n)}\widehat{\Psi}^{(n)}\right)}{p_{BL,1}\sum_{k=1}^{K}\lambda_k}\left(\left(\sum_{k=1}^{K}\frac{\dot{\eta}_{k,n-M}}{\dot{\eta}_{k,n-M}+...+\dot{\eta}_{k,N}}\mathbf{e}(K-k+1)\right)\otimes\mathbf{e}\right). \tag{6.3.27}$$

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

$$P\{W_{BL,>1}=l_n\}=\frac{\left(\sum_{m=n+1}^{M+N}\left(\mathbf{v}_{+}^{(m)}\mathcal{L}_{l_{m-1},l_m}^{\mathcal{K}^{(m)}}\Psi^{(m)}+\mathbf{v}_{-}^{(m)}\widetilde{\mathcal{L}}_{l_{m-1},l_m}^{\widehat{\mathcal{K}}^{(m)}}\right)\left(\mathbf{e}\otimes\left(\sum_{k=1}^{K}\eta_{k,n-M}D_k\mathbf{e}\right)\otimes\mathbf{e}\right)\right)}{p_{BL,>1}\sum_{k=1}^{K}\lambda_k}. \tag{6.3.28}$$

Proof. Similar to Proposition 6.3. □

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

$$\begin{aligned}\mathbb{E}[W_{B,F}] &= \frac{1}{p_{B,F}\sum_{k=1}^{K}\lambda_k}\Bigg(\sum_{n=M+1}^{M+N}\left(\mathbf{u}_+^{(n)}\mathcal{M}_{l_{n-1},l_n}^{\mathcal{K}^{(n)}}+\mathbf{u}_-^{(n)}\widetilde{\mathcal{M}}_{l_{n-1},l_n}^{\widehat{\mathcal{K}}^{(n)}}\widehat{\Psi}^{(n)}\right)\hat{\boldsymbol{\delta}}(+,+)\\ &\qquad+\left(\mathbf{u}_+^{(n)}\mathcal{M}_{l_{n-1},l_n}^{\mathcal{K}^{(n)}}\Psi^{(n)}+\mathbf{u}_-^{(n)}\widetilde{\mathcal{M}}_{l_{n-1},l_n}^{\widehat{\mathcal{K}}^{(n)}}\right)\hat{\boldsymbol{\delta}}(+,-,n)\Bigg).\end{aligned} \tag{6.3.29}$$

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

$$\mathbb{E}[W_B]=p_{B,F}\mathbb{E}[W_{B,F}]+p_{BL,1}\mathbb{E}[W_{BL,1}]+p_{BL,>1}\mathbb{E}[W_{BL,>1}]. \tag{6.3.30}$$

Queue lengths of buyers-sellers

Proposition 6.8. *The mean queue length of buyers is given by*

$$\begin{aligned}\mathbb{E}[q_B(t)] &= \sum_{n=M+1}^{M+N}\left(\mathbf{v}_+^{(n)}\mathcal{L}_{l_{n-1},l_n}^{\mathcal{K}^{(n)}}+\mathbf{v}_-^{(n)}\widetilde{\mathcal{L}}_{l_{n-1},l_n}^{\widehat{\mathcal{K}}^{(n)}}\widehat{\Psi}^{(n)}\right)\mathbf{e}\\ &+\sum_{n=M+1}^{M+N}\sum_{m=M+1}^{n-1}\int_{l_{n-1}}^{l_n}\mathbf{f}_B^{(n)}(x)\left(I\otimes e^{D(x-l_m)}\otimes I\right)\mathrm{d}x\\ &\qquad\times\left(I\otimes\left(\left(\sum_{k=1}^{K}\xi_{k,m}\lambda_k\right)b_mI+(e^{Db_m}-I)(D-\mathbf{e}\boldsymbol{\theta}_b)^{-1}\left(\sum_{k=1}^{K}\xi_{k,m}D_k\right)\right)\otimes I\right)\mathbf{e}\\ &+\sum_{n=M+1}^{M+N}\int_{l_{n-1}}^{l_n}\mathbf{f}_B^{(n)}(x)\Bigg(I\otimes\Bigg(\left(\sum_{k=1}^{K}\xi_{k,n}\lambda_k\right)(x-l_{n-1})I\\ &\qquad+(e^{D(x-l_{n-1})}-I)(D-\mathbf{e}\boldsymbol{\theta}_b)^{-1}\left(\sum_{k=1}^{K}\xi_{k,n}D_k\right)\Bigg)\otimes I\Bigg)\mathbf{e}\mathrm{d}x.\end{aligned} \tag{6.3.31}$$

Note that the expression of the queue length of buyers in this queueing model is the same as the one in Chapter 5 Proposition 5.4.

6.3.3 Summary of Queueing Quantities

We summarize all important queueing quantities in Table 6.2 to help readers quickly find the meaning and equations of these quantities.

	Buyer-Seller level quantities	Order level quantities
Density of the age process	$\mathbf{f}^{(n)}(x)$ (6.2.16)	$\mathbf{f}_B^{(n)}(k,x)$ (6.2.19)
Type of the queue	p_B (6.2.21)	$p_B(k)$ (6.2.20)
Matching rate	ω_B (6.3.21)	ω^o (6.3.1)
Abandonment probabilities	$P_{B,F}$, $P_{B,L}$ (6.3.23), $P_{BL,>1}$, $P_{BL,1}$ (6.3.24)	$P^o_{B,F}$, $P^o_{B,L}$ (6.3.8), $P^o_{BL,1}$ & $P^o_{BL,>1}$ (6.3.9)
Sojourn time of filled buyers	$W_{B,F}$ (6.3.25), $\mathbb{E}[W_{B,F}]$ (6.3.29)	$W^o_{B,F}$ (6.3.10), $\mathbb{E}[W^o_{B,F}]$ (6.3.14)
Sojourn time of lost buyers	$W_{BL,1}$ (6.3.27), $W_{BL,>1}$ (6.3.28)	$W^o_{BL,1}$ (6.3.12), $W^o_{BL,>1}$ (6.3.13)
Mean sojourn time	$\mathbb{E}[W_B]$ (6.3.30)	$\mathbb{E}[W^o_B]$ (6.3.15)
Queue lengths	$\mathbb{E}[q_B(t)]$(6.3.31)	$\mathbb{E}[q^o_B(t)]$ (6.3.20)

Table 6.2: Summary of queueing quantities in Chapter 6

6.3.4 Numerical Example

Example 6.1. (continued) We use the parameters provided in Example 6.1 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 6.3 and Figure 6.3, 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	$p^o_{B,F}$	$p^o_{BL,1}$	$p^o_{BL,>1}$	$\mathbb{E}[W^o_{B,F}]$	$\mathbb{E}[W^o_{B,L}]$	$\mathbb{E}[W^o_B]$	$\mathbb{E}[q^o_B(t)]$	$P\{W^o_{B,F}=0\}$
	8.1017	0.9778	0.0006	0.0216	0.2893	1.1253	0.3079	2.5510	0.7131
Buyer	ω_B	$p_{B,F}$	$p_{BL,1}$	$p_{BL,>1}$	$\mathbb{E}[W_{B,F}]$	$\mathbb{E}[W_{B,L}]$	$\mathbb{E}[W_B]$	$\mathbb{E}[q_B(t)]$	$P\{W_{B,F}=0\}$
	4.7720	0.9825	0.0004	0.0172	0.3030	1.1227	0.3174	1.5417	0.7056
Seller order	ω^o	$p^o_{S,F}$	$p^o_{SL,1}$	$p^o_{SL,>1}$	$\mathbb{E}[W^o_{S,F}]$	$\mathbb{E}[W^o_{S,L}]$	$\mathbb{E}[W^o_S]$	$\mathbb{E}[q^o_S(t)]$	$P\{W^o_{S,F}=0\}$
	8.1017	0.9002	0.0000	0.0998	0.9456	2.0069	1.0515	9.4635	0.2869
Seller	ω_S	$p_{S,F}$	$p_{SL,1}$	$p_{SL,>1}$	$\mathbb{E}[W_{S,F}]$	$\mathbb{E}[W_{S,L}]$	$\mathbb{E}[W_S]$	$\mathbb{E}[q_S(t)]$	$P\{W_{S,F}=0\}$
	5.2139	0.9201	0.0000	0.0799	0.9964	2.0137	1.0777	6.1067	0.2759

Table 6.3: Queueing quantities for Example 6.1

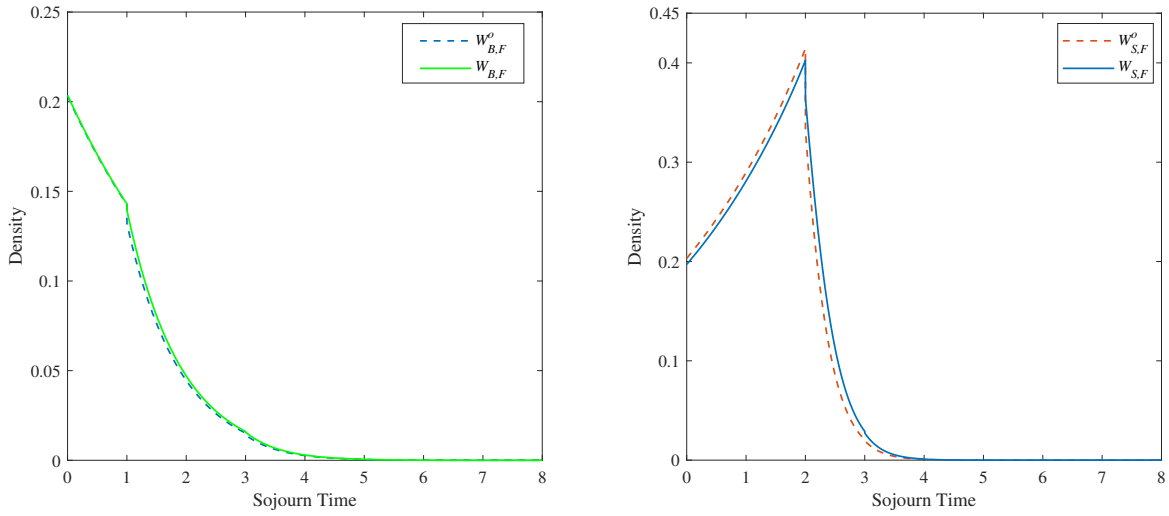

Figure 6.3: The stationary density functions of W^o_{BF},W_{BF},W^o_{SF} and W_{SF} for Example 6.1

Example 6.2. A double-sided queue with maximum batch size $K = 4$, and $M = N = 5$. All the input parameters are presented in Table 6.4. We assume that the batch size of buyers can only be 4, while the batch size of sellers can be 1 or 2. The batch size of a buyer can be less than 4 when the buyer is at the head of the queue due to partial matching, so we still define the abandonment time distributions for buyers at the head of the queue with less than 4 orders.

The probability that there is a buyer queue is $p_B = 0.2222$ and the probability that there is a seller queue is $p_S = 0.7748$. We can find the queueing quantities for both order

level and buyer-seller level in Table 6.5 and Figure 6.4.

Buyer Arrival	Batch Size	$\tilde{l}_1 = 5$	$\tilde{l}_2 = 10$	$\tilde{l}_3 = 15$	$\tilde{l}_4 = 20$	$\tilde{l}_5 = \infty$
$D_0 = \begin{pmatrix} -7, & 2 \\ 3, & -5 \end{pmatrix}$	$D_1 = \begin{pmatrix} 0, & 0 \\ 0, & 0 \end{pmatrix}$	$\dot{\eta}_{1,1} = 0$	$\dot{\eta}_{1,2} = 0$	$\dot{\eta}_{1,3} = 0$	$\dot{\eta}_{1,4} = 0.1$	$\dot{\eta}_{1,5} = 0.9$
	$D_2 = \begin{pmatrix} 0, & 0 \\ 0, & 0 \end{pmatrix}$	$\dot{\eta}_{2,1} = 0$	$\dot{\eta}_{2,2} = 0$	$\dot{\eta}_{2,3} = 0.1$	$\dot{\eta}_{2,4} = 0.1$	$\dot{\eta}_{2,5} = 0.8$
	$D_3 = \begin{pmatrix} 0, & 0 \\ 0, & 0 \end{pmatrix}$	$\dot{\eta}_{3,1} = 0$	$\dot{\eta}_{3,2} = 0$	$\dot{\eta}_{3,3} = 0$	$\dot{\eta}_{3,4} = 0.1$	$\dot{\eta}_{3,5} = 0.9$
	$D_4 = \begin{pmatrix} 5, & 0 \\ 1, & 1 \end{pmatrix}$	$\eta_{4,1} = 0.1$ $\dot{\eta}_{4,1} = 0.1$	$\eta_{4,2} = 0.1$ $\dot{\eta}_{4,2} = 0.1$	$\eta_{4,3} = 0.1$ $\dot{\eta}_{4,3} = 0.1$	$\eta_{4,4} = 0.1$ $\dot{\eta}_{4,4} = 0.1$	$\eta_{4,5} = 0.6$ $\dot{\eta}_{4,5} = 0.6$
Seller Arrival	Batch Size	$\hat{l}_1 = 5$	$\hat{l}_2 = 10$	$\hat{l}_3 = 15$	$\hat{l}_4 = 20$	$\hat{l}_5 = \infty$
$\hat{D}_0 = \begin{pmatrix} -50, & 6 \\ 1, & -5 \end{pmatrix}$	$\hat{D}_1 = \begin{pmatrix} 25, & 2 \\ 1, & 1 \end{pmatrix}$	$\hat{\eta}_{1,1} = 0.2$ $\dot{\hat{\eta}}_{1,1} = 0$	$\hat{\eta}_{1,2} = 0.2$ $\dot{\hat{\eta}}_{1,2} = 0$	$\hat{\eta}_{1,3} = 0.2$ $\dot{\hat{\eta}}_{1,3} = 0$	$\hat{\eta}_{1,4} = 0.2$ $\dot{\hat{\eta}}_{1,4} = 0.2$	$\hat{\eta}_{1,5} = 0.2$ $\dot{\hat{\eta}}_{1,5} = 0.8$
	$\hat{D}_2 = \begin{pmatrix} 17, & 0 \\ 0, & 2 \end{pmatrix}$	$\hat{\eta}_{2,1} = 0.1$ $\dot{\hat{\eta}}_{2,1} = 0$	$\hat{\eta}_{2,2} = 0$ $\dot{\hat{\eta}}_{2,2} = 0$	$\hat{\eta}_{2,3} = 0$ $\dot{\hat{\eta}}_{2,3} = 0$	$\hat{\eta}_{2,4} = 0$ $\dot{\hat{\eta}}_{2,4} = 0.1$	$\hat{\eta}_{2,5} = 0.9$ $\dot{\hat{\eta}}_{2,5} = 0.9$

Table 6.4: Parameters of Example 6.2

Buyer order	ω^o	$p^o_{B,F}$	$p^o_{BL,1}$	$p^o_{BL,>1}$	$\mathbb{E}[W^o_{B,F}]$	$\mathbb{E}[W^o_{B,L}]$	$\mathbb{E}[W^o_B]$	$\mathbb{E}[q^o_B(t)]$	$P\{W^o_{B,F} = 0\}$
	15.9111	0.9944	0.0000	0.0056	0.7343	5.3778	0.7601	12.1611	0.7686
Buyer	ω_B	$p_{B,F}$	$p_{BL,1}$	$p_{BL,>1}$	$\mathbb{E}[W_{B,F}]$	$\mathbb{E}[W_{B,L}]$	$\mathbb{E}[W_B]$	$\mathbb{E}[q_B(t)]$	$P\{W_{B,F} = 0\}$
	3.9778	0.9944	0.0000	0.0056	0.7554	5.3783	0.7811	3.1245	0.7639
Seller order	ω^o	$p^o_{S,F}$	$p^o_{SL,1}$	$p^o_{SL,>1}$	$\mathbb{E}[W^o_{S,F}]$	$\mathbb{E}[W^o_{S,L}]$	$\mathbb{E}[W^o_S]$	$\mathbb{E}[q^o_S(t)]$	$P\{W^o_{S,F} = 0\}$
	15.9111	0.9359	0.0000	0.0641	4.0606	5.5133	4.1537	70.6127	0.2126
Seller	ω_S	$p_{S,F}$	$p_{SL,1}$	$p_{SL,>1}$	$\mathbb{E}[W_{S,F}]$	$\mathbb{E}[W_{S,L}]$	$\mathbb{E}[W_S]$	$\mathbb{E}[q_S(t)]$	$P\{W_{S,F} = 0\}$
	11.1149	0.9262	0.0000	0.0738	4.0191	5.6314	4.1380	49.6564	0.2120

Table 6.5: Queueing quantities for Example 6.2

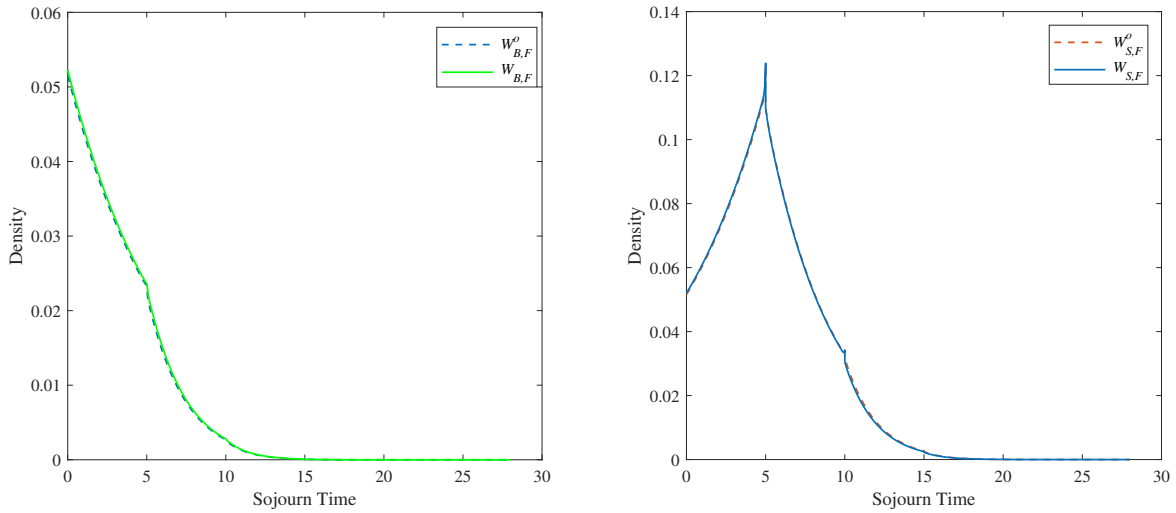

Figure 6.4: The stationary density functions of W^o_{BF},W_{BF},W^o_{SF} and W_{SF} for Example 6.2

From these two examples, we can observe 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 partial matching. In addition, we ran a large number of numerical examples and found that the density functions of W^o_{BF} and W_{BF} are quite close, as are the density functions of W^o_{SF} and W_{SF}.

6.4 Application of the Model to Vaccine Inventory Systems

In this section, we illustrate a potential application of the proposed queueing model to analyze the performance of vaccination services and inventory systems in a hypothetical setting, which may be close to reality for particular vaccination practices. For this application, as illustrated in Figure 6.5, we consider that vaccines stochastically arrive 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 expected to be administrated to patients randomly arriving at 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 a 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 the vaccine clinic complies with our definition of the double-sided queue with $BMAP$.

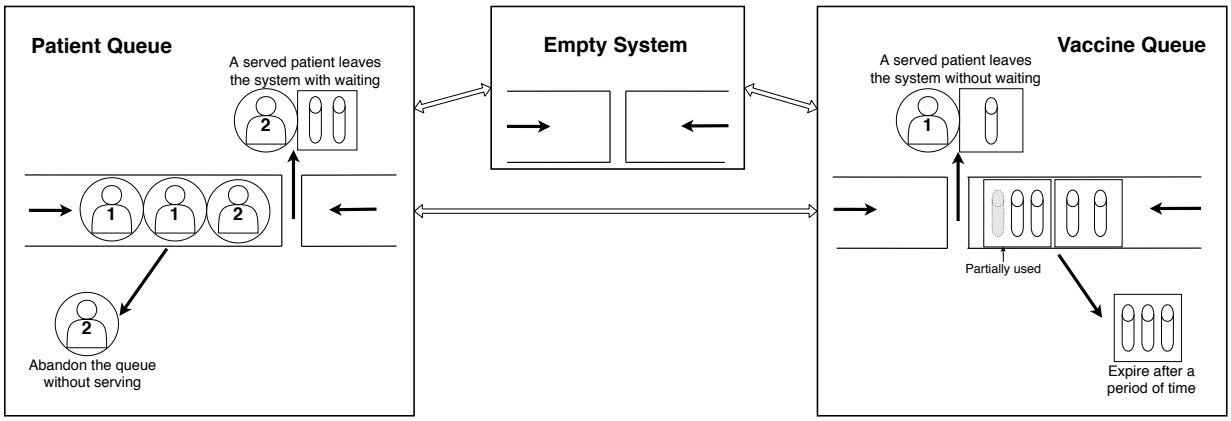

Figure 6.5: Diagram of the vaccine inventory system

This 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 [19, 119]. 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. Similarly, vaccine arrivals refer to randomly timed replenishments in batches, while abandonment refers to the 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 expires 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, 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) [94]. Therefore, the proposed model may represent (to some extend) the vaccination performances at least in those two cases.

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 is 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, and it can be expressed as a $BMAP$ as follows,

$$D_0 = -\hat{\lambda} + p^y\hat{\lambda}; \quad D_k = \begin{pmatrix} y \\ k \end{pmatrix} (1-p)^k p^{y-k}\hat{\lambda}, \quad \text{for} \quad k = 1, 2, ..., y. \tag{6.4.1}$$

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 (6.4.2),

$$\hat{D}_0 = -\hat{\mu}; \quad \hat{D}_1 = (1-r)\hat{\mu}; \quad \hat{D}_2 = r\hat{\mu}. \tag{6.4.2}$$

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

Table 6.6: Queueing quantities for the vaccine inventory system

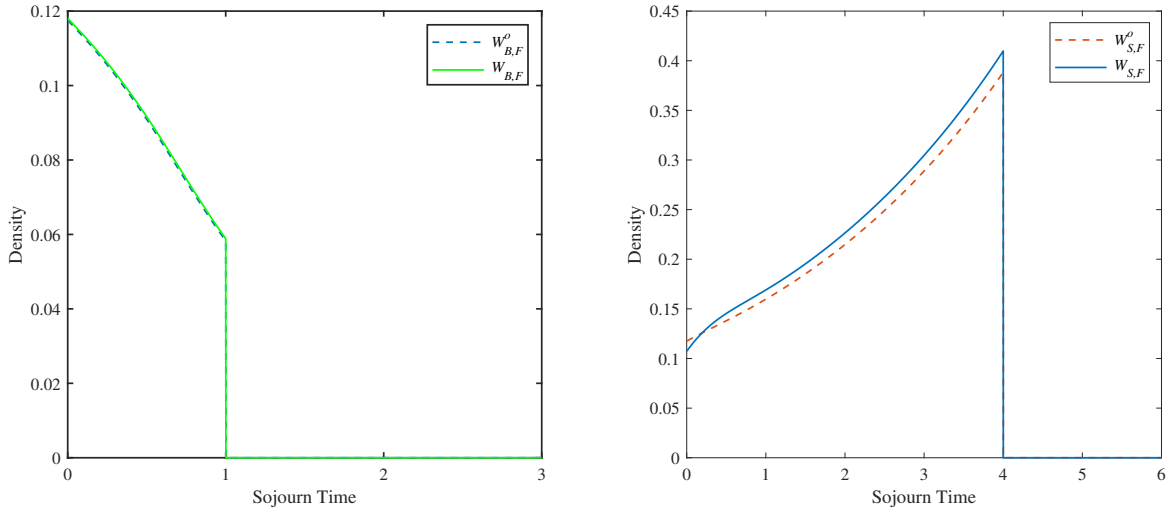

Figure 6.6: The stationary density functions of W^o_{BF},W_{BF},W^o_{SF} and W_{SF}

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 6.6 and Figure 6.6. From the results in Table 6.6, 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 6.7). When $\hat{\lambda} = 0.85$, 84.79% of vaccine doses are being used and 88.70% of patients' demands can be filled.

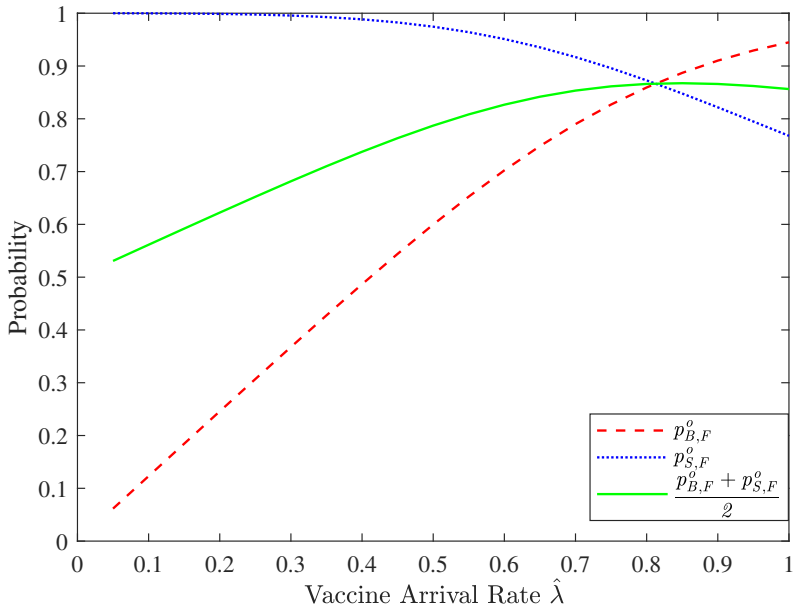

Figure 6.7: Matching probabilities with varying vaccine arrival rate.

6.4.1 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 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

$$\beta = [0.9, 0.1], \quad T = \begin{pmatrix} -9/2 & 0 \\ 0 & -1/8 \end{pmatrix}; \tag{6.4.3}$$

and the hyperexponential distribution for vaccines is

$$\beta = [0.8, 0.2], \quad T = \begin{pmatrix} -2 & 0 \\ 0 & -1/18 \end{pmatrix}. \tag{6.4.4}$$

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
Abandonemnt 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 6.7: Comparison of different abandonment distributions

From Table 6.7 we can see that the distribution of abandonment times has significant impacts on the queueing performance (e.g., matching probabilities), and with the increasing of the coefficient of variation of the abandonment distribution of either side, 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, the increasing of variation of the abandonment times increases the probability that 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 $MMPP$,

$$\hat{D}_0 = \begin{pmatrix} -16 & 2 \\ 1 & -1.5 \end{pmatrix}; \hat{D}_1 = \begin{pmatrix} 9.8 & 0 \\ 0 & 0.35 \end{pmatrix}; \hat{D}_2 = \begin{pmatrix} 4.2 & 0 \\ 0 & 0.15 \end{pmatrix}. \tag{6.4.5}$$

In this process, with probability of 1/3 the arrival rate is 14, with 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 (6.4.6) as the arrival process. In this process, with probability of 25% the arrival rate is 3, with probability of 75%, the arrival rate is 1/3.

$$D_0 = \begin{pmatrix} -6 + 0.2^{10} \times 3 & 3 \\ 1 & -4/3 + 0.2^{10} \times \frac{1}{3} \end{pmatrix};$$
$$D_k = \begin{pmatrix} \binom{10}{k} 0.8^k 0.2^{10-k} 3 & 0 \\ 0 & \binom{10}{k} 0.8^k 0.2^{10-k} \frac{1}{3} \end{pmatrix}, \quad \text{for} \quad k = 1, 2, ..., 10. \tag{6.4.6}$$

Second, we consider an ideal situation with constant interarrival times for the vaccine arrival process. We use a technique called Erlangization to approximate the constant interarrival times by Erlang distribution $Erlang(m_b, m_b\hat{\lambda})$[15, 100]. To obtain high accuracy, m_b needs to be relatively large (we consider $m_b = 50, 30, 10$ in this example). From Table 6.8, we can see that the matching probabilities for both sides are better if m_b is larger, which indicates that the situation with constant interarrival times for vaccine arrivals could result in the best performance. But constant vaccine replenishments time 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 6.8: Comparison of different arrival processes

The result of the sensitivity analysis on the arrival process given in Table 6.8 also shows that the non-stationary Poisson arrival processes (i.e., $MMPP$) result in worse 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 time distributions (normally with high variance) are shown to better fit the arrival processes in reality [119]. Therefore, the performance of the proposed model in terms of matching rate under more general distributions worse than 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 in such regions more accurately.

6.5 Summary

In this chapter, we consider a double-sided queueing model with $BMAP$ and finite discrete abandonment times. Our contributions are two-fold.

First, the model is quite general as $BMAP$ can approximate any stochastic arrival process and the finite discrete distribution can approximate general distributions. We assume that the abandonment times of the customers on both sides depend on their position in the queue and their batch size, which is a quite practical assumption. We use multi-layer $MMFF$ processes to analyze the queueing model and compute a number of queueing quantities such as the matching rates, fill rates, abandonment probabilities, distributions of sojourn times, and the mean queue lengths.

Second, we apply our model to a hypothetical 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 observe 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 uncertainty of the patients abandonment and non-stationary arrivals when designing the vaccine inventory system and to maintain a stable vaccine inventory system.

Chapter 7

Concluding Remarks

In this chapter, we summarize the main contributions in this thesis and give a brief discussion on several future research topics related to this thesis.

7.1 Summary

In this thesis, we first studied the basic theory on the multi-layer $MMFF$ processes. Specifically, we reviewed and refined the existing theory on multi-layer $MMFF$ processes and developed an efficient algorithm for computing the joint stationary distributions. Then we used three applications to queueing models to demonstrate the applicability of this approach. A number of interesting quantities were obtained for these queueing models, such as the abandonment probabilities, distributions of waiting times and the mean queue lengths. All three queueing models are fairly general and cover many interesting and challenging cases including i) models without abandonment; ii) models with only zero patience inputs (i.e., balking); iii) models with a constant abandonment time; and iv) the mixtures of all of them.

For the $MAP/PH/K + GI$ queueing model, several ideas in applied probability and queueing theory were put together with the multi-layer $MMFF$ processes to develop an algorithm for evaluating the queueing performances. Some of the queueing quantities

obtained in this model, such as the probability and waiting time of customers abandoning at the head of the queue or before reaching the head of the queue, are not easy to derive by other methods. We also demonstrated the efficiency of our algorithm by several numerical examples with moderately large numbers of servers.

For the double-sided queues with $MMAP$ and abandonment, the contributions are two-fold. First, the queueing model with multiple types of inputs and abandonment is fairly general. Second, we constructed a multi-layer $MMFF$ process to analyze the model and obtained a variety of queueing quantities, including both aggregate quantities and quantities for individual types, which can be useful to gain insight into the stochastic model of interest.

For the double-sided queues with $BMAP$ and abandonment, the contribution is mainly its generality and practicality. The generality of the model is guaranteed by the $BMAP$ and the discrete abandonment times. In addition, the abandonment time of a batch in this model depends on its batch size and its position in the queue, which makes the model more practical for inventory systems, such as the shelf life after opening in the perishable inventory systems. We applied this model to a vaccine inventory system and evaluated the system's performance. We considered various system settings and compared the performance under different levels of uncertainty.

7.2 Future Research

7.2.1 Stationary Distribution with Zero Mean-Drift

In Chapter 3, the expression of the joint density function in Theorem 3.2 does not include the case with $\zeta^{(n)} = 0$. Although this issue has been discussed briefly in [77], we hope to give computational details for multi-layer $MMFF$ processes with $\zeta^{(n)} = 0$, for some $n = 2, 3, ..., N-1$, in future research.

7.2.2 Other Queueing Models

In Chapter 4, we studied the $MAP/PH/K+GI$ queue, and we considered two double-sided queues with $MMAP$ and $BMAP$, respectively in Chapter 5 and Chapter 6. A natural extension is to consider $MMAP$ or $BMAP$ in the $MAP/PH/K+GI$ queueing model, e.g, a generalized queueing model with multiple types of customers (i.e., $MMAP[K]/PH[K]/N+G[K]$ queue). This model can be combined with some data analytic techniques to solve the emergency department abandonment problem, which will be discussed in Subsection 7.2.4.

Second, we plan to consider the $MAP/PH/K+GI$ queue with customer priorities in future research. This extension may require advanced applied probability tools such as the two-dimensional $MMFF$ processes. A more powerful tool may be found by combining multi-dimensional $MMFF$ processes and multi-layer $MMFF$ processes. More explorations are required for this research direction.

Last, double-sided queues have become an increasingly interesting research topic in recent years with the emerging of the sharing economy, for example, ridesharing, bicycle-sharing, online rental and online lending [34, 42]. 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 [51]. It may be an interesting direction to consider batch arrivals and priority in the double-sided queueing model. Based on the work of [1, 2], the double-sided queueing model with a type-dependent matching mechanism could also be an interesting topic for future research.

7.2.3 Applications to Perishable Inventory Systems

In Chapter 6, we studied a vaccine inventory system as a case study, however, there are some limitations in our vaccine inventory system. First, we did not consider the wastage of the remaining doses in an opened multi-dose vaccine vial at the end of the day open (i.e., open vial wastage), while open vial wastage is a major contributor to the vaccine wastage [79]. 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 [19], 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.

7.2.4 Utilizing the Proposed Models for Healthcare Data Analytics

In the healthcare system, abandonment or leaving without being seen (LWBS) is an undesirable problem in the emergency department. Abandonment can be viewed as an aspect of patient behavior and also a critical component of queueing models. Numerous studies have been conducted looking at the factors involved in LWBS in the emergency department, including age, gender, triage, chief complaint, etc [35, 57, 80].

Kaplan-Meier analysis and the survival tree algorithms are commonly used in the biostatistics field and also used in our other research projects on Amyotrophic lateral sclerosis (ALS) patients management. Inspired by these applications, we propose to estimate the abandonment time distributions of the patients in the emergency department and classify patients based on their abandonment times and contributing risk factors. First, we can use the Kaplan-Meier estimation to get the empirical distribution for the patient abandonment times. Empirical distribution makes no assumption about the underlying distribution of the abandonment time. Another advantage of the Kaplan-Meier estimation is that we can deal with right-censored data, a common data type in most emergency departments. Second, we can do patients classification based on abandonment time distribution and associated factors, which can improve the performance evaluation results. We can get the number of types of patient classes from the data by the survival tree algorithms.

With the empirical abandonment time distribution and multiple types of patients, the analysis of the queueing system becomes difficult. However, we have shown that our queueing models are able to handle both multiple types of customers and discrete aban-

donment time distribution. In addition, there are many existing algorithms to do parameters fitting for $MMAP$ and PH distributions [16, 40, 69]. Therefore, we can develop an $MMAP[K]/PH[K]/N+G[K]$ queueing model to evaluate the performance in the system. We can also get the performance for a specific type of patients, e.g., the waiting time and loss probability of senior male patients with ESI 3 triage.

References

[1] Ivo Adan, Igor Kleiner, Rhonda Righter, and Gideon Weiss. FCFS parallel service systems and matching models. *Performance Evaluation*, 127:253–272, 2018.

[2] Ivo Adan and Gideon Weiss. Exact FCFS matching rates for two infinite multitype sequences. *Operations Research*, 60(2):475–489, 2012.

[3] Philipp Afèche, Adam Diamant, and Joseph Milner. Double-sided batch queues with abandonment: Modeling crossing networks. *Operations Research*, 62(5):1179–1201, 2014.

[4] Soohan Ahn, Andrei L Badescu, and Vaidyanathan Ramaswami. Time dependent analysis of finite buffer fluid flows and risk models with a dividend barrier. *Queueing Systems*, 55(4):207–222, 2007.

[5] Soohan Ahn and Vaidyanathan Ramaswami. Fluid flow models and queues—a connection by stochastic coupling. *Stochastic Models*, 19(3):325–348, 2003.

[6] Soohan Ahn and Vaidyanathan Ramaswami. Transient analysis of fluid flow models via stochastic coupling to a queue. *Stochastic Models*, 20(1):71–101, 2004.

[7] Soohan Ahn and Vaidyanathan Ramaswami. Duality results for Markov-modulated fluid flow models. *Journal of Applied Probability*, 48(A):309–318, 2011.

[8] Mustafa Akan, Oguzhan Alagoz, Baris Ata, Fatih Safa Erenay, and Adnan Said. A broader view of designing the liver allocation system. *Operations Research*, 60(4):757–770, 2012.

[9] Attahiru Sule Alfa and Qi-Ming He. Point of queue size change analysis of the *PH/PH/k* system with heterogeneous servers. *Operations Research Letters*, 45(6):581–584, 2017.

[10] CJ Ancker Jr and AV Gafarian. Some queuing problems with balking and reneging. *I. Operations Research*, 11(1):88–100, 1963.

[11] David Anick, Debasis Mitra, and Man Mohan Sondhi. Stochastic theory of a data-handling system with multiple sources. *Bell System Technical Journal*, 61(8):1871–1894, 1982.

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

[13] Søren Asmussen. *Applied Probability and Queues*, volume 51. Springer Science & Business Media, New York, second edition, 2003.

[14] Søren Asmussen. Levy processes, phase-type distributions, and martingales. *Stochastic Models*, 30(4):443–468, 2014.

[15] Soren Asmussen, Florin Avram, and Miguel Usabel. Erlangian approximations for finite-horizon ruin probabilities. *ASTIN Bulletin: The Journal of the IAA*, 32(2):267–281, 2002.

[16] Søren Asmussen, Olle Nerman, and Marita Olsson. Fitting phase-type distributions via the EM algorithm. *Scandinavian Journal of Statistics*, 23(4):419–441, 1996.

[17] Søren Asmussen and Colm Art O'Cinneide. Representations for matrix-geometric and matrix-exponential steady-state distributions with applications to many-server queues. *Stochastic Models*, 14(1-2):369–387, 1998.

[18] Florin Avram and Miguel Usabel. Ruin probabilities and deficit for the renewal risk model with phase-type interarrival times. *Astin Bulletin*, 34(2):315–332, 2004.

[19] Zahra Azadi, Harsha Gangammanavar, and Sandra Eksioglu. Developing childhood vaccine administration and inventory replenishment policies that minimize open vial wastage. *Annals of Operations Research*, 292(1):215–247, 2020.

[20] Andrei Badescu, Lothar Breuer, Ana Da Silva Soares, Guy Latouche, Marie-Ange Remiche, and David Stanford. Risk processes analyzed as fluid queues. *Scandinavian Actuarial Journal*, 105(2):127–141, 2005.

[21] Andrei Badescu, Steve Drekic, and David Landriault. Analysis of a threshold dividend strategy for a *MAP* risk model. *Scandinavian Actuarial Journal*, 2007(4):227–247, 2007.

[22] Andrei Badescu, Steve Drekic, and David Landriault. On the analysis of a multi-threshold Markovian risk model. *Scandinavian Actuarial Journal*, 2007(4):248–260, 2007.

[23] Andrei Badescu and David Landriault. Moments of the discounted dividends in a threshold-type Markovian risk process. *Brazilian Journal of Probability and Statistics*, 21(1):13–25, 2007.

[24] Andrei Badescu and David Landriault. Recursive calculation of the dividend moments in a multi-threshold risk model. *North American Actuarial Journal*, 12(1):74–88, 2008.

[25] Andrei Badescu and David Landriault. 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, 2009.

[26] Andrei L Badescu, Lothar Breuer, Steve Drekic, Guy Latouche, and David A Stanford. The surplus prior to ruin and the deficit at ruin for a correlated risk process. *Scandinavian Actuarial Journal*, 2005(6):433–445, 2005.

[27] Shaul K Bar-Lev, Onno Boxma, Britt Mathijsen, and David Perry. A blood bank model with perishable blood and demand impatience. *Stochastic Systems*, 7(2):237–263, 2017.

[28] Nigel G Bean and Małgorzata M O'Reilly. A stochastic two-dimensional fluid model. *Stochastic Models*, 29(1):31–63, 2013.

[29] Nigel G Bean, Małgorzata M O'Reilly, and Zbigniew Palmowski. Yaglom limit for stochastic fluid models. *Advances in Applied Probability*, 53(3), 2021.

[30] Nigel G Bean, Małgorzata M O'Reilly, and Jane E Sargison. A stochastic fluid flow model of the operation and maintenance of power generation systems. *IEEE Transactions on Power Systems*, 25(3):1361–1374, 2010.

[31] Nigel G Bean, Małgorzata M O'reilly, and Peter G Taylor. Hitting probabilities and hitting times for stochastic fluid flows: The bounded model. *Probability in the Engineering and Informational Sciences*, 23(1):121–147, 2009.

[32] Nigel G Bean and Małgorzata M O'Reilly. Performance measures of a multi-layer Markovian fluid model. *Annals of Operations Research*, 160(1):99–120, 2008.

[33] Nigel G Bean, Małgorzata M O'Reilly, and Peter G Taylor. Hitting probabilities and hitting times for stochastic fluid flows. *Stochastic Processes and Their Applications*, 115(9):1530–1556, 2005.

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

[35] Ehsan Bolandifar, Nicole DeHoratius, Tava Olsen, and Jennifer Wiler. An empirical study of the behavior of patients who leave the emergency department without being seen. *Journal of Operations Management*, 65(5):430–446, 2019.

[36] Nam Kyoo Boots and Henk Tijms. An $M/M/c$ queue with impatient customers. *TOP: An Official Journal of the Spanish Society of Statistics and Operations Research*, 7(2):213–220, 1999.

[37] Onno J Boxma, Israel David, David Perry, and Wolfgang Stadje. A new look at organ transplantation models and double matching queues. *Probability in the Engineering and Informational Sciences*, 25(2):135–155, 2011.

[38] Lothar Breuer and Dieter Baum. *An Introduction to Queueing Theory and Matrix-Analytic Methods*. Springer Science & Business Media, 2005.

[39] Lothar Breuer, Alexander Dudin, and Valentina Klimenok. A retrial *BMAP/PH/n* system. *Queueing Systems*, 40(4):433–457, 2002.

[40] Peter Buchholz, Peter Kemper, and Jan Kriege. Multi-class markovian arrival processes and their parameter fitting. *Performance Evaluation*, 67(11):1092–1106, 2010.

[41] Bong Dae Choi, Bara Kim, and Dongbi Zhu. *MAP/M/c* queue with constant impatient time. *Mathematics of Operations Research*, 29(2):309–325, 2004.

[42] Maxime C Cohen. Big data and service operations. *Production and Operations Management*, 27(9):1709–1723, 2018.

[43] Ana da Silva Soares and Guy Latouche. Further results on the similarity between fluid queues and *QBD*s. In *Matrix-Analytic Methods: Theory and Applications*, pages 89–106. World Scientific, 2002.

[44] Ana Da Silva Soares and Guy Latouche. A matrix-analytic approach to fluid queues with feedback control. *International Journal of Simulation. Systems, Science and Technology*, 6(1-2):4–12, 2005.

[45] Ana da Silva Soares and Guy Latouche. Matrix-analytic methods for fluid queues with finite buffers. *Performance Evaluation*, 63(4-5):295–314, 2006.

[46] Ana da Silva Soares and Guy Latouche. Fluid queues with level dependent evolution. *European Journal of Operational Research*, 196(3):1041–1048, 2009.

[47] JG Dai and Shuangchi He. Customer abandonment in many-server queues. *Mathematics of Operations Research*, 35(2):347–362, 2010.

[48] JG Dai and Shuangchi He. Queues in service systems: Customer abandonment and diffusion approximations. In *Transforming Research into Action*, pages 36–59. INFORMS, 2011.

[49] JG Dai, Shuangchi He, and Tolga Tezcan. Many-server diffusion limits for *G/Ph/n+ GI* queues. *Annals of Applied Probability*, 20(5):1854–1890, 2010.

[50] Eline De Cuypere and Dieter Fiems. Performance evaluation of a kitting process. In *International Conference on Analytical and Stochastic Modeling Techniques and Applications*, pages 175–188. Springer, 2011.

[51] Adam Diamant and Opher Baron. Double-sided matching queues: Priority and impatient customers. *Operations Research Letters*, 47(3):219–224, 2019.

[52] Tessa Dzial, Lothar Breuer, Ana da Silva Soares, Guy Latouche, and Marie-Ange Remiche. Fluid queues to solve jump processes. *Performance Evaluation*, 62(1-4):132–146, 2005.

[53] Nicky van Foreest, Michel Mandjes, and Werner Scheinhardt. Analysis of a feedback fluid model for heterogeneous TCP sources. *Stochastic Models*, 19(3):299–324, 2003.

[54] David Freedman. *Approximating Countable Markov Chains*. Springer Science & Business Media, 2012.

[55] Noah Gans, Ger Koole, and Avishai Mandelbaum. Telephone call centers: Tutorial, review, and research prospects. *Manufacturing & Service Operations Management*, 5(2):79–141, 2003.

[56] Ofer Garnett, Avishai Mandelbaum, and Martin Reiman. Designing a call center with impatient customers. *Manufacturing & Service Operations Management*, 4(3):208–227, 2002.

[57] Jenna M Geers, Kalyan S Pasupathy, Kimberly K Lovik, Janet L Finley, Thomas R Hellmich, Gomathi Marisamy, David M Nestler, Annie T Sadosty, Mustafa Y Sir, and Heather A Heaton. Characterization of emergency department abandonment using a real-time location system. *The American Journal of Emergency Medicine*, 38(4):759–762, 2020.

[58] Chun-Hua Guo. Nonsymmetric algebraic Riccati equations and Wiener–Hopf factorization for M-matrices. *SIAM Journal on Matrix Analysis and Applications*, 23(1):225–242, 2001.

[59] Chun-Hua Guo. A note on the minimal nonnegative solution of a nonsymmetric algebraic Riccati equation. *Linear Algebra and its Applications*, 357(1-3):299–302, 2002.

[60] Xiao-Xia Guo, Wen-Wei Lin, and Shu-Fang Xu. A structure-preserving doubling algorithm for nonsymmetric algebraic Riccati equation. *Numerische Mathematik*, 103(3):393–412, 2006.

[61] Frank A Haight. Queueing with reneging. *Metrika*, 2(1):186–197, 1959.

[62] Qi-Ming He. *Fundamentals of Matrix-Analytic Methods*, volume 365. Springer, 2014.

[63] Qi-Ming He and Attahiru Sule Alfa. Construction of Markov chains for discrete time *MAP/PH/K* queues. *Performance Evaluation*, 93:17–26, 2015.

[64] Qi-Ming He and Attahiru Sule Alfa. Space reduction for a class of multidimensional Markov chains: A summary and some applications. *INFORMS Journal on Computing*, 30(1):1–10, 2017.

[65] Qi-Ming He and Marcel F Neuts. Markov chains with marked transitions. *Stochastic Processes and Their Applications*, 74(1):37–52, 1998.

[66] Qi-Ming He and Haoran Wu. Multi-layer *MMFF* processes and the *MAP/PH/K+GI* queue: Theory and algorithm. *Queueing Models and Service Management*, 3(1):37–87, 2020.

[67] Qi-Ming He, Hao Zhang, and Qingqing Ye. An *M/PH/K* queue with constant impatient time. *Mathematical Methods of Operations Research*, 87(1):139–168, 2018.

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

[69] Gábor Horváth and Hiroyuki Okamura. A fast *EM* algorithm for fitting marked Markovian arrival processes with a new special structure. In *European Workshop on Performance Engineering*, pages 119–133. Springer, 2013.

[70] Gábor Horváth and Benny Van Houdt. A multi-layer fluid queue with boundary phase transitions and its application to the analysis of multi-type queues with general customer impatience. In *Quantitative Evaluation of Systems (QEST), 2012 Ninth International Conference on Quantitative Evaluation of Systems*, pages 23–32. IEEE Computer Society Press, 2012.

[71] Guang-Hui Hsu and Qi-Ming He. The distribution of the first passage time for the Markov processes of *GI/M/1* type. *Communications in Statistics. Stochastic Models*, 7(3):397–417, 1991.

[72] Rouba Ibrahim and Ward Whitt. Real-time delay estimation in overloaded multi-server queues with abandonments. *Management Science*, 55(10):1729–1742, 2009.

[73] Ken'ichi Kawanishi. *QBD* approximations of a call center queueing model with general patience distribution. *Computers & Operations Research*, 35(8):2463–2481, 2008.

[74] David G Kendall. Some problems in the theory of queues. *Journal of the Royal Statistical Society: Series B (Methodological)*, 13(2):151–173, 1951.

[75] Bara Kim and Jeongsim Kim. A single server queue with Markov modulated service rates and impatient customers. *Performance Evaluation*, 83:1–15, 2015.

[76] Che Soong Kim, Vilena V Mushko, and Alexander N Dudin. Computation of the steady state distribution for multi-server retrial queues with phase type service process. *Annals of Operations research*, 201(1):307–323, 2012.

[77] Guy Latouche and Giang Nguyen. Analysis of fluid flow models. *Queueing Models and Service Management*, 1(2):1–29, 2018.

[78] Guy Latouche and Vaidyanathan Ramaswami. *Introduction to Matrix Analytic Methods in Stochastic Modeling*, volume 5. SIAM, 1999.

[79] Bruce Y Lee, Bryan A Norman, Tina-Marie Assi, Sheng-I Chen, Rachel R Bailey, Jayant Rajgopal, Shawn T Brown, Ann E Wiringa, and Donald S Burke. Single versus

multi-dose vaccine vials: an economic computational model. *Vaccine*, 28(32):5292–5300, 2010.

[80] David R Li, Jesse J Brennan, Allyson A Kreshak, Edward M Castillo, and Gary M Vilke. Patients who leave the emergency department without being seen and their follow-up behavior: a retrospective descriptive analysis. *The Journal of Emergency Medicine*, 57(1):106–113, 2019.

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

[82] Xin Liu, Qi Gong, and Vidyadhar G Kulkarni. Diffusion models for double-ended queues with renewal arrival processes. *Stochastic Systems*, 5(1):1–61, 2015.

[83] David M Lucantoni. New results on the single server queue with a batch Markovian arrival process. *Communications in Statistics. Stochastic Models*, 7(1):1–46, 1991.

[84] Avi Mandelbaum and Sergey Zeltyn. The impact of customers' patience on delay and abandonment: some empirically-driven experiments with the $M/M/n+ G$ queue. *OR Spectrum*, 26(3):377–411, 2004.

[85] Michel Mandjes, Debasis Mitra, and Werner Scheinhardt. Models of network access using feedback fluid queues. *Queueing Systems*, 44(4):365–398, 2003.

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

[87] Masakiyo Miyazawa. Markov modulated fluid network process: Tail asymptotics of the stationary distribution. *Stochastic Models*, 37(1):127–167, 2021.

[88] Marcel F Neuts. Probability distributions of phase type. In *Liber Amicorum Prof. Emeritus H. Florin*, pages 173–206. Department of Mathematics, University of Louvain, 1975.

[89] Marcel F Neuts. A versatile Markovian point process. *Journal of Applied Probability*, 16(4):764–779, 1979.

[90] Marcel F Neuts. *Structured Stochastic Matrices of M/G/1 Type and Their Applications*. Dekker, 1989.

[91] Marcel F Neuts. *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach*. Courier Corporation, 1994.

[92] Małgorzata M O'Reilly and Werner Scheinhardt. Stationary distributions for a class of Markov-modulated tandem fluid queues. *Stochastic Models*, 33(4):524–550, 2017.

[93] David Perry and Wolfgang Stadje. Perishable inventory systems with impatient demands. *Mathematical Methods of Operations Research*, 50(1):77–90, 1999.

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

[95] Satheesh Ramachandran and Dursun Delen. Performance analysis of a kitting process in stochastic assembly systems. *Computers & Operations Research*, 32(3):449–463, 2005.

[96] Vadyanathan Ramaswami. The *N/G/1* queue and its detailed analysis. *Advances in Applied Probability*, 12(1):222–261, 1980.

[97] Vaidyanathan Ramaswami. Independent Markov processes in parallel. *Stochastic Models*, 1(3):419–432, 1985.

[98] Vaidyanathan Ramaswami. Matrix analytic methods for stochastic fluid flows. In *Teletraffic Engineering in a Competitive World (Proceedings of the 16th International Teletraffic Congress)*, pages 1019–1030. Elsevier Science B.V., 1999.

[99] Vaidyanathan Ramaswami and David M Lucantoni. Algorithms for the multi-server queue with phase type service. *Stochastic Models*, 1(3):393–417, 1985.

[100] Vaidyanathan Ramaswami, Douglas G Woolford, and David A Stanford. The Erlangization method for Markovian fluid flows. *Annals of Operations Research*, 160(1):215–225, 2008.

[101] S Subba Rao. Queuing models with balking, reneging, and interruptions. *Operations Research*, 13(4):596–608, 1965.

[102] S Subba Rao. Queuing with balking and reneging in *M/G/1* systems. *Metrika*, 12(1):173–188, 1967.

[103] Leonard CG Rogers. Fluid models in queueing theory and Wiener-Hopf factorization of Markov chains. *Annals of Applied Probability*, 4(2):390–413, 1994.

[104] Aviva Samuelson, Andrew Haigh, Małgorzata M O'Reilly, and Nigel G Bean. Stochastic model for maintenance in continuously deteriorating systems. *European Journal of Operational Research*, 259(3):1169–1179, 2017.

[105] Matthieu Simon. SIR epidemics with stochastic infectious periods. *Stochastic Processes and Their Applications*, 130(7):4252–4274, 2020.

[106] Pradip Som, WE Wilhelm, and Ralph L Disney. Kitting process in a stochastic assembly system. *Queueing Systems*, 17(3-4):471–490, 1994.

[107] Misa Takahashi, Hideo Ōsawa, and Takehisa Fujisawa. On a synchronization queue with two finite buffers. *Queueing Systems*, 36(1-3):107–123, 2000.

[108] Misa Takahashi and Yukio Takahashi. 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*, pages 375–390, 2000.

[109] Caglar Tunc and Nail Akar. Markov fluid queue model of an energy harvesting IoT device with adaptive sensing. *Performance Evaluation*, 111:1–16, 2017.

[110] Benny Van Houdt. 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, 2012.

[111] Dietmar Wagner. Analysis of mean values of a multi-server model with non-preemptive priorities and non-renewal input. *Stochastic Models*, 13(1):67–84, 1997.

[112] Kuo-Hsiung Wang and Ying-Chung Chang. Cost analysis of a finite $M/M/R$ queueing system with balking, reneging, and server breakdowns. *Mathematical Methods of Operations Research*, 56(2):169–180, 2002.

[113] Wei-Guo Wang, Wei-Chao Wang, and Ren-Cang Li. Alternating-directional doubling algorithm for m-matrix algebraic riccati equations. *SIAM Journal on Matrix Analysis and Applications*, 33(1):170–194, 2012.

[114] Ward Whitt. Improving service by informing customers about anticipated delays. *Management Science*, 45(2):192–207, 1999.

[115] Ward Whitt. Engineering solution of a basic call-center model. *Management Science*, 51(2):221–235, 2005.

[116] Haoran Wu and Qi-Ming He. Double-sided queues with marked Markovian arrival processes and abandonment. *Stochastic Models*, 37(1):23–58, 2021.

[117] Wei Xiong and Tayfur Altiok. An approximation for multi-server queues with deterministic reneging times. *Annals of Operations Research*, 172(1):143, 2009.

[118] Wei Xiong, David Jagerman, and Tayfur Altiok. *M/G/1* queue with deterministic reneging times. *Performance Evaluation*, 65(3-4):308–316, 2008.

[119] Wanfei Yang, Monika Parisi, Betsy J Lahue, Md Jasim Uddin, and David Bishai. 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, 2014.

[120] Mehmet Akif Yazıcı and Nail Akar. Analysis of continuous feedback Markov fluid queues and its applications to modeling optical burst switching. In *Proceedings of the 2013 25th International Teletraffic Congress (ITC)*, pages 1–8. IEEE, 2013.

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

APPENDICES

Appendix A

Newton's Method to the Quadratic Riccati Equations

In this Appendix, we present the Newton's method in [58] to obtain the minimal nonnegative solution Ψ of the quadratic Riccati equation

$$C_+^{-1}T_{+-} + C_+^{-1}T_{++}\Psi + \Psi C_-^{-1}T_{--} + \Psi C_-^{-1}T_{-+}\Psi = 0. \tag{A.0.1}$$

We give the computational steps for solving this equation, while the dual equation of $\widehat{\Psi}$ in Equation (3.1.9) can be solved similarly.

1. Let $A = C_+^{-1}T_{++}$, $B = C_+^{-1}T_{+-}$, $C = C_-^{-1}T_{-+}$ and $D = C_-^{-1}T_{--}$;
2. Find $a = \max_i a_{ii}$, $d = \max_i d_{ii}$ and let $A_\gamma = A + aI$, $D_\gamma = D + dI$;
3. Initialize $\Psi_0 = 0$ and $n = 0$;
4. Compute $\Psi_{n+1} = (A_\gamma\Psi_n + \Psi_n D_\gamma + \Psi_n B\Psi_n + C)/(a + d)$ iteratively. Use $||\Psi_{n+1} - \Psi_n||_1 < 10^{-15}$ as the stopping criteria.

Given the condition that $\zeta \neq 0$, $\{\Psi_n, n \geq 0\}$ converges to Ψ quadratically.

More efficient algorithms called doubling algorithms can solve for Ψ and $\widehat{\Psi}$ simultaneously, such as the Structure-preserving Doubling Algorithm in [60] and Alternating-Directional Doubling Algorithm in [113]. Since those algorithms are faster due to fewer computations for Ψ and $\widehat{\Psi}$ together, they typically require only half as much time as the Newton's method.

Appendix B

Lemmas

B.1 Evaluation of Several Integrals

Closed form expressions given in this Appendix are partially obtained in [66]. We present them here for convenience and completeness. Define

$$
\begin{aligned}
&\mathcal{L}_{a,b}^{\mathcal{K}} = \int_a^b \exp(\mathcal{K}(x-a))\mathrm{d}x; \qquad \widetilde{\mathcal{L}}_{a,b}^{\mathcal{K}} = \int_a^b \exp(\mathcal{K}(b-x))\mathrm{d}x; \\
&\mathcal{M}_{a,b}^{\mathcal{K}} = \int_a^b x\exp(\mathcal{K}(x-a))\mathrm{d}x; \quad \widetilde{\mathcal{M}}_{a,b}^{\mathcal{K}} = \int_a^b x\exp(\mathcal{K}(b-x))\mathrm{d}x; \\
&\mathcal{M}_{a,b}^{(\mathcal{K},2)} = \int_a^b x^2\exp(\mathcal{K}(x-a))\mathrm{d}x; \quad \widetilde{\mathcal{M}}_{a,b}^{(\mathcal{K},2)} = \int_a^b x^2\exp(\mathcal{K}(b-x))\mathrm{d}x; \\
&\mathcal{M}_{a,b}^{(\mathcal{K},n)} = \int_a^b x^n\exp(\mathcal{K}(x-a))\mathrm{d}x; \quad \widetilde{\mathcal{M}}_{a,b}^{(\mathcal{K},n)} = \int_a^b x^n\exp(\mathcal{K}(b-x))\mathrm{d}x; \\
&\qquad\qquad \text{for} \quad n = 3,4,5,... \\
&\mathcal{L}_{a,b}^{(\mathcal{K},D)} = \int_a^b \exp(\mathcal{K}(x-a))\left(I\otimes e^{D(x-a)}\otimes I\right)\mathrm{d}x; \\
&\widetilde{\mathcal{L}}_{a,b}^{(\mathcal{K},D)} = \int_a^b \exp(\mathcal{K}(b-x))\widehat{\Psi}^{(n)}\left(I\otimes e^{D(x-a)}\otimes I\right)\mathrm{d}x.
\end{aligned}
\tag{B.1.1}
$$

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

$$\begin{aligned}
\mathcal{L}_{a,b}^{\mathcal{K}} = \widetilde{\mathcal{L}}_{a,b}^{\mathcal{K}} &= \mathcal{K}^{-1}(e^{\mathcal{K}(b-a)} - I); \\
\mathcal{M}_{a,b}^{\mathcal{K}} &= \mathcal{K}^{-1}\left(\mathcal{K}^{-1} - aI + (bI - \mathcal{K}^{-1})e^{\mathcal{K}(b-a)}\right); \\
\widetilde{\mathcal{M}}_{a,b}^{\mathcal{K}} &= \mathcal{K}^{-1}\left(-\mathcal{K}^{-1} - bI + (aI + \mathcal{K}^{-1})e^{\mathcal{K}(b-a)}\right); \\
\mathcal{M}_{a,b}^{(\mathcal{K},2)} &= \mathcal{K}^{-1}\left(b^2 e^{\mathcal{K}(b-a)} - a^2 I - 2\mathcal{M}_{a,b}^{\mathcal{K}}\right); \\
\widetilde{\mathcal{M}}_{a,b}^{(\mathcal{K},2)} &= \mathcal{K}^{-1}\left(a^2 e^{\mathcal{K}(b-a)} - b^2 I + 2\widetilde{\mathcal{M}}_{a,b}^{\mathcal{K}}\right); \\
\mathcal{M}_{a,b}^{(\mathcal{K},n)} &= \mathcal{K}^{-1}\left(b^n e^{\mathcal{K}(b-a)} - a^n I - n\mathcal{M}_{a,b}^{(\mathcal{K},n-1)}\right); \\
\widetilde{\mathcal{M}}_{a,b}^{(\mathcal{K},n)} &= \mathcal{K}^{-1}\left(a^n e^{\mathcal{K}(b-a)} - b^n I + n\widetilde{\mathcal{M}}_{a,b}^{(\mathcal{K},n-1)}\right),
\end{aligned} \tag{B.1.2}$$

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

$$\begin{aligned}
&\mathcal{K}\mathcal{L}_{a,b}^{(\mathcal{K},D)} + \mathcal{L}_{a,b}^{(\mathcal{K},D)}(I \otimes D \otimes I) = e^{\mathcal{K}(b-a)}\left(I \otimes e^{D(b-a)} \otimes I\right) - I; \\
&\mathcal{K}\widetilde{\mathcal{L}}_{a,b}^{(\mathcal{K},D)} - \widetilde{\mathcal{L}}_{a,b}^{(\mathcal{K},D)}(I \otimes D \otimes I) = e^{\mathcal{K}(b-a)}\widehat{\Psi}^{(n)} - \widehat{\Psi}^{(n)}\left(I \otimes e^{D(b-a)} \otimes I\right).
\end{aligned} \tag{B.1.3}$$

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

$$\begin{aligned}
\mathcal{L}_{a,b}^{\mathcal{K}} = \widetilde{\mathcal{L}}_{a,b}^{\mathcal{K}} &= (\mathcal{K} - \mathbf{v}_R\mathbf{v}_L)^{-1}\left(e^{\mathcal{K}(b-a)} - I\right) + (b-a)\mathbf{v}_R\mathbf{v}_L; \\
\mathcal{M}_{a,b}^{\mathcal{K}} &= (\mathcal{K} - \mathbf{v}_R\mathbf{v}_L)^{-1}\left(be^{\mathcal{K}(b-a)} - aI\right) + \frac{(b^2-a^2)}{2}\mathbf{v}_R\mathbf{v}_L \\
&\quad -(\mathcal{K} - \mathbf{v}_R\mathbf{v}_L)^{-2}\left(e^{\mathcal{K}(b-a)} - I\right) + (b-a)\mathbf{v}_R\mathbf{v}_L; \\
\widetilde{\mathcal{M}}_{a,b}^{\mathcal{K}} &= (\mathcal{K} - \mathbf{v}_R\mathbf{v}_L)^{-1}\left(ae^{\mathcal{K}(b-a)} - bI\right) + \frac{(b^2-a^2)}{2}\mathbf{v}_R\mathbf{v}_L \\
&\quad +(\mathcal{K} - \mathbf{v}_R\mathbf{v}_L)^{-2}\left(e^{\mathcal{K}(b-a)} - I\right) - (b-a)\mathbf{v}_R\mathbf{v}_L; \\
\mathcal{M}_{a,b}^{(\mathcal{K},2)} &= (\mathcal{K} - \mathbf{v}_R\mathbf{v}_L)^{-1}\left(b^2 e^{\mathcal{K}(b-a)} - a^2 I - 2\mathcal{M}_{a,b}^{\mathcal{K}}\right) + \frac{(b^3-a^3)}{3}\mathbf{v}_R\mathbf{v}_L; \\
\widetilde{\mathcal{M}}_{a,b}^{(\mathcal{K},2)} &= (\mathcal{K} - \mathbf{v}_R\mathbf{v}_L)^{-1}\left(a^2 e^{\mathcal{K}(b-a)} - b^2 I + 2\widetilde{\mathcal{M}}_{a,b}^{\mathcal{K}}\right) + \frac{(b^3-a^3)}{3}\mathbf{v}_R\mathbf{v}_L; \\
\mathcal{M}_{a,b}^{(\mathcal{K},n)} &= (\mathcal{K} - \mathbf{v}_R\mathbf{v}_L)^{-1}\left(b^n e^{\mathcal{K}(b-a)} - a^n I - n\mathcal{M}_{a,b}^{(\mathcal{K},n-1)}\right) + \frac{(b^{n+1}-a^{n+1})}{n+1}\mathbf{v}_R\mathbf{v}_L; \\
\widetilde{\mathcal{M}}_{a,b}^{(\mathcal{K},n)} &= (\mathcal{K} - \mathbf{v}_R\mathbf{v}_L)^{-1}\left(a^n e^{\mathcal{K}(b-a)} - b^n I + n\widetilde{\mathcal{M}}_{a,b}^{(\mathcal{K},n-1)}\right) + \frac{(b^{n+1}-a^{n+1})}{n+1}\mathbf{v}_R\mathbf{v}_L,
\end{aligned} \tag{B.1.4}$$

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

$$\begin{aligned}
&(\mathcal{K}-\mathbf{v}_R\mathbf{v}_L)\mathcal{L}_{a,b}^{(\mathcal{K},D)}+\mathcal{L}_{a,b}^{(\mathcal{K},D)}(I\otimes D\otimes I)=e^{\mathcal{K}(b-a)}\left(I\otimes e^{D(b-a)}\otimes I\right)-I-\mathbf{v}_R\mathbf{v}_L L_1;\\
&(\mathcal{K}-\mathbf{v}_R\mathbf{v}_L)\widetilde{\mathcal{L}}_{a,b}^{(\mathcal{K},D)}-\widetilde{\mathcal{L}}_{a,b}^{(\mathcal{K},D)}(I\otimes D\otimes I)=e^{\mathcal{K}(b-a)}\widehat{\Psi}^{(n)}-\widehat{\Psi}^{(n)}\left(I\otimes e^{D(b-a)}\otimes I\right)-\mathbf{v}_R\mathbf{v}_L\widehat{\Psi}^{(n)}L_1,
\end{aligned}\tag{B.1.5}$$

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

Proof. We only consider $\mathcal{L}_{a,b}^{M}$, $\mathcal{M}_{a,b}^{\mathcal{K}}$, $\mathcal{M}_{a,b}^{(\mathcal{K},2)}$ and $\mathcal{M}_{a,b}^{(\mathcal{K},n)}$. We can easily have following equations:

$$\begin{aligned}
\mathcal{K}\int_a^b e^{\mathcal{K}(x-a)}\mathrm{d}x &= \int_a^b \mathrm{d}e^{\mathcal{K}(x-a)}=e^{\mathcal{K}(b-a)}-I;\\
\mathcal{K}\int_a^b xe^{\mathcal{K}(x-a)}\mathrm{d}x &= \int_a^b x\mathrm{d}e^{\mathcal{K}(x-a)}=be^{\mathcal{K}(b-a)}-aI-\int_a^b e^{\mathcal{K}(x-a)}\mathrm{d}x;\\
\mathcal{K}\int_a^b x^2e^{\mathcal{K}(x-a)}\mathrm{d}x &= \int_a^b x^2\mathrm{d}e^{\mathcal{K}(x-a)}=b^2e^{\mathcal{K}(b-a)}-a^2I-2\int_a^b xe^{\mathcal{K}(x-a)}\mathrm{d}x;\\
\mathcal{K}\int_a^b x^ne^{\mathcal{K}(x-a)}\mathrm{d}x &= \int_a^b x^n\mathrm{d}e^{\mathcal{K}(x-a)}=b^ne^{\mathcal{K}(b-a)}-a^nI-n\int_a^b x^{n-1}e^{\mathcal{K}(x-a)}\mathrm{d}x;\\
\mathbf{v}_R\mathbf{v}_L\int_a^b e^{\mathcal{K}(x-a)}\mathrm{d}x &= \mathbf{v}_R\mathbf{v}_L(b-a);\\
\mathbf{v}_R\mathbf{v}_L\int_a^b xe^{\mathcal{K}(x-a)}\mathrm{d}x &= \mathbf{v}_R\mathbf{v}_L\frac{(b^2-a^2)}{2};\\
\mathbf{v}_R\mathbf{v}_L\int_a^b x^2e^{\mathcal{K}(x-a)}\mathrm{d}x &= \mathbf{v}_R\mathbf{v}_L\frac{(b^3-a^3)}{3};\\
\mathbf{v}_R\mathbf{v}_L\int_a^b x^ne^{\mathcal{K}(x-a)}\mathrm{d}x &= \mathbf{v}_R\mathbf{v}_L\frac{(b^{n+1}-a^{n+1})}{n+1}.
\end{aligned}\tag{B.1.6}$$

If $\mathcal{K}$ is non-singular, the results are obtained directly from the first four equations. If $\mathcal{K}$ is singular, then $\mathcal{K}-\mathbf{v}_R\mathbf{v}_L$ is non-singular. The results are obtained by routine calculations using all the above equations. Results for $\widetilde{\mathcal{L}}_{a,b}^{\mathcal{K}}$, $\widetilde{\mathcal{M}}_{a,b}^{\mathcal{K}}$, $\widetilde{\mathcal{M}}_{a,b}^{(\mathcal{K},2)}$ and $\widetilde{\mathcal{M}}_{a,b}^{(\mathcal{K},n)}$ can be obtained similarly.

The proof of $\mathcal{L}_{a,b}^{(\mathcal{K},D)}$ and $\widetilde{\mathcal{L}}_{a,b}^{(\mathcal{K},D)}$ are similar, details are omitted. □

B.2 The Probability Generating Function of *MAPs*

The probability generating functions (PGF) of the number of arrivals for MAP, $BMAP$ and $MMAP$ given in this Appendix are well-known (See [62]). We present them here for convenience and completeness.

For MAP with matrix representation (D_0, D_1), let $\{N(t), t \geq 0\}$ be the number of arrivals by time t and $P^*(z,t)$ be the conditional PGF of $N(t)$, for $z \geq 0$. According to Theorem 2.3.1 in [62], we have

$$P^*(z,t) = \sum_{n=o}^{\infty} z^n P(n,t) = \exp\{(D_0 + zD_1)t\}. \tag{B.2.1}$$

By Theorem 2.3.2 in [62], if the underlying Markov chain is irreducible and the initial distribution is $\boldsymbol{\alpha}$, we have

$$\mathbb{E}[N(t)] = \lambda t + \boldsymbol{\alpha}(\exp(Dt) - I)(D - \mathbf{e}\boldsymbol{\theta})^{-1} D_1 \mathbf{e}, t \geq 0, \tag{B.2.2}$$

where $D = D_o + D_1$ and λ is the stationary arrival rate.

For $BMAP$ with matrix representation $(D_0, D_1, ..., D_K)$, let $\{N_B(t), t \geq 0\}$ be the number of arrivals by time t and $P_B^*(z,t)$ be the PGF of $N_B(t)$, for $z \geq 0$. According to Theorem 2.4.1 in [62], we have

$$P_B^*(z,t) = \exp\{(D_0 + zD_1 + z^2 D_2 + ... + z^K D_K)t\}, \tag{B.2.3}$$

and according to Theorem 2.4.2 in [62], given the initial distribution of the underlying Markov chain $\boldsymbol{\alpha}$, we have

$$\mathbb{E}[N_B(t)] = \lambda t + \boldsymbol{\alpha}(\exp(Dt) - I)(D - \mathbf{e}\boldsymbol{\theta})^{-1} \left(\sum_{j=1}^{K} j D_j \mathbf{e} \right), t \geq 0, \tag{B.2.4}$$

where $D = D_0 + D_1 + .. + D_K$ is irreducible and has stationary distribution vector $\boldsymbol{\theta}$ and $\lambda = \boldsymbol{\theta} \left(\sum_{j=1}^{K} j D_j \right) \mathbf{e}$ is the stationary arrival rate.

For $MMAP$ with matrix representation $(D_0, D_1, ..., D_K)$, if we want to get the total number of arrivals by time t, we can convert it into an MAP with representation $(D_0, D_1 + D_2 + ... + D_K)$. We can also get the joint PGF for the numbers of arrivals by time t for each type. Let $P^*_M(\mathbf{z}, t)$ be the joint PGF and $N_{M,k}(t)$ be the number of arrivals of type k customers, $k = 1, 2, ..., K$. According to Theorem 2.5.1 in [62], we have

$$P^*_M(\mathbf{z}, t) = \exp\{(D_0 + z_1 D_1 + z_2 D_2 + ... + z_K D_K)t\}, \tag{B.2.5}$$

where $\mathbf{z} = (z_1, z_2, ..., z_K)$. Given the initial state $\boldsymbol{\alpha}$, we have

$$\mathbb{E}[N_{M,k}(t)] = \lambda_k t + \boldsymbol{\alpha}(\exp(Dt) - I)(D - \mathbf{e}\boldsymbol{\theta})^{-1} D_k \mathbf{e}, t \geq 0, \tag{B.2.6}$$

where $D = D_0 + D_1 + .. + D_K$ is irreducible and has stationary distribution vector $\boldsymbol{\theta}$ and $\lambda_k = \boldsymbol{\theta} D_k \mathbf{e}$ is the arrival rate of type k customers.

Appendix C

Important Notations

In this appendix, we summarize important notations of each chapter from Chapter 3 to Chapter 6 in following tables. The notations of Chapter 3 can be applied to all the following chapters.

Indices:	
t	time ($t \geq 0$)
N	number of layers/borders ($n \in \{1, ..., N\}$)
Random Variables:	
$X(t)$	the fluid level at time t
$\phi(t)$	the state of the underlying Markov chain at time t
Parameters: (Note	that we add superscript (n) to represent Layer n parameters)
c_i	changing rate of the fluid level when $\phi(t) = i$
$\mathcal{S}$	state space of the underlying Markov chain
$\mathcal{S}_+, \mathcal{S}_-, \mathcal{S}_0$	$\mathcal{S}_+ = \{i \in \mathcal{S} : \ c_i > 0\}$; $\mathcal{S}_- = \{i \in \mathcal{S} : \ c_i < 0\}$; $\mathcal{S}_0 = \{i \in \mathcal{S} : \ c_i = 0\}$
$\mathbf{c}$, $\mathbf{c}_+$, $\mathbf{c}_-$	vectors $\mathbf{c} = \{c_i, i \in \mathcal{S}\}$; $\mathbf{c}_+ = \{c_i, i \in \mathcal{S}_+\}$; $\mathbf{c}_- = \{c_i, i \in \mathcal{S}_-\}$
ζ	mean drift of the fluid level
C_+, C_-	$C_+ = \text{diag}(\mathbf{c}_+)$; $C_- = -\text{diag}(\mathbf{c}_-)$
Q	the infinitesimal generator of the underlying Markov chain
Q_{++}, Q_{+-}, Q_{+0}	the transition from $\mathcal{S}_+$ to $\mathcal{S}_+, \mathcal{S}_-$ and $\mathcal{S}_0$
Q_{-+}, Q_{--}, Q_{-0}	the transition from $\mathcal{S}_-$ to $\mathcal{S}_+, \mathcal{S}_-$ and $\mathcal{S}_0$
Q_{0+}, Q_{0-}, Q_{00}	the transition from $\mathcal{S}_0$ to $\mathcal{S}_+, \mathcal{S}_-$ and $\mathcal{S}_0$
T	the censored underlying Markov chain
Basic Quantities:	(Note that we add superscript (n) to represent Layer n quantities)
Ψ, $\widehat{\Psi}$	the transition of the state of Q at regenerative epochs
$\mathcal{U}$, $\widehat{\mathcal{U}}$	a continuous time Markov chain related to the minimal (maximal) of the fluid flow process.
$\mathcal{K}$, $\widehat{\mathcal{K}}$	associated with numbers of visits to certain fluid level and state during some first passage periods

Table C.1: Important notations in Chapter 3

Multi-layer Parameters:	
l_n	Border n
(l_{n-1}, l_n)	Layer n
$\mathcal{S}_b^{(n)}$	state space of the underlying Markov chain on Border n
$Q_b^{(n)}$, $Q_{b+}^{(n)}$, $Q_{b-}^{(n)}$	generator on Border n
$P_{+b+}^{(n)}$, $P_{+b-}^{(n)}$, $P_{+bb}^{(n)}$	transition probabilities when approaching Border n from above
$P_{-b+}^{(n)}$, $P_{-b-}^{(n)}$, $P_{-bb}^{(n)}$	transition probabilities when approaching Border n from below
Multi-layer Quantities:	
$\Lambda_{++}^{(b-a)}$	the probabilities for the process to go from level a to level b before returning to level a.
$\widehat{\Lambda}_{--}^{(b-a)}$	the probabilities for the process to go from level b to level a before returning to level b
$\Psi_{+-}^{(b-a)}$	similar to Ψ except that the process does not reach fluid level b and the process starts in fluid level a
$\widehat{\Psi}_{-+}^{(b-a)}$	similar to $\widehat{\Psi}$ except that the process does not reach fluid level a and the process starts in fluid level b
Joint Stationary Distribution:	
$p_j^{(n)}$	border probabilities
$\mathbf{p}^{(n)}$	$\mathbf{p}^{(n)} = (p_j^{(n)} : \ j \in \mathcal{S}_b^{(n)})$, for $n = 1, 2, ..., N-1$;
$\pi_j^{(n)}(x)$	density function
$\boldsymbol{\pi}^{(n)}(x)$	$\boldsymbol{\pi}^{(n)}(x) = (\pi_j^{(n)}(x) : \ j \in \mathcal{S}^{(n)})$, for $n = 1, 2, ..., N$ and $-\infty < x < \infty$
$\mathbf{u}_+^{(n)}$, $\mathbf{u}_-^{(n)}$	coefficients in the density function of $MMFF$ processes
c_{norm}	normalization factor of $MMFF$ processes

Table C.2: Important notations in Chapter 3 (Continued)

Indices:	
m_a	order of Markovian arrival process
m_s	order of phase-type distributed service time ($i = 1, 2, \ldots, m_s$)
N	number of possible abandonment epochs ($n = 1, ..., N$)
K	number of servers
Random Variables:	
$a(t)$	the age of the customer at the head of the queue at time t
$I_a(t)$	the phase of the customer arrival process at time t
$I_{(a)}(t)$	the phase of the customer arrival process right after the arrival of the customer at the head of the queue
$n_i(t)$	the number of servers whose service phase is i at time t
τ	the abandonment time
Parameters:	
(D_0, D_1)	Markovian arrival processes
λ	the (average) customer arrival rate
η_n	the probability that abandonment time is l_n
$(\boldsymbol{\beta}, T)$	service time representation
μ_s	the service rate
Quantities:	
$\mathbf{f}^{(n)}(x)$	density function of the age process
$\hat{\mathbf{p}}_k^{(1)}$	border probabilities of the age process
$\mathbf{v}_+^{(n)}$, $\mathbf{v}_-^{(n)}$	coefficients of the density function of the age process
$\hat{c}_{norm}$	normalization factor of the age process
$p_S, p_L, p_{L,1}, p_{L,>1}$	abandonment probabilities
$W_S, W_{L,1}, W_{L,>1}$	waiting times
$q_S(t), q_W(t), q_{tot}(t)$	queue lengths

Table C.3: Important notations in Chapter 4

Indices:	
m_a, m_b	order of the marked Markovian arrival process of passengers (taxis)
N, M	number of possible abandonment epochs ($n = 1, ..., N; m = 1, ..., M$)
K	number of types of passengers ($k = 1, ..., K$)
H	number of types of taxis ($h = 1, ..., H$)
Random Variables:	
$a(t)$, $a_P(t)$, $a_T(t)$	the age of the passenger (taxi) at the head of the queue at time t
$I_a(t)$, $I_b(t)$	the phase of the passenger (taxi) arrival process at time t
$I_{(a)}(t)$, $I_{(b)}(t)$	the phase of the passenger (taxi) arrival process right after the arrival of the passenger (taxi) at the head of the queue
$s(t)$	the type of the passenger (taxi) at the head of the queue at time t
τ_k, $\hat{\tau}_h$	the abandonment time for type k passengers (type h taxis)
Parameters:	
$(D_0, ..., D_K)$	marked Markovian arrival processes of passengers
λ_k	the (average) type k passenger arrival rate
λ	the total average arrival rate of passengers $\lambda = \sum_{k=1}^{K} \lambda_k$
$\eta_{k,n}$	the probability that abandonment time is $\tilde{l}_n$ for type k passengers
$(B_0, ..., B_H)$	marked Markovian arrival processes of taxis
μ_h	the (average) type h taxis arrival rate
μ	the total average arrival rate of taxis $\mu = \sum_{h=1}^{H} \mu_h$
$\hat{\eta}_{h,m}$	the probability that abandonment time is $\hat{l}_m$ for type h taxis
Quantities: (Note that some quantities are omitted here, check Table 5.1 for detailed summary)	
ω, $\omega_P(k)$, $\omega(k,h)$	matching rates
$p_{P,S}, p_{P,L}, p_{PL,1}, p_{PL,>1}$	abandonment probabilities of passengers
$W_{P,S}, W_{PL,1}, W_{PL,>1}$	waiting times of passengers
$q_P(t), q_P(k,t)$	queue lengths

Table C.4: Important notations in Chapter 5

Indices:	
m_b, m_s	order of the batch Markovian arrival process of buyers (sellers)
N, M	number of possible abandonment epochs ($n = 1, ..., N; m = 1, ..., M$)
K	maximum batch size of orders ($k = 1, ..., K$)
Random Variables:	
$a(t)$, $a_B(t)$, $a_S(t)$	the age of the buyer (seller) at the head of the queue at time t
$I_a(t)$, $I_b(t)$	the phase of the passenger (taxi) arrival process at time t
$I_{(a)}(t)$, $I_{(b)}(t)$	the phase of the passenger (taxi) arrival process right after the arrival of the passenger (taxi) at the head of the queue
$s(t)$, $s_B(t)$, $s_S(t)$	the remaining batch size of the buyer (seller) at the head of the queue at time t
τ_k, $\hat{\tau}_h$	the abandonment time of a buyer (seller) with batch size k before reaching the head of the queue
$\dot{\tau}_k$, $\dot{\hat{\tau}}_k$	the abandonment time of a buyer (seller) with batch size k after becoming the head of the queue
Parameters:	
$(D_0, ..., D_K)$	batch Markovian arrival processes of buyers
λ_k	the (average) type k buyers arrival rate
λ	the arrival rate of buyer orders $\lambda = \sum_{k=1}^{K} k\lambda_k$
$\eta_{k,n}$, $\dot{\eta}_{k,n}$	the probability that abandonment time is $\tilde{l}_n$, for size k buyers
$(B_0, ..., B_K)$	batch Markovian arrival processes of sellers
μ_k	the (average) type k sellers arrival rate
μ	the arrival rate of seller orders $\mu = \sum_{k=1}^{K} k\mu_k$
$\hat{\eta}_{k,m}$, $\dot{\hat{\eta}}_{k,m}$	the probability that abandonment time is $\hat{l}_m$, for size k sellers
Quantities: (Note that we add superscript "o" to represent order level quantities)	
ω_B	matching rates of buyers
$p_{B,F}, p_{B,L}, p_{BL,1}, p_{BL,>1}$	loss probabilities of buyers
$W_{B,F}, W_{BL,1}, W_{BL,>1}$	sojourn times of buyers
$q_B(t)$	queue lengths

Table C.5: Important notations in Chapter 6