STUDIES
ON
ALGORITHMIC ANALYSIS OF QUEUES
WITH BATCH MARKOVIAN ARRIVAL STREAMS

HIROYUKI MASUYAMA

STUDIES
ON
ALGORITHMIC ANALYSIS OF QUEUES
WITH BATCH MARKOVIAN ARRIVAL STREAMS

HIROYUKI MASUYAMA

STUDIES
ON
ALGORITHMIC ANALYSIS OF QUEUES
WITH BATCH MARKOVIAN ARRIVAL STREAMS

by

HIROYUKI MASUYAMA

Submitted in partial fulfillment of
the requirement for the degree of
DOCTOR OF INFORMATICS
(Applied Mathematics and Physics)

KYOTO UNIVERSITY
KYOTO 606–8501, JAPAN
NOVEMBER 2003

Preface

Algorithmic analysis of queues, pioneered by Neuts [Neut79] in the mid-1970's, focuses on developing numerical algorithms to compute performance measures of interest in queueing models, rather than on obtaining their formal solutions. Its basic idea is to introduce appropriate auxiliary variables and represent queueing models as Markov chains (or Markov processes) whose structures are suitable for algorithmic analysis of their transient and steady-state solutions. Before the era of the high-performance computer, algorithmic analysis did not gain much attention because the resulting numerical algorithms demand considerable computer performance from a practical viewpoint. As computer performance is improving, however, algorithmic analysis becomes important in queueing research.

It is Markovian arrival process (MAP) [Luca90] which must not be forgotten when discussing algorithmic analysis of queues. MAP is a natural extension of Poisson process, in which the arrival rate is governed by a continuous-time Markov chain with finite states. MAP and its extensions (e.g. batch MAP [Luca91] and marked MAP [He96, He01]) have enough flexibilities for practical use and good representations for algorithmic analysis. For the last decade, many researchers have studied queues with MAP and its extensions and made a great contribution to algorithmic analysis in queueing research. In most of those works, however, service times are assumed to be independent and identically distributed (i.i.d.).

This thesis studies queues with batch Markovian arrival streams. The respective arrival streams are batch MAPs (BMAPs), which are extensions of MAP to allow batch arrivals, and they may have different service time distributions. Queues with batch Markovian arrival streams are so flexible that they can represent most of the queues studied in the past as special cases. In chapter 2, we establish a numerical algorithm to compute the stationary joint queue length distribution in a FIFO single-server queue by extending Takine's recent results [Taki01a, Taki01b, Taki01c]. In a very similar way to chapter 2, chapter 3 constructs a numerical algorithm for the stationary joint queue length distribution in a FIFO single-server queue with service interruptions. In chapter 4, we consider an infinite-server queue. Assuming phase-type service time distributions, we obtain an explicit and numerically feasible formula for time-dependent binomial moments of the queue length. Finally, chapter 5 studies a processor-sharing queue, focusing on computing the sojourn time distribution. Although a single arrival stream and exponential services are assumed there, the processor-sharing queue has been recognized to be very hard to obtain results suitable for the computation of the sojourn time distribution.

The main contribution of this thesis is to develop numerical algorithms to compute performance

measures of interest in several queues with stream-dependent service times. Although the results in this thesis is only one small step in such research, the author hopes that they will be helpful to further research in this field.

Hiroyuki Masuyama
November 2003

Acknowledgment

This thesis would not have been completed without the help of many people, whom I would like to acknowledge here.

First of all, I wish to express my gratitude to Associate Professor Tetsuya Takine of Kyoto University for his constant encouragement and instruction since my graduate student days. I had not studied queueing theory until I became a graduate student, but, his seminar on queueing theory and Markov process let me find the pleasure of studying them and decided to pursue research in this field. Also, I deeply appreciate his helpful comments on an earlier draft of this thesis. Without his support, none of this work could have been completed.

I am indebted to Professor Masao Fukushima of Kyoto University for his insightful suggestions and comments. Although I had not many opportunities to receive his direct instruction, his valuable advice greatly improved my research. Further, I wish to thank other members of Fukushima Laboratory, including Assistant Professor Nobuo Yamashita of Kyoto University, for providing a pleasant working environment. I feel happy to have studied at Fukushima Laboratory for five years.

I would also like to express my great appreciation to Asuka Ikeda, who will be my wife. For six years, she has been always encouraging me to overcome difficulties encountered in advancing my research. Without her moral support, I could not complete this thesis.

Finally, my special thanks are due to my parents for their financial assistance and heartfelt encouragement.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This chapter provides materials to be required in the following chapters and the brief survey of previous works on queues fed by multiple arrival streams with different service time distributions. Throughout this thesis, we denote matrices and vectors by bold capital letters and bold small letters, respectively, and the empty sum is defined as zero.

1.1 Markovian Input Process

This section introduces an input process with batch Markovian arrival streams having different service time distributions. In this input process, customer arrivals from each arrival stream follow a BMAP [Luca91], which is an extension of a MAP [Luca90]. Therefore, after discussing MAP and BMAP [Luca91] in subsection 1.1.1, we formally define the input process with batch Markovian arrival streams in subsection 1.1.2.

1.1.1 Markovian arrival process

We begin with the definition of MAP. We consider a time homogeneous, stationary Markov chain with finite states $\mathcal{M} = \{1, \ldots, M\}$, which is assumed to be irreducible. The Markov chain is called the underlying Markov chain hereafter.

MAP is defined as follows. The underlying Markov chain stays in state i ($i \in \mathcal{M}$) for an exponential interval of time with mean μ_i^{-1}. When the sojourn time in state i has elapsed, there are two possibilities: (1) With probability $d_{i,j}$, a transition to state j ($j \in \mathcal{M}$) happens with an arrival. (2) With probability $c_{i,j}$ ($j \in \mathcal{M}$, $j \neq i$), a transition to state j happens without an arrival. Note that

$$\sum_{\substack{j \in \mathcal{M} \\ j \neq i}} c_{i,j} + \sum_{j \in \mathcal{M}} d_{i,j} = 1, \qquad \text{for all } i \in \mathcal{M}.$$

For the later use, we introduce two matrices, $\boldsymbol{C}$ and $\boldsymbol{D}$. Let $\boldsymbol{C}$ denote an $M \times M$ matrix whose (i,j)th $(i, j \in \mathcal{M})$ element $C_{i,j}$ is given by

$$C_{i,j} = \begin{cases} -\mu_i, & \text{if } i = j, \\ \mu_i c_{i,j}, & \text{otherwise.} \end{cases}$$

Let $\boldsymbol{D}$ denote an $M \times M$ nonnegative matrix whose (i,j)th $(i,j \in \mathcal{M})$ element $D_{i,j}$ is given by

$$D_{i,j} = \mu_i d_{i,j}.$$

Note that the infinitesimal generator of the underlying Markov chain is given by $\boldsymbol{C}+\boldsymbol{D}$. We define $\boldsymbol{\pi}$ as the stationary probability vector of the underlying Markov chain. Because of the finite state space $\mathcal{M}$ and the irreducibility of the underlying Markov chain, $\boldsymbol{\pi}$ is uniquely determined. Note that $\boldsymbol{\pi}$ satisfies

$$\boldsymbol{\pi}(\boldsymbol{C}+\boldsymbol{D}) = \boldsymbol{0}, \qquad \boldsymbol{\pi}\boldsymbol{e} = 1, \tag{1.1}$$

where $\boldsymbol{e}$ denotes a column vector whose elements are all equal to one. The arrival rate λ is then given by

$$\lambda = \boldsymbol{\pi}\boldsymbol{D}\boldsymbol{e}.$$

We define $N(t)$ and $S(t)$ as the number of arrivals in time interval $(0,t]$ and the state of the underlying Markov chain at time t, respectively. Further, let $\boldsymbol{N}(t,n)$ denote an $M \times M$ matrix whose (i,j)th $(i,j \in \mathcal{M})$ element represents

$$\Pr\left[N(t) = n, S(t) = j \mid N(0) = 0, S(0) = i\right].$$

$\boldsymbol{N}(t,n)$'s $(t \geq 0, n = 0,1,\ldots)$ satisfy the forward Chapman-Kolmogorov equations:

$$\begin{aligned} \frac{d}{dt}\boldsymbol{N}(t,n) &= \boldsymbol{N}(t,n)\boldsymbol{C} + \boldsymbol{N}(t,n-1)\boldsymbol{D}, \qquad t \geq 0,\ n = 1,2,\ldots, \\ \boldsymbol{N}(0,0) &= \boldsymbol{I}, \end{aligned}$$

where $\boldsymbol{I}$ denotes an identity matrix with appropriate dimension. Thus, the z-transform $\boldsymbol{N}^*(t,z)$ of $\boldsymbol{N}(t,n)$ is given by

$$\boldsymbol{N}^*(t,z) = \exp\left[(\boldsymbol{C}+z\boldsymbol{D})t\right].$$

As mentioned above, MAP can be characterized by $(\boldsymbol{C},\boldsymbol{D})$.

In what follows, we show some special cases of MAP.

Example 1.1 Poisson process. *For $\boldsymbol{C} = -\lambda$ and $\boldsymbol{D} = \lambda$, MAP is reduced to a Poisson process with arrival rate λ.*

Example 1.2 Phase-type (PH) renewal process. *PH renewal process is defined in a very similar way to MAP except that the state of the underlying Markov chain immediately after an arrival is chosen according to a $1 \times M$ probability vector $\boldsymbol{\beta}$, independent of its state immediately before the arrival. Namely, $\boldsymbol{D} = (-\boldsymbol{T}\boldsymbol{e})\boldsymbol{\beta}$, where we use, by convention, $\boldsymbol{T}$ for $\boldsymbol{C}$ to distinguish PH renewal process from MAP. Note that the ith $(i \in \mathcal{M})$ element of $M \times 1$ vector $(-\boldsymbol{T}\boldsymbol{e})$ represents the arrival (renewal) rate given the underlying Markov chain is in state i. Note also that inter-arrival (renewal) times are i.i.d. according to distribution function $\psi(x)$:*

$$\psi(x) = 1 - \boldsymbol{\beta}\exp(\boldsymbol{T}x)\boldsymbol{e},$$

which is called the phase-type distribution with representation $(\boldsymbol{\beta},\boldsymbol{T})$.

Example 1.3 Markov modulated Poisson process (MMPP). *MMPP is a natural extension of a Poisson process, which has the arrival rate modulated by an M-state irreducible Markov chain. To put it more concretely, the arrival rate takes M values $\lambda_1, \ldots \lambda_M$, and is equal to λ_j whenever the Markov chain is in state j. Let $\boldsymbol{R}$ denote the infinitesimal generator of the underlying Markov chain. Let $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots \lambda_M)$. MMPP can then be characterized by $\boldsymbol{R}$ and $\boldsymbol{\Lambda}$. Note here that this MMPP is equivalent to a MAP with representation $(\boldsymbol{C}, \boldsymbol{D})$, where*

$$\boldsymbol{C} = \boldsymbol{R} - \boldsymbol{\Lambda}, \qquad \boldsymbol{D} = \boldsymbol{\Lambda}.$$

We next discuss BMAP. The definition of BMAP is very similar to that of MAP except that multiple arrivals may happen simultaneously. Given the underlying Markov chain is in state i ($i \in \mathcal{M}$), it changes its state to state j ($j \in \mathcal{M}$) at rate $\mu_i d_{i,j}(n)$ and n arrivals occur at the same time. Besides, transitions with no arrivals are driven by $\boldsymbol{C}$, as well as MAP. Note here that $d_{i,j}(n)$'s satisfy

$$\sum_{\substack{j \in \mathcal{M} \\ j \neq i}} c_{i,j} + \sum_{j \in \mathcal{M}} \sum_{n=1}^{\infty} d_{i,j}(n) = 1, \qquad \text{for all } i \in \mathcal{M},$$

where $c_{i,j}$ denotes the (i,j)th element of $\boldsymbol{C}$. Let $\boldsymbol{D}(n)$ denote an $M \times M$ matrix whose (i,j)th ($i, j \in \mathcal{M}$) element $D_{i,j}(n)$ is given by

$$D_{i,j}(n) = \mu_i d_{i,j}(n).$$

Note that $\boldsymbol{C} + \sum_{n=1}^{\infty} \boldsymbol{D}(n)$ is the infinitesimal generator of the underlying Markov chain. Note further that the arrival rate λ is given by

$$\lambda = \boldsymbol{\pi} \sum_{n=1}^{\infty} n \boldsymbol{D}(n) \boldsymbol{e}, \tag{1.2}$$

where $\boldsymbol{\pi}$ denotes the stationary probability vector of the underlying Markov chain. Note also that $\boldsymbol{N}(t,n)$'s for BMAP satisfy

$$\begin{aligned} \frac{d}{dt}\boldsymbol{N}(t,n) &= \boldsymbol{N}(t,n)\boldsymbol{C} + \sum_{m=1}^{n} \boldsymbol{N}(t,n-m)\boldsymbol{D}(m), \qquad t \geq 0,\ n = 1,2,\ldots, \\ \boldsymbol{N}(0,0) &= \boldsymbol{I}. \end{aligned}$$

It then follows from the above equations that the z-transform $\boldsymbol{N}^*(t,z)$ of $\boldsymbol{N}(t,n)$ is given by

$$\boldsymbol{N}^*(t,z) = \exp\left[(\boldsymbol{C} + \boldsymbol{D}^*(z))\, t\right],$$

where

$$\boldsymbol{D}^*(z) = \sum_{n=1}^{\infty} z^n \boldsymbol{D}(n).$$

Thus, BMAP can be characterized by $(\boldsymbol{C}, \boldsymbol{D}(n))$.

1.1.2 Input process with batch Markovian arrival streams

It is assumed that there exist K ($K \geq 1$) arrival streams. Customers arriving from the kth ($k \in \mathcal{K} = \{1, \ldots, K\}$) stream is called *class* k customers. Service times of class k customers are assumed to be i.i.d. according to a distribution function $H_k(x)$ ($x \geq 0$) with finite mean h_k.

As well as in subsection 1.1.1, we consider the underlying Markov chain with the finite state space $\mathcal{M} = \{1, \ldots, M\}$ and assume its irreducibility. Customer arrivals happen in the following way. The underlying Markov chain stays in state i ($i \in \mathcal{M}$) for an exponential interval of time with mean μ_i^{-1}, and then changes its state to state j ($j \in \mathcal{M}$) with probability $\sigma_{i,j}$, where $\sum_{j \in \mathcal{M}} \sigma_{i,j} = 1$ for all i ($i \in \mathcal{M}$). When a transition from state i to state j happens, n_k ($k \in \mathcal{K}$) class k customers simultaneously arrive at the queue with probability $\sigma_{i,j}(n_1, \ldots, n_K)/\sigma_{i,j}$, where $\sigma_{i,i}(0, \ldots, 0)$ is assumed to be zero for all i ($i \in \mathcal{M}$), and $\sigma_{i,j}(n_1, \ldots, n_K)$ satisfies

$$\sigma_{i,j} = \sum_{n_1=1}^{\infty} \cdots \sum_{n_K=1}^{\infty} \sigma_{i,j}(n_1, \ldots, n_K).$$

The assumption $\sigma_{i,i}(0, \ldots, 0) = 0$ ($\forall i \in \mathcal{M}$) implies that at least one customer arrives whenever a transition from any state i to itself happens.

We now introduce some notations to describe the above input process. We define $\boldsymbol{C}$ as an $M \times M$ matrix whose (i,j)th element $C_{i,j}$ is given by

$$C_{i,j} = \begin{cases} -\mu_i, & i = j, \\ \sigma_{i,j}(\boldsymbol{0})\mu_i, & \text{otherwise}, \end{cases}$$

where $\boldsymbol{0}$ denotes a vector of zeros with appropriate dimension. We also define $\mathcal{Z}$ and $\mathcal{Z}^+$ as

$$\begin{aligned} \mathcal{Z} &= \{(n_1, \ldots, n_K);\ n_k = 0, 1, \ldots \text{ for all } k \in \mathcal{K}\}, \\ \mathcal{Z}^+ &= \mathcal{Z} - \{\boldsymbol{0}\}, \end{aligned}$$

respectively. Let $\boldsymbol{n}$ denote a $1 \times K$ nonnegative vector $(n_1, \ldots, n_K)$. We then define $\boldsymbol{D}(\boldsymbol{n})$ ($\boldsymbol{n} \in \mathcal{Z}^+$) as an $M \times M$ matrix whose (i,j)th element $D_{i,j}(\boldsymbol{n})$ is given by

$$D_{i,j}(\boldsymbol{n}) = \sigma_{i,j}(\boldsymbol{n})\mu_i, \qquad \boldsymbol{n} \in \mathcal{Z}^+.$$

When a state transition driven by $\boldsymbol{D}(\boldsymbol{n})$ occurs, n_k customers of each class k simultaneously arrive. On the other hand, when a state transition driven by $\boldsymbol{C}$ occurs, no customers arrive. Note here that the infinitesimal generator of the underlying Markov chain is given by $\boldsymbol{C} + \boldsymbol{D}$, where $\boldsymbol{D}$ is defined as

$$\boldsymbol{D} = \sum_{\boldsymbol{n} \in \mathcal{Z}^+} \boldsymbol{D}(\boldsymbol{n}). \tag{1.3}$$

Thus, the counting process of arrivals is characterized by $(\boldsymbol{C}, \boldsymbol{D}(\boldsymbol{n}))$.

Example 1.4 *Suppose that*

$$\boldsymbol{D}(\boldsymbol{n}) = \boldsymbol{O} \qquad \text{if } \boldsymbol{n} \in \mathcal{Z} - \{\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_K\},$$

where $\boldsymbol{O}$ *denotes a matrix of zeros and* $\boldsymbol{e}_k$ *denotes the* k*th unit vector:*

$$\boldsymbol{e}_k = (\overbrace{0,\ldots,0,1,0,\ldots,0}^{K}). \tag{1.4}$$
$$\qquad\quad kth$$

This assumption means that at most one arrival from an arrival stream happens at any time. This special case is called marked MAP (MMAP) in [He96].

Finally, we consider the marginal arrival process of the kth ($k \in \mathcal{K}$) stream. We define $\widetilde{\boldsymbol{C}}_k$ and $\widetilde{\boldsymbol{D}}_k(n)$ ($n = 1, 2, \ldots$) as

$$\widetilde{\boldsymbol{C}}_k = \boldsymbol{C} + \sum_{\substack{\boldsymbol{n}\in\mathcal{Z}^+ \\ n_k=0}} \boldsymbol{D}_k(\boldsymbol{n}), \qquad \widetilde{\boldsymbol{D}}_k(n) = \sum_{\substack{\boldsymbol{n}\in\mathcal{Z}^+ \\ n_k=n}} \boldsymbol{D}_k(\boldsymbol{n}),$$

respectively. Note that $\widetilde{\boldsymbol{C}}_k$ and $\widetilde{\boldsymbol{D}}_k(n)$ satisfy

$$\widetilde{\boldsymbol{C}}_k + \sum_{n=1}^{\infty} \widetilde{\boldsymbol{D}}_k(n) = \boldsymbol{C} + \boldsymbol{D}.$$

When a transition driven by $\widetilde{\boldsymbol{D}}_k(n)$ happens, n customers of class k arrive at the queue. Besides, a transition driven by $\widetilde{\boldsymbol{C}}_k$ happens without a class k customer. Thus, the marginal arrival process of the kth stream is a BMAP with representation $(\widetilde{\boldsymbol{C}}_k, \widetilde{\boldsymbol{D}}_k(n))$. We define λ_k ($k \in \mathcal{K}$) as the arrival rate of class k customers, i.e.,

$$\lambda_k = \boldsymbol{\pi} \sum_{n=1}^{\infty} n\widetilde{\boldsymbol{D}}_k(n)\boldsymbol{e},$$

where $\boldsymbol{\pi}$ denotes the invariant probability vector for $\boldsymbol{C} + \boldsymbol{D}$. Let ρ_k denote the utilization factor of class k customers, i.e., $\rho_k = \lambda_k h_k$. Note that the overall arrival rate λ and utilization factor ρ are given by

$$\lambda = \sum_{k\in\mathcal{K}} \lambda_k, \qquad \rho = \sum_{k\in\mathcal{K}} \rho_k,$$

respectively. Note also that in a work-conserving single-server queue, $\rho < 1$ ensures that the system is stable [Loyn62].

1.2 Markov Chains Skip Free to One Direction

This section summarizes analytical results on time-homogeneous Markov chains with the skip free to one direction and spatial homogeneity properties. Markov chains with these properties are classified into two cases, i.e., the skip free to the left case and to the right case, which are natural extensions of embedded Markov chains in an M/G/1 queue and GI/M/1 queue, respectively. Such Markov chains now play a vital role in algorithmic analysis of the queue length distribution. In what follows, we first introduce Markov chains with the skip free to the right/left properties and briefly discuss the steady-state solutions of these two types of Markov chains in subsections 1.2.1 and 1.2.2.

In a single-server queue, let X_k^- and D_k denote the number of customers immediately before the kth arrival and the number of departures during the kth inter-arrival time, respectively. We then have

$$X_{k+1}^- = \max(X_k^- + 1 - D_{k+1}, 0), \qquad k = 0, 1, \ldots.$$

We now consider a GI/M/1 queue, where inter-arrival times are i.i.d. according to a distribution function $G(x)$ with mean λ^{-1} and service times are exponential with mean μ^{-1}. In such a queue, D_k's are i.i.d. random variables, and hence the sequence $\{X_k^-; k = 0, 1, \ldots\}$ of random variables is a Markov chain with transition probability matrix:

$$\begin{bmatrix} f_0 & e_0 & 0 & 0 & \cdots \\ f_1 & e_1 & e_0 & 0 & \cdots \\ f_2 & e_2 & e_1 & e_0 & \cdots \\ f_3 & e_3 & e_2 & e_1 & \cdots \\ f_4 & e_4 & e_3 & e_2 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix},$$

where

$$e_n = \int_0^\infty e^{-\mu x} \frac{(\mu x)^n}{n!} dG(x), \qquad f_n = \sum_{m=n+1}^{\infty} e_m.$$

If $\rho = \lambda/\mu < 1$, the Markov chain $\{X_k^-; k = 0, 1, \ldots\}$ has the stationary distribution. Assuming $\rho < 1$, we define $q_{\mathrm{GI/M/1}}(n)$ $(n = 0, 1, \ldots)$ as $\lim_{k\to\infty} \Pr[X_k^- = n]$. Then $q_{\mathrm{GI/M/1}}(n)$ is given by

$$q_{\mathrm{GI/M/1}}(n) = \gamma^n(1 - \gamma), \qquad n = 0, 1, \ldots, \tag{1.5}$$

where γ is a positive number satisfying $\gamma = \sum_{n=0}^{\infty} e_n \gamma^n$ and $\gamma < 1$. The assumption $\rho < 1$ ensures that γ is uniquely determined.

On the other hand, in a single-server queue, let X_k^+ and A_k denote the number of customers immediately after the kth departure and the number of arrivals during the kth service time. We then have

$$X_{k+1}^+ = \max(X_k^+ - 1, 0) + A_{k+1}.$$

We consider an M/G/1 queue, where arrivals follow Poisson process with a rate λ and service times are i.i.d. according to a distribution $H(x)$ with mean μ^{-1}. In the M/G/1 queue, A_k's are i.i.d. random variables, and hence $\{X_k^+; k = 0, 1, \ldots\}$ is a Markov chain with transition probability matrix:

$$\begin{bmatrix} a_0 & a_1 & a_2 & a_3 & \cdots \\ a_0 & a_1 & a_2 & a_3 & \cdots \\ 0 & a_0 & a_1 & a_2 & \cdots \\ 0 & 0 & a_0 & a_1 & \cdots \\ 0 & 0 & 0 & a_0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix},$$

where

$$a_n = \int_0^\infty e^{-\lambda x} \frac{(\lambda x)^n}{n!} dH(x).$$

If $\rho = \lambda/\mu < 1$, the Markov chain $\{X_k^+; k = 0, 1, \ldots\}$ has the stationary distribution. Under the assumption $\rho < 1$, we define $q_{\mathrm{M/G/1}}(n)$ $(n = 0, 1, \ldots)$ as $\lim_{k\to\infty} \Pr[X_k^+ = n]$. Then $q_{\mathrm{M/G/1}}(n)$'s are recursively determined in the following way.

$$\begin{aligned} q_{\mathrm{M/G/1}}(0) &= 1 - \rho, \\ q_{\mathrm{M/G/1}}(n) &= \left[q_{\mathrm{M/G/1}}(0)\overline{a}_n + \sum_{m=1}^{n-1} q_{\mathrm{M/G/1}}(m)\overline{a}_{n+1-m} \right] (1 - \overline{a}_1)^{-1}, \qquad n = 1, 2, \ldots, \end{aligned}$$

where

$$\overline{a}_n = \sum_{m=n}^{\infty} a_m.$$

Note that Markov chains $\{X_k^-; k = 0, 1, \ldots\}$ and $\{X_k^+; k = 0, 1, \ldots\}$ have the following two properties. One is the spatial homogeneity except for the boundary part (e.g., if $X_k^+ \geq 1$, X_k^+ increases by $n - 1$ with probability a_n). The other is the skip free to one direction property. To put it more concretely, the Markov chain $\{X_k^-; k = 0, 1, \ldots\}$ (resp. $\{X_k^+; k = 0, 1, \ldots\}$) can increase (resp. decrease) at most by one when a transition happens. This is called the *skip free to the right* (resp. left) property.

Next we consider a GI/PH/1 queue and a PH/G/1 queue, where the service time distribution and the inter-arrival time distribution are of phase-type with representation $(\boldsymbol{\beta}_H, \boldsymbol{T}_H)$ and $(\boldsymbol{\beta}_A, \boldsymbol{T}_A)$, respectively. Let $G(x)$ and $H(x)$ denote the inter-arrival distribution in the GI/PH/1 queue and the service time distribution in the PH/G/1 queue, respectively. Note that D_k's in the GI/PH/1 queue and A_k's in the PH/G/1 queue are non-i.i.d. random variables. However, D_{k+1} (resp. A_{k+1}) is conditionally independent of the past history $\{D_l; l = 1, \ldots, k\}$ (resp. $\{A_l; l = 1, \ldots, k\}$), given the state J_k^- (resp. J_k^+) of the phase-type distribution $(\boldsymbol{\beta}_H, \boldsymbol{T}_H)$ (resp. $(\boldsymbol{\beta}_A, \boldsymbol{T}_A)$) immediately before (resp. after) the kth arrival (resp. departure). Thus, $\{(X_k^-, J_k^-); k = 0, 1, \ldots\}$ is a bivariate Markov chain whose transition probability matrix is given by

$$\boldsymbol{P}_{\mathrm{GI/PH/1}} = \begin{bmatrix} \boldsymbol{F}_0 & \boldsymbol{E}_0 & \boldsymbol{O} & \boldsymbol{O} & \cdots \\ \boldsymbol{F}_1 & \boldsymbol{E}_1 & \boldsymbol{E}_0 & \boldsymbol{O} & \cdots \\ \boldsymbol{F}_2 & \boldsymbol{E}_2 & \boldsymbol{E}_1 & \boldsymbol{E}_0 & \cdots \\ \boldsymbol{F}_3 & \boldsymbol{E}_3 & \boldsymbol{E}_2 & \boldsymbol{E}_1 & \cdots \\ \boldsymbol{F}_4 & \boldsymbol{E}_4 & \boldsymbol{E}_3 & \boldsymbol{E}_2 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \tag{1.6}$$

where $\boldsymbol{E}_n$'s $(n = 0, 1, \ldots)$ satisfy

$$\sum_{n=0}^{\infty} z^n \boldsymbol{E}_n = \int_0^\infty \exp\left[(\boldsymbol{T}_H + z(-\boldsymbol{T}_H \boldsymbol{e})\boldsymbol{\beta}_H)\, x \right] dG(x),$$

and

$$\boldsymbol{F}_n = \sum_{m=n+1}^{\infty} \boldsymbol{E}_m \boldsymbol{e}\boldsymbol{\beta}_H, \qquad n = 0, 1, \ldots.$$

Similarly, $\{(X_k^+, J_k^+); k = 0, 1, \ldots\}$ is a bivariate Markov chain whose transition probability matrix is given by

$$\boldsymbol{P}_{\mathrm{PH/G/1}} = \begin{bmatrix} \boldsymbol{B}_0 & \boldsymbol{B}_1 & \boldsymbol{B}_2 & \boldsymbol{B}_3 & \cdots \\ \boldsymbol{A}_0 & \boldsymbol{A}_1 & \boldsymbol{A}_2 & \boldsymbol{A}_3 & \cdots \\ \boldsymbol{O} & \boldsymbol{A}_0 & \boldsymbol{A}_1 & \boldsymbol{A}_2 & \cdots \\ \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{A}_0 & \boldsymbol{A}_1 & \cdots \\ \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{A}_0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \tag{1.7}$$

where $\boldsymbol{A}_n$'s $(n = 0, 1, \ldots)$ satisfy

$$\sum_{n=0}^{\infty} z^n \boldsymbol{A}_n = \int_0^{\infty} \exp\left[(\boldsymbol{T}_A + z(-\boldsymbol{T}_A \boldsymbol{e})\boldsymbol{\beta}_A)\, x\right] dH(x),$$

and

$$\boldsymbol{B}_n = \boldsymbol{e}\boldsymbol{\beta}_A \boldsymbol{A}_n, \qquad n = 0, 1, \ldots.$$

We can see from (1.6) and (1.7) that these two bivariate Markov chains have the spatial homogeneity (except for the boundary part) and the skip free to one direction property with respect to the first variables, i.e., X_k^- and X_k^+. As mentioned above, the queue length processes in queues we often encounter are reduced to bivariate Markov chains with these two properties. In general, if a bivariate Markov chain has a transition probability matrix with the same structure as

$$\boldsymbol{P}_{\mathrm{R}} = \begin{bmatrix} \boldsymbol{F}_0 & \widehat{\boldsymbol{E}}_0 & \boldsymbol{O} & \boldsymbol{O} & \cdots \\ \boldsymbol{F}_1 & \boldsymbol{E}_1 & \boldsymbol{E}_0 & \boldsymbol{O} & \cdots \\ \boldsymbol{F}_2 & \boldsymbol{E}_2 & \boldsymbol{E}_1 & \boldsymbol{E}_0 & \cdots \\ \boldsymbol{F}_3 & \boldsymbol{E}_3 & \boldsymbol{E}_2 & \boldsymbol{E}_1 & \cdots \\ \boldsymbol{F}_4 & \boldsymbol{E}_4 & \boldsymbol{E}_3 & \boldsymbol{E}_2 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \tag{1.8}$$

it is called a Markov chain of G/M/1 type. Further, a bivariate Markov chain has a transition probability matrix with the same structure as

$$\boldsymbol{P}_{\mathrm{L}} = \begin{bmatrix} \boldsymbol{B}_0 & \boldsymbol{B}_1 & \boldsymbol{B}_2 & \boldsymbol{B}_3 & \cdots \\ \widehat{\boldsymbol{A}}_0 & \boldsymbol{A}_1 & \boldsymbol{A}_2 & \boldsymbol{A}_3 & \cdots \\ \boldsymbol{O} & \boldsymbol{A}_0 & \boldsymbol{A}_1 & \boldsymbol{A}_2 & \cdots \\ \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{A}_0 & \boldsymbol{A}_1 & \cdots \\ \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{A}_0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \tag{1.9}$$

it is called a Markov chain of M/G/1 type. These two Markov chains can be considered as extensions of the Markov chains describing the queue length processes in the GI/M/1 queue and the M/G/1 queue, respectively, in the sense that each element of a transition probability matrix is replaced by a block matrix. This is the origin of their names.

Remark 1.1 *Some researchers have studied G/M/1-type and M/G/1-type Markov chains with a tree structure. For G/M/1-type Markov chains with a tree structure, matrix product-form solutions are obtained by Yeung and Sengupta [Yeun94]. As for M/G/1-type Markov chains with a tree structure, He [He00a, He00b, He03a] studied conditions for them to be positive recurrent, null recurrent or transient. A related work is found in [He03b], which studied sufficient conditions for stability or instability of a single-server queue with LCFS preemptive-repeat service discipline. Further, Takine et al. [Taki95] established a paradigm to compute the steady-state solution of the M/G/1-type Markov chain with a tree structure under the assumption that it is irreducible and positive recurrent. That paradigm is an extension of the M/G/1 paradigm [Neut89].*

1.2.1 Markov chains of G/M/1 type

In this subsection, we consider a G/M/1-type Markov chain $\{(X_k^-, J_k^-); k = 0, 1, \ldots\}$ with a transition probability matrix $\boldsymbol{P}_{\mathrm{R}}$ in (1.8), where $\boldsymbol{E}_n$, $\boldsymbol{F}_n$ $(n = 0, 1, \ldots)$ and $\widehat{\boldsymbol{E}}_0$ denotes $M_1^- \times M_1^-$, $M_{\min(n,1)}^- \times M_0^-$ and $M_0^- \times M_1^-$ matrices, respectively.

We define $\boldsymbol{E}$ as

$$\boldsymbol{E} = \sum_{n=0}^{\infty} \boldsymbol{E}_n.$$

Note here that $\boldsymbol{E}$ is a stochastic matrix.

We now assume the following:

Assumption 1.1 *$\boldsymbol{E}$ is irreducible.*

We then define $\boldsymbol{\nu}_E$ as a $1 \times M_1^-$ vector satisfying

$$\boldsymbol{\nu}_E \boldsymbol{E} = \boldsymbol{\nu}_E, \qquad \boldsymbol{\nu}_E \boldsymbol{e} = 1.$$

Note that Assumption 1.1 ensures that $\boldsymbol{\nu}_E$ is uniquely determined. We also make the following assumption:

Assumption 1.2

(a) The G/M/1-type Markov chain $\{(X_k^-, J_k^-); k = 0, 1, \ldots\}$ is irreducible, and

(b) $\boldsymbol{\nu}_E \sum_{n=1}^{\infty} n \boldsymbol{E}_n \boldsymbol{e} > 1.$

Remark 1.2 *The irreducible G/M/1-type Markov chain $\{(X_k^-, J_k^-); k = 0, 1, \ldots\}$ is positive recurrent if and only if Assumption 1.4 (b) holds (See Corollary 3.3 on page 319 in [Asmu03]).*

We define X^- and J^- as generic random variables for X_k^- and J_k^-, respectively, in steady sate. Let $\boldsymbol{x}^- = (\boldsymbol{x}_0^-, \boldsymbol{x}_1^-, \ldots)$ denote the steady-sate solution for the G/M/1-type Markov chain $\{(X_k^-, J_k^-); k = 0, 1, \ldots\}$, where $\boldsymbol{x}_n^-$ denotes a $1 \times M^-_{\min(n,1)}$ vector whose jth element represents $\Pr[X^- = n, J^- = j]$. Let $\boldsymbol{R}$ denote the minimal nonnegative solution [Neut81] of

$$\boldsymbol{R} = \sum_{n=0}^{\infty} \boldsymbol{R}^n \boldsymbol{E}_n.$$

$\boldsymbol{R}$ is called a *rate matrix* and has the following probabilistic meanings [Neut81]. We consider a cycle of visits to the state set $\{(n, \nu); \nu = 1, \ldots, M_1^-\}$ $(n \geq 1)$. The (i, j)th element of $\boldsymbol{R}$ then represents the mean number of visits on state $(n+1, j)$ during a cycle that starts with state (n, i). In terms of rate matrix $\boldsymbol{R}$, $\boldsymbol{x}^- = (\boldsymbol{x}_0^-, \boldsymbol{x}_1^-, \ldots)$ is given in the following way.

Proposition 1.1 ([Neut81]) *$\boldsymbol{x}_0^-$ and $\boldsymbol{x}_1^-$ are determined by*

$$\begin{aligned}
\boldsymbol{x}_0^- &= \boldsymbol{x}_0^- \boldsymbol{F}_0 + \boldsymbol{x}_1^- \sum_{m=1}^{\infty} \boldsymbol{R}^{m-1} \boldsymbol{F}_m, \\
\boldsymbol{x}_1^- &= \boldsymbol{x}_0^- \widehat{\boldsymbol{E}}_0 + \boldsymbol{x}_1^- \sum_{m=1}^{\infty} \boldsymbol{R}^{m-1} \boldsymbol{E}_m, \\
\boldsymbol{x}_0^- \boldsymbol{e} &+ \boldsymbol{x}_1^- (\boldsymbol{I} - \boldsymbol{R})^{-1} \boldsymbol{e} = 1,
\end{aligned}$$

and for $n = 2, 3, \ldots$,

$$\boldsymbol{x}_n^- = \boldsymbol{x}_1^- \boldsymbol{R}^{n-1}. \tag{1.10}$$

The $\boldsymbol{x}_n^-$ can be considered as a matrix version of the geometric solution (1.5) for the GI/M/1 queue, and it is called the *matrix geometric* solution.

1.2.2 Markov chains of M/G/1 type

In this subsection, we consider an M/G/1-type Markov chain $\{(X_k^+, J_k^+); k = 0, 1, \ldots\}$ with a transition probability matrix $\boldsymbol{P}_{\mathrm{L}}$ in (1.9), where $\boldsymbol{A}_n$, $\boldsymbol{B}_n$ $(n = 0, 1, \ldots)$ and $\widehat{\boldsymbol{A}}_0$ denotes $M_1^+ \times M_1^+$, $M_0^+ \times M^+_{\min(n,1)}$ and $M_1^+ \times M_0^+$ matrices, respectively. We define $\boldsymbol{\mu}_A$ and $\boldsymbol{\mu}_B$ as

$$\boldsymbol{\mu}_A = \sum_{n=1}^{\infty} n \boldsymbol{A}_n \boldsymbol{e}, \qquad \boldsymbol{\mu}_B = \sum_{n=1}^{\infty} n \boldsymbol{B}_n \boldsymbol{e},$$

respectively. We also define $\boldsymbol{A}$ and $\boldsymbol{B}$ as

$$\boldsymbol{A} = \sum_{n=0}^{\infty} \boldsymbol{A}_n, \qquad \boldsymbol{B} = \sum_{n=1}^{\infty} \boldsymbol{B}_n,$$

respectively. Note that $\boldsymbol{A}$ is a stochastic matrix and $\boldsymbol{B}_0 \boldsymbol{e} + \boldsymbol{B}\boldsymbol{e} = \boldsymbol{e}$.

We now assume the following:

Assumption 1.3 *$\boldsymbol{A}$ is irreducible.*

Let $\boldsymbol{\nu}_A$ denote a $1 \times M_1^+$ vector satisfying

$$\boldsymbol{\nu}_A \boldsymbol{A} = \boldsymbol{\nu}_A, \qquad \boldsymbol{\nu}_A \boldsymbol{e} = 1.$$

Note that Assumption 1.3 ensures that $\boldsymbol{\nu}_A$ is uniquely determined. We define ρ_A as

$$\rho_A = \boldsymbol{\nu}_A \boldsymbol{\mu}_A.$$

In what follows, we also make the following assumption:

Assumption 1.4

(a) The M/G/1-type Markov chain $\{(X_k^+, J_k^+); k = 0, 1, \ldots\}$ *is irreducible,*

(b) $\boldsymbol{\mu}_B$ *is a finite vector, and*

(c) $\rho_A < 1$.

Remark 1.3 *The irreducible M/G/1-type Markov chain* $\{(X_k^+, J_k^+); k = 0, 1, \ldots\}$ *is positive recurrent if and only if Assumption 1.4 (b) and (c) hold (See Proposition 3.1 on page 318 in [Asmu03]).*

We define X^+ and J^+ as generic random variables for X_k^+ and J_k^+, respectively, in steady state. Let $\boldsymbol{x}^+ = (\boldsymbol{x}_0^+, \boldsymbol{x}_1^+, \ldots)$ denote the steady-state solution for the M/G/1-type Markov chain $\{(X_k^+, J_k^+); k = 0, 1, \ldots\}$, where $\boldsymbol{x}_n^+$ denotes a $1 \times M_{\min(n,1)}^+$ vector whose jth element represents $\Pr[X^+ = n, J^+ = j]$. The numerical algorithm to compute the steady-state solution $\boldsymbol{x}^+ = (\boldsymbol{x}_0^+, \boldsymbol{x}_1^+, \ldots)$ is called the M/G/1 paradigm, where an $M_1^+ \times M_1^+$ matrix $\boldsymbol{G}$ plays a vital role [Neut89]. The (i, j)th element of $\boldsymbol{G}$ represents the probability that for $n \geq 2$, the first passage time from state (n, i) to any state in $\{(n-1, 1), \ldots, (n-1, M_1^+)\}$ ends with state $(n-1, j)$. The $\boldsymbol{G}$ is often called the fundamental matrix. It is known that $\boldsymbol{G}$ is the minimal nonnegative solution of [Neut89]

$$\boldsymbol{G} = \sum_{n=0}^{\infty} \boldsymbol{A}_n \boldsymbol{G}^n.$$

The computation of $\boldsymbol{G}$ is referred to [Neut89, Luca91]. In terms of $\boldsymbol{G}$, we define $M_0^+ \times M_0^+$ matrix $\boldsymbol{K}$ as

$$\boldsymbol{K} = \boldsymbol{B}_0 + \sum_{m=1}^{\infty} \boldsymbol{B}_m \boldsymbol{G}^{m-1} \left[\boldsymbol{I} - \sum_{l=1}^{\infty} \boldsymbol{A}_l \boldsymbol{G}^{l-1} \right]^{-1} \widehat{\boldsymbol{A}}_0. \tag{1.11}$$

Further, let $\boldsymbol{\kappa}$ denote a $1 \times M_0^+$ vector satisfying

$$\boldsymbol{\kappa} \boldsymbol{K} = \boldsymbol{\kappa}, \qquad \boldsymbol{\kappa} \boldsymbol{e} = 1.$$

We now show a recursive formula for the computation of $\boldsymbol{x}_n^+$'s $(n = 0, 1, \ldots)$.

Proposition 1.2 ([Rama88, Sche90]) $\boldsymbol{x}_0^+$ *is given by*

$$\boldsymbol{x}_0^+ = \left[1 + \frac{\boldsymbol{\kappa}}{1-\rho_A}\left\{\boldsymbol{\mu}_B + \left(\boldsymbol{B} - \sum_{l=1}^{\infty}\boldsymbol{B}_l\boldsymbol{G}^l\right)[\boldsymbol{I} - \boldsymbol{A} + \boldsymbol{e}\boldsymbol{\nu}_A]^{-1}\boldsymbol{\mu}_A\right\}\right]^{-1}\boldsymbol{\kappa},$$

and $\boldsymbol{x}_n^+$*'s* $(n = 1, 2, \ldots)$ *are recursively determined by*

$$\boldsymbol{x}_n^+ = \left[\boldsymbol{x}_0^+\overline{\boldsymbol{B}}_n + \sum_{l=1}^{n-1}\boldsymbol{x}_l^+\overline{\boldsymbol{A}}_{n-l+1}\right][\boldsymbol{I} - \overline{\boldsymbol{A}}_1]^{-1},$$

where

$$\overline{\boldsymbol{B}}_\nu = \sum_{m=0}^{\infty}\boldsymbol{B}_{m+\nu}\boldsymbol{G}^m, \qquad \overline{\boldsymbol{A}}_\nu = \sum_{m=0}^{\infty}\boldsymbol{A}_{m+\nu}\boldsymbol{G}^m.$$

1.3 Markov Processes Skip Free to One Direction

This section considers time-homogeneous Markov processes with the skip free to one direction and spatial homogeneity properties, which are continuous analogs of Markov chains discussed in section 1.2. These Markov processes are also classified into two cases, i.e., the skip free to the left case and to the right case, and they are very important models to analyze the waiting time distribution and the virtual waiting time distribution in single-server queues.

1.3.1 Markov processes skip free to the right

This subsection considers a bivariate Markov process $\{(X_{\mathrm{R}}(t), J_{\mathrm{R}}(t)); t \geq 0\}$ with the skip free to the right property. We formally define it as follows:

(a) $X_{\mathrm{R}}(t)$ takes values on $[0, \infty)$ and $J_{\mathrm{R}}(t)$ takes integer values on $\mathcal{M} = \{1, \ldots, M\}$. Further, $X_{\mathrm{R}}(t)$ is skip free to the right and increases at a liner rate of one if there are no downward jumps.

(b) Given $(X_{\mathrm{R}}(t), J_{\mathrm{R}}(t)) = (x, i)$ $(x > 0, i \in \mathcal{M})$, the bivariate Markov process can change its state to somewhere between $(x - y, j)$ and $(x - y + dy, j)$ $(0 \leq y < x, j \in \mathcal{M})$ at a rate $dE_{i,j}(y)$. Let $\boldsymbol{E}(y)$ denote an $M \times M$ matrix whose (i, j)th element is given by $E_{i,j}(y)$. To avoid trivialities, $E_{i,i}(0) = 0$ $(\forall i \in \mathcal{M})$ is assumed. Further, $\boldsymbol{E}(\infty)$ is assumed to be an irreducible matrix.

(c) Given $(X_{\mathrm{R}}(t), J_{\mathrm{R}}(t)) = (x, i)$ $(x > 0, i \in \mathcal{M})$, the bivariate Markov process can change its state to state $(0, j)$ $(j \in \mathcal{M})$ at rate $F_{\mathrm{jump},i,j}(x)$. Let $\boldsymbol{F}_{\mathrm{jump}}(x)$ denote an $M \times M$ matrix whose (i, j)th element is given by $F_{\mathrm{jump},i,j}(x)$. It is assumed that $\boldsymbol{F}_{\mathrm{jump}}(\infty) = \boldsymbol{O}$.

We define $\boldsymbol{E}_{\mathrm{jump}}(x)$ $(x \geq 0)$ and $\boldsymbol{E}_{\mathrm{jump}}(\infty)$ as

$$\begin{aligned}\boldsymbol{E}_{\mathrm{jump}}(x) &= \boldsymbol{E}(x) - \boldsymbol{E}(0),\\ \boldsymbol{E}_{\mathrm{jump}} &= \boldsymbol{E}(\infty) - \boldsymbol{E}(0),\end{aligned}$$

respectively. Note that these two matrices are nonnegative. We also define $\boldsymbol{E}_{\rm slope}$ as an $M \times M$ matrix whose (i,j)th element $\boldsymbol{E}_{{\rm slope},i,j}$ is given by

$$\boldsymbol{E}_{{\rm slope},i,j} = \begin{cases} E_{i,j}(0), & \text{if } i \neq j, \\ -\sum_{\nu \in \mathcal{M}} E_{i,\nu}(\infty), & \text{otherwise.} \end{cases}$$

Note that $\boldsymbol{E}_{\rm slope} + \boldsymbol{E}_{\rm jump}$ is the infinitesimal generator of an irreducible Markov chain. Thus, $\{(X_{\rm R}(t), J_{\rm R}(t)); t \geq 0\}$ can be characterized by $\boldsymbol{E}_{\rm slope}$, $\boldsymbol{E}_{\rm jump}(x)$ and $\boldsymbol{F}_{\rm jump}(x)$ as shown in Figure 1.1

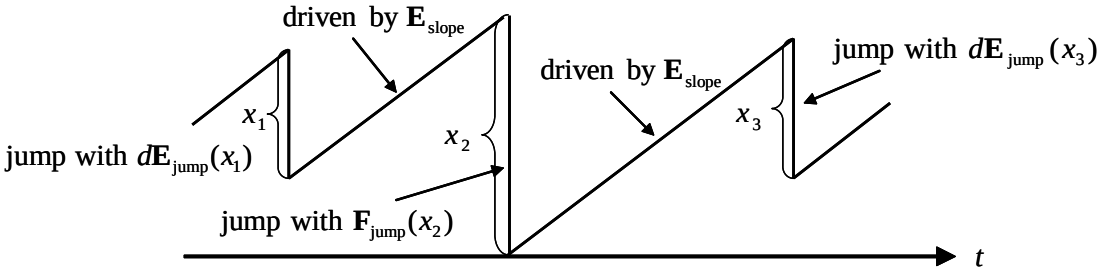

Figure 1.1: The bivariate Markov process with the skip free to the right property.

Proposition 1.3 [Seng89] *A necessary and sufficient condition for the irreducible Markov process* $\{(X_{\rm R}(t), J_{\rm R}(t)); t \geq 0\}$ *to be positive recurrent is*

$$\boldsymbol{\eta}_E \int_0^\infty x d\boldsymbol{E}_{\rm jump}(x) \boldsymbol{e} > 1,$$

where $\boldsymbol{\eta}_E$ *denotes the* $1 \times M$ *invariant probability vector for* $\boldsymbol{E}_{\rm slope} + \boldsymbol{E}_{\rm jump}$.

In the rest of this subsection, we assume that $\{(X_{\rm R}(t), J_{\rm R}(t)); t \geq 0\}$ is positive recurrent and discuss the steady-state density of $\{(X_{\rm R}(t), J_{\rm R}(t)); t \geq 0\}$.

Let $\boldsymbol{\pi}_{\rm R}(x)$ denote a $1 \times M$ vector whose jth element $\pi_{{\rm R},j}(x)$ represents

$$\pi_{{\rm R},j}(x)dx = \lim_{t \to \infty} \frac{1}{t} \int_0^t 1\{x \leq X_{\rm R}(\tau) \leq x + dx, J_{\rm R}(\tau) = j\} d\tau,$$

where $1\{\chi\}$ denotes an indicator function of event χ Then, standard flow conservation arguments yield

$$\boldsymbol{\pi}_{\rm R}(x) = \boldsymbol{\pi}_{\rm R}(x - \Delta x)(\boldsymbol{I} + \boldsymbol{E}_{\rm slope}\Delta x) + \int_0^\infty \boldsymbol{\pi}_{\rm R}(x+u) d\boldsymbol{E}_{\rm jump}(u)\Delta x + o(\Delta x), \qquad x > 0, \tag{1.12}$$

and

$$\boldsymbol{\pi}_{\rm R}(0) = \int_0^\infty \boldsymbol{\pi}_{\rm R}(u) d\boldsymbol{F}_{\rm jump}(u).$$

Rearranging terms in (1.12), dividing both sides by Δx and taking their limits as $\Delta x \to 0$, we have

$$\frac{d}{dx}\boldsymbol{\pi}_{\rm R}(x) = \boldsymbol{\pi}_{\rm R}(x)\boldsymbol{E}_{\rm slope} + \int_0^\infty \boldsymbol{\pi}_{\rm R}(x+u) d\boldsymbol{E}_{\rm jump}(u), \qquad x > 0.$$

Further, from the total probability law, we have

$$\int_0^\infty \boldsymbol{\pi}_{\mathrm{R}}(x)\boldsymbol{e}\,dx = 1.$$

We now introduce an $M \times M$ irreducible matrix $\boldsymbol{\Theta}$, which characterizes $\boldsymbol{\pi}_{\mathrm{R}}(x)$ $(x \geq 0)$. We consider the sequence of matrices $\{\boldsymbol{\Theta}_n; n = 0, 1, \ldots\}$ with $\boldsymbol{\Theta}_0 = \boldsymbol{E}_{\mathrm{slope}}$ and

$$\boldsymbol{\Theta}_n = \boldsymbol{E}_{\mathrm{slope}} + \int_0^\infty \exp(\boldsymbol{\Theta}_{n-1}x)d\boldsymbol{E}_{\mathrm{jump}}(x), \qquad n = 1, 2, \ldots.$$

It is known [Seng89] that if $\{(X_{\mathrm{R}}(t), J_{\mathrm{R}}(t)); t \geq 0\}$ is positive recurrent, $\boldsymbol{\Theta} = \lim_{n\to\infty} \boldsymbol{\Theta}_n$ exists and the limit matrix $\boldsymbol{\Theta}$ is the minimal solution of

$$\boldsymbol{\Theta} = \boldsymbol{E}_{\mathrm{slope}} + \int_0^\infty \exp(\boldsymbol{\Theta}x)d\boldsymbol{E}_{\mathrm{jump}}(x).$$

Note here that $\boldsymbol{\Theta}$ is non-singular and has strictly negative diagonal elements and nonnegative off-diagonal elements [Seng89].

Proposition 1.4 [Seng89] *The steady-state density $\boldsymbol{\pi}_{\mathrm{R}}(x)$ $(x \geq 0)$ is given by*

$$\boldsymbol{\pi}_{\mathrm{R}}(x) = \boldsymbol{\pi}_{\mathrm{R}}(0)\exp(\boldsymbol{\Theta}x), \qquad x > 0,$$

where $\boldsymbol{\pi}_{\mathrm{R}}(0)$ satisfies

$$\boldsymbol{\pi}_{\mathrm{R}}(0) = \boldsymbol{\pi}_{\mathrm{R}}(0)\int_0^\infty \exp(\boldsymbol{\Theta}x)\boldsymbol{F}_{\mathrm{jump}}(x)dx, \qquad \boldsymbol{\pi}_{\mathrm{R}}(0)(-\boldsymbol{\Theta})^{-1}\boldsymbol{e} = 1.$$

We close this subsection by showing an important application of the above result.

We consider the waiting time distribution in a stationary FIFO GI/PH/1 queue. Service times are assumed to be i.i.d. according to a phase-type distribution with representation $(\boldsymbol{\beta}, \boldsymbol{T})$, where $\boldsymbol{\beta}$ denotes a $1 \times M$ probability vector and $\boldsymbol{T}$ denotes an $M \times M$ irreducible and defective generator. We define $\boldsymbol{\tau}$ as $\boldsymbol{\tau} = -\boldsymbol{T}\boldsymbol{e}$. Let $G(x)$ denote the inter-arrival time distribution. We assume that $G(0+) = 0$. For the stationary GI/PH/1 queue, we define a bivariate process $\{(W(t), S(t)); t \geq 0\}$, where $W(t)$ and $S(t)$ denotes the attained waiting time and the phase of service, respectively, at time t. If no customers are present in the system at time t, $W(t)$ and $S(t)$ are set to zero. Note that the $\{W(t); t \geq 0\}$ increases at a rate of one as long as a customer is in service and jumps downward when a customer in the server departs from the system. The amount of the downward jump is equal to the inter-arrival time between the departing customer and the customer served next to him, unless the departure corresponds to the end of a busy period. Otherwise, $W(t)$ jumps to zero.

We now construct another process by deleting all times when $W(t)$ and $J(t)$ are equal to zero and gluing together all non-zero realizations of $\{(W(t), S(t)); t \geq 0\}$. It is easy to see that the resulting process is equivalent to $\{(X_{\mathrm{R}}(t), J_{\mathrm{R}}(t)); t \geq 0\}$ with

$$\begin{aligned}
\boldsymbol{E}_{\mathrm{slope}} &= \boldsymbol{T}, \\
\boldsymbol{E}_{\mathrm{jump}}(x) &= \boldsymbol{\tau}\boldsymbol{\beta}G(x), \qquad (1.13) \\
\boldsymbol{F}_{\mathrm{jump}}(x) &= \boldsymbol{\tau}\boldsymbol{\beta}(1 - G(x)), \qquad (1.14)
\end{aligned}$$

where $\widetilde{\boldsymbol{T}}$ denote a diagonal matrix whose jth diagonal element is given by the (j,j)th element of $\boldsymbol{T}$. Thus, from Proposition 1.4, the steady-state density $\boldsymbol{\pi}_{\rm R}(x)$ $(x>0)$ is given by

$$\boldsymbol{\pi}_{\rm R}(x) = -\boldsymbol{\alpha}\boldsymbol{\Theta}\exp(\boldsymbol{\Theta}x), \qquad x>0, \tag{1.15}$$

where $\boldsymbol{\alpha}$ is a invariant probability vector for $(\boldsymbol{T}+\boldsymbol{\tau}\boldsymbol{\beta})$ and $\boldsymbol{\Theta}$ satisfies

$$\boldsymbol{\Theta} = \boldsymbol{T} + \int_0^\infty \exp(\boldsymbol{\Theta}x)dG(x)\boldsymbol{\tau}\boldsymbol{\beta}. \tag{1.16}$$

Let t_k denote the time when the kth downward jump happens, i.e., the kth customer departs from the system. Note that the waiting time of $(k+1)$th customer is equivalent to $X_{\rm R}(t_k+)$. It is clear from (1.13) and (1.14) that $X_{\rm R}(t_k+)$ and $J_{\rm R}(t_k+)$ are independent. Thus, the waiting time distribution $W_q(x)$ is given by

$$W_q(x) = \int_0^\infty \left(\int_0^{x+u} \frac{\boldsymbol{\pi}_{\rm R}(y)\boldsymbol{\tau}}{\boldsymbol{\alpha}\boldsymbol{\tau}} dy \right) dG(u), \tag{1.17}$$

Substituting (1.15) and (1.16) into (1.17), we obtain

$$W_q(x) = 1 - \frac{1}{\mu}\exp(\boldsymbol{\Theta}x)(\boldsymbol{\Theta}-\boldsymbol{T})\boldsymbol{e},$$

where $\mu = \boldsymbol{\alpha}\boldsymbol{\tau}$ is the reciprocal of the mean service time.

1.3.2 Markov processes skip free to the left

In this subsection, we consider a bivariate Markov process $\{(X_{\rm L}(t), J_{\rm L}(t)); t \geq 0\}$ with the skip free to the left property, which is formally defined as follows:

(a) $X_{\rm L}(t)$ takes values on $[0,\infty)$ and $J_{\rm L}(t)$ takes integer values on $\mathcal{M} = \{1,\ldots,M\}$. Further, $X_{\rm L}(t)$ is skip free to the left and decreases at a liner rate of one if there are no upward jumps.

(b) Given $(X_{\rm L}(t), J_{\rm L}(t)) = (x,i)$ $(x>0, i \in \mathcal{M})$, the bivariate Markov process can change its state to somewhere between $(x+y,j)$ and $(x+y+dy,j)$ $(y \geq 0, j \in \mathcal{M})$ at a rate $dA_{i,j}(y)$. Let $\boldsymbol{A}(y)$ denote an $M \times M$ matrix whose (i,j)th element is given by $A_{i,j}(y)$. To avoid trivialities, $A_{i,i}(0) = 0$ $(\forall i \in \mathcal{M})$ is assumed. Further, $\boldsymbol{A}(\infty)$ is assumed to be an irreducible matrix.

(c) Given $(X_{\rm L}(t), J_{\rm L}(t)) = (0,i)$ $(i \in \mathcal{M})$, the bivariate Markov process can change its state to somewhere between (y,j) and $(y+dy,j)$ $(y \geq 0, j \in \mathcal{M})$ with probability $dB_{{\rm jump},i,j}(y)$. Let $\boldsymbol{B}_{\rm jump}(y)$ denote an $M \times M$ matrix whose (i,j)th element is given by $B_{{\rm jump},i,j}(y)$, where $\boldsymbol{B}_{\rm jump}(\infty)$ is assumed to be a stochastic matrix. This means that an upward jump happens with probability one when $X_{\rm L}(t) = 0$.

For later use, we introduce some matrices. We define $\boldsymbol{A}_{\rm jump}(x)$ $(x \geq 0)$ and $\boldsymbol{A}_{\rm jump}$ as

$$\begin{aligned} \boldsymbol{A}_{\rm jump}(x) &= \boldsymbol{A}(x) - \boldsymbol{A}(0), \\ \boldsymbol{A}_{\rm jump} &= \boldsymbol{A}(\infty) - \boldsymbol{A}(0), \end{aligned}$$

respectively. Note that these two matrices are nonnegative. We also define $\boldsymbol{A}_{\rm slope}$ as an $M \times M$ matrix whose (i,j)th element $\boldsymbol{A}_{{\rm slope},i,j}$ is given by

$$\boldsymbol{A}_{{\rm slope},i,j} = \begin{cases} A_{i,j}(0), & \text{if } i \neq j, \\ -\sum_{\nu \in \mathcal{M}} A_{i,\nu}(\infty), & \text{otherwise.} \end{cases}$$

Note that $\boldsymbol{A}_{\rm slope}$ has strictly negative diagonal elements and nonnegative off-diagonal elements. Note also that $\boldsymbol{A}_{\rm slope} + \boldsymbol{A}_{\rm jump}$ is an irreducible generator of a Markov chain and hence has the invariant probability vector $\boldsymbol{\eta}_A$. Thus, $\{(X_{\rm L}(t), J_{\rm L}(t)); t \geq 0\}$ can be characterized by $\boldsymbol{A}_{\rm slope}$, $\boldsymbol{A}_{\rm jump}(x)$ and $\boldsymbol{B}_{\rm jump}(x)$, as shown in Figure 1.2.

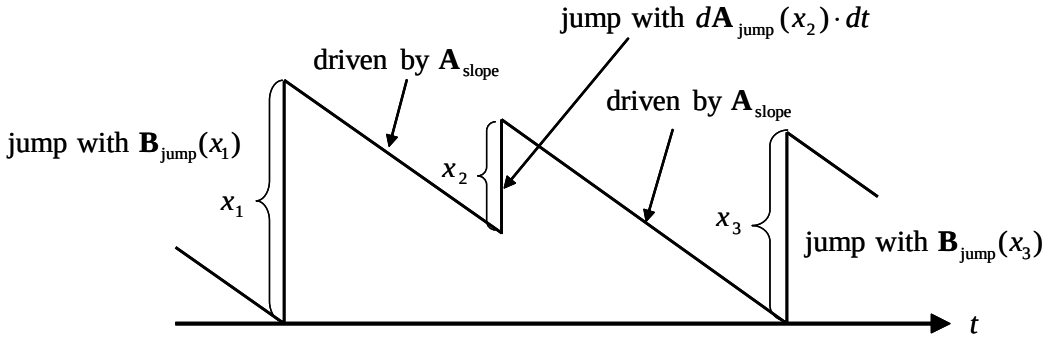

Figure 1.2: The bivariate Markov process with the skip free to the left property.

In what follows, we discuss the steady-state behavior of $\{(X_{\rm L}(t), J_{\rm L}(t)); t \geq 0\}$. To do so, we first show a necessary and sufficient condition for its positive recurrence.

Proposition 1.5 [Taki96] *A necessary and sufficient condition for the irreducible Markov process* $\{(X_{\rm L}(t), J_{\rm L}(t)); t \geq 0\}$ *to be positive recurrent is*

$$\boldsymbol{\eta}_A \overline{\boldsymbol{a}}_{\rm jump} < 1, \quad \textit{and} \quad \left[\int_0^\infty x d\boldsymbol{B}_{\rm jump}(x)\right]_i < \infty \ (i \in \mathcal{M}),$$

where

$$\overline{\boldsymbol{a}}_{\rm jump} = \int_0^\infty x d\boldsymbol{A}_{\rm jump}(x)\boldsymbol{e}.$$

Hereafter, we assume that $\{(X_{\rm L}(t), J_{\rm L}(t)); t \geq 0\}$ is positive recurrent. Let $\boldsymbol{\pi}_{\rm L}(x)$ denote a $1 \times M$ vector whose jth element $\pi_{{\rm L},j}(x)$ represents

$$\pi_{{\rm L},j}(x)dx = \lim_{t\to\infty} \frac{1}{t} \int_0^t 1\{x \leq X_{\rm L}(\tau) \leq x + dx, J_{\rm L}(\tau) = j\} d\tau.$$

We define $\boldsymbol{\Pi}_{\rm L}(x)$ as

$$\boldsymbol{\Pi}_{\rm L}(x) = \int_0^x \boldsymbol{\pi}_{\rm L}(y)dy.$$

Then, standard flow conservation arguments yield

$$\begin{aligned}\boldsymbol{\Pi}_{\rm L}(x) &= [\boldsymbol{\Pi}_{\rm L}(x+\Delta x)-\boldsymbol{\Pi}_{\rm L}(\Delta x)]\,(\boldsymbol{I}+\boldsymbol{A}_{\rm slope}\Delta x)+\int_0^x \boldsymbol{\Pi}_{\rm L}(x+\Delta x-y)d\boldsymbol{A}_{\rm jump}(y)\Delta x \\ &\quad +\boldsymbol{\pi}_{\rm L}(0)\Delta x\boldsymbol{B}_{\rm jump}(x)+o(\Delta x). \end{aligned} \tag{1.18}$$

Rearranging terms in (1.18), dividing both sides by Δx and taking their limits as $\Delta x \to 0$, we have

$$\frac{d}{dx}\boldsymbol{\Pi}_{\rm L}(x)+\boldsymbol{\Pi}_{\rm L}(x)\boldsymbol{A}_{\rm slope}+\int_0^x \boldsymbol{\Pi}_{\rm L}(x-y)d\boldsymbol{A}_{\rm jump}(y)-\boldsymbol{\pi}_{\rm L}(0)(\boldsymbol{I}-\boldsymbol{B}_{\rm jump}(x))=\boldsymbol{0}.$$

Finally, taking the Laplace-Stielties transforms (LSTs) of both sides of the above equation and noting $s\boldsymbol{I}+\boldsymbol{A}_{\rm slope}+\boldsymbol{A}^*_{\rm jump}(s)$ is non-singular for $\mathrm{Re}(s)>0$ [Taki94a], we obtain the following result.

Proposition 1.6 [Taki96] *The LST $\boldsymbol{\pi}^*_{\rm L}(s)$ of $\boldsymbol{\pi}_{\rm L}(x)$ is given by*

$$\boldsymbol{\pi}^*_{\rm L}(s)=\boldsymbol{\pi}_{\rm L}(0)[\boldsymbol{I}-\boldsymbol{B}^*_{\rm jump}(s)][s\boldsymbol{I}+\boldsymbol{A}_{\rm slope}+\boldsymbol{A}^*_{\rm jump}(s)]^{-1}, \qquad \mathrm{Re}(s)>0, \tag{1.19}$$

where $\boldsymbol{A}^*_{\rm jump}(s)$ and $\boldsymbol{B}^*_{\rm jump}(s)$ denote the LSTs of $\boldsymbol{A}_{\rm jump}(x)$ and $\boldsymbol{B}_{\rm jump}(x)$, respectively.

Next, we determine the boundary vector $\boldsymbol{\pi}_{\rm L}(0)$. To do so, we introduce some notations. Let $\boldsymbol{Q}_{\rm L}$ denote an $M\times M$ irreducible matrix representing an infinitesimal generator of a certain Markov chain, which satisfies

$$\boldsymbol{Q}_{\rm L}=\boldsymbol{A}_{\rm slope}+\int_0^\infty d\boldsymbol{A}_{\rm jump}(x)e^{\boldsymbol{Q}_{\rm L}x}.$$

Let $\boldsymbol{\kappa}_{\rm L}$ denote a $1\times M$ probability vector satisfying $\boldsymbol{\kappa}_{\rm L}\boldsymbol{Q}_{\rm L}=\boldsymbol{0}$. We define $\boldsymbol{r}(x)$ as an $M\times 1$ vector whose jth element represents the mean first passage time from state (x,i) to the set of states $\{(0,j); j\in\mathcal{M}\}$. Then $\boldsymbol{r}(x)$ is given by [Taki94a]

$$\boldsymbol{r}(x)=(x\boldsymbol{e}\boldsymbol{\kappa}_{\rm L}-e^{\boldsymbol{Q}_{\rm L}x}+\boldsymbol{I})\,[(\boldsymbol{e}-\overline{\boldsymbol{a}}_{\rm jump})\,\boldsymbol{\kappa}_{\rm L}-\boldsymbol{A}_{\rm slope}-\boldsymbol{A}_{\rm jump}]^{-1}\,\boldsymbol{e}.$$

Proposition 1.7 [Taki96] *The boundary vector $\boldsymbol{\pi}_{\rm L}(0)$ is given by*

$$\boldsymbol{\pi}_{\rm L}(0)=c\cdot\boldsymbol{\pi}_{\rm L,0},$$

where the $1\times M$ vector $\boldsymbol{\pi}_{\rm L,0}$ and the constant c are uniquely determined by

$$\boldsymbol{\pi}_{\rm L,0}=\boldsymbol{\pi}_{\rm L,0}\int_0^\infty d\boldsymbol{B}_{\rm jump}(x)\exp(\boldsymbol{Q}_{\rm L}x), \qquad \boldsymbol{\pi}_{\rm L,0}\boldsymbol{e}=1,$$

and

$$c=\frac{1}{\boldsymbol{\pi}_{\rm L,0}\displaystyle\int_0^\infty d\boldsymbol{B}_{\rm jump}(x)\boldsymbol{r}(x)},$$

respectively.

We now provide an important application of the above results. We consider the virtual waiting time (i.e, the amount of unfinished work in the system) in a stationary single-server queue with the input process introduced in subsection 1.1.2. Namely, the input process has K batch Markovian arrival streams, which are characterized by $\boldsymbol{C}$, $\boldsymbol{D}(\boldsymbol{n})$ and K different service time distributions $H_1(x), \ldots, H_K(x)$. We define $\overline{\boldsymbol{D}}(x)$ as

$$\overline{\boldsymbol{D}}(x) = \sum_{\boldsymbol{n} \in \mathcal{Z}} \boldsymbol{D}(\boldsymbol{n}) H_1^{(n_1)} * \cdots * H_K^{(n_K)}(x), \tag{1.20}$$

where $H_k^{(1)}(x) = H_k(x)$ and $H_k^{(n)}(x)$ $(n = 2, 3, \ldots)$ denotes the n-fold convolution of $H_k(x)$ with itself, and where $H * G(x)$ denotes the convolution of two distribution functions $H(x)$ and $G(x)$ $(x \geq 0)$, i.e.,

$$H * G(x) = \int_0^x dH(y) G(x - y).$$

Let $V(t)$ and $S(t)$ denote the virtual waiting time and the state of the underlying Markov chain that governs the input process, respectively, at time t. Let $\boldsymbol{v}(x)$ denote an $M \times M$ vector whose jth element represents

$$\lim_{t \to \infty} \frac{1}{t} \int_0^t 1\{V(\tau) \leq x, S(\tau) = j\} d\tau.$$

Also, let $\boldsymbol{v}^*(s)$ denote the LST of $\boldsymbol{v}(x)$. We derive $\boldsymbol{v}^*(s)$ as follows.

We first construct the censored process obtained by observing $\{(V(t), S(t)); t \geq 0\}$ only when $\{V(t) > 0\}$. This censored process is equivalent to $\{(X_{\rm L}(t), J_{\rm L}(t)); t \geq 0\}$ with

$$\boldsymbol{A}_{\rm slope} = \boldsymbol{C}, \quad \boldsymbol{A}_{\rm jump}(x) = \overline{\boldsymbol{D}}(x), \quad \boldsymbol{B}_{\rm jump}(x) = (-\boldsymbol{C})^{-1} \overline{\boldsymbol{D}}(x).$$

Let $\boldsymbol{\kappa}$ denote the invariant probability vector for $\boldsymbol{Q}$ satisfying

$$\boldsymbol{Q} = \boldsymbol{C} + \int_0^\infty d\overline{\boldsymbol{D}}(x) \exp(\boldsymbol{Q}x).$$

Note that $\boldsymbol{Q}$ can be considered as the irreducible generator of the censored Markov chain obtained by observing the underlying Markov chain only when the server is busy [Taki94a]. It then follows from Propositions 1.6 and 1.7 that

$$\boldsymbol{\pi}_{\rm L}^*(s) = c \boldsymbol{\pi}_{\rm L,0} [\boldsymbol{I} - (-\boldsymbol{C})^{-1} \overline{\boldsymbol{D}}^*(s)][s\boldsymbol{I} + \boldsymbol{C} + \overline{\boldsymbol{D}}^*(s)]^{-1}, \qquad {\rm Re}(s) > 0, \tag{1.21}$$

where

$$\boldsymbol{\pi}_{\rm L,0} = \frac{\boldsymbol{\kappa}(-\boldsymbol{C})}{\boldsymbol{\kappa}(-\boldsymbol{C})\boldsymbol{e}}, \qquad c = \frac{(1-\rho)\boldsymbol{\kappa}(-\boldsymbol{C})\boldsymbol{e}}{\rho}.$$

Note here that $\boldsymbol{\pi}_{\rm L}^*(s)$ in (1.21) represents the conditional LST of the stationary virtual waiting time distribution, given that the server is busy. Note also that the stationary probability vector of the underlying Markov chain provided that the server is idle is given by

$$\boldsymbol{\kappa} = \frac{\boldsymbol{\pi}_{\rm L,0}(-\boldsymbol{C})^{-1}}{\boldsymbol{\pi}_{\rm L,0}(-\boldsymbol{C})^{-1}\boldsymbol{e}}.$$

Further, since the fraction of time when the server is idle (resp. busy) is given by $1-\rho$ (resp. ρ), we obtain

$$\boldsymbol{v}^*(s) = (1-\rho)\boldsymbol{\kappa} + \rho\boldsymbol{\pi}_{\mathrm{L}}^*(s). \tag{1.22}$$

From (1.21) and (1.22), we obtain the following result.

Proposition 1.8 ([Taki94a]) *$\boldsymbol{v}^*(s)$ is given by*

$$\boldsymbol{v}^*(s)\left[s\boldsymbol{I} + \boldsymbol{C} + \overline{\boldsymbol{D}}^*(s)\right] = s(1-\rho)\boldsymbol{\kappa}, \qquad \mathrm{Re}(s) > 0. \tag{1.23}$$

We finally show the recursive formula for derivatives of $\boldsymbol{v}^*(s)$ evaluated at $s = 0+$. We define $\overline{\boldsymbol{v}}^{(n)}$ as

$$\overline{\boldsymbol{v}}^{(n)} = \lim_{s\to 0+} \frac{(-1)^n}{n!}\frac{d^n}{ds^n}\boldsymbol{v}^*(s), \qquad n = 1, 2, \ldots,$$

and $\overline{\boldsymbol{v}}^{(0)} = \boldsymbol{\pi}$, where $\boldsymbol{\pi}$ is the invariant probability vector for $(\boldsymbol{C}+\boldsymbol{D})$.

Proposition 1.9 [Taki94a] *The $\overline{\boldsymbol{v}}^{(n)}$ is recursively determined by the following recursion:*

$$\begin{aligned}
\boldsymbol{x}^{(1)} &= \left[\boldsymbol{\pi}(\overline{\boldsymbol{D}}^{(1)} - \boldsymbol{I}) + (1-\rho)\boldsymbol{\kappa}\right](\boldsymbol{e}\boldsymbol{\pi} - \boldsymbol{C} - \boldsymbol{D})^{-1},\\
\overline{\boldsymbol{v}}^{(1)}\boldsymbol{e} &= \frac{1}{1-\rho}\left[\boldsymbol{\pi}\overline{\boldsymbol{D}}^{(2)}\boldsymbol{e} + \boldsymbol{x}^{(1)}\overline{\boldsymbol{D}}^{(1)}\boldsymbol{e}\right],\\
\overline{\boldsymbol{v}}^{(1)} &= \overline{\boldsymbol{v}}^{(1)}\boldsymbol{e}\boldsymbol{\pi} + \boldsymbol{x}^{(1)},
\end{aligned}$$

and for $n = 2, 3, \ldots$,

$$\begin{aligned}
\boldsymbol{x}^{(n)} &= \left[\overline{\boldsymbol{v}}^{(n-1)}(\overline{\boldsymbol{D}}^{(1)} - \boldsymbol{I}) + \sum_{k=2}^{n}\overline{\boldsymbol{v}}^{(n-k)}\overline{\boldsymbol{D}}^{(k)}\right](\boldsymbol{e}\boldsymbol{\pi} - \boldsymbol{C} - \boldsymbol{D})^{-1},\\
\overline{\boldsymbol{v}}^{(n)}\boldsymbol{e} &= \frac{1}{1-\rho}\left[\sum_{k=2}^{n+1}\overline{\boldsymbol{v}}^{(n+1-k)}\overline{\boldsymbol{D}}^{(k)}\boldsymbol{e} + \boldsymbol{x}^{(n)}\overline{\boldsymbol{D}}^{(1)}\boldsymbol{e}\right],\\
\overline{\boldsymbol{v}}^{(n)} &= \overline{\boldsymbol{v}}^{(n)}\boldsymbol{e}\boldsymbol{\pi} + \boldsymbol{x}^{(n)},
\end{aligned}$$

where $\overline{\boldsymbol{D}}^{(n)}$ is defined as

$$\overline{\boldsymbol{D}}^{(n)} = \lim_{s\to 0+}\frac{(-1)^n}{n!}\frac{d^n}{ds^n}\overline{\boldsymbol{D}}^*(s). \qquad n = 1, 2, \ldots.$$

1.4 Queues with Multiple Arrival Streams

In this section, we survey previous works on queues with multiple non-Poissonian arrival streams having different service time distributions. No papers have analyzed such queues with full generality until Takine and Hasegawa's work [Taki94a], which studied the workload in a MAP/G/1 queue with state-dependent service times. Related studies on the workload can be found in [Asmu91, Zhu91, Mach93, Taki96]. As for the queue length, Regterschot and de Smit [Regt86] derived the probability generating function of the total queue length in the MMPP/G/1-FIFO queue with state-dependent service times. Further, for a MAP/G/1 nonpreemptive priority queue with class-dependent service

times, Takine [Taki99] derived the moments and the mass function of the marginal queue length of each class.

To our best of our knowledge, Takine [Taki01a] performed, for the first time, the full-dress analysis of the joint queue length distribution in queues with multiple non-Poissonian arrival streams having different service time distributions. In [Taki01a], an invariance relationship between the time-average joint queue length distribution and the customer-average joint queue length distribution at departures was shown for queues with the Markovian input process described in subsection 1.1.2. Based on this invariance relationship, a distributional form of Little's law for FIFO queues with simple arrivals (i.e., no batch arrivals happen) was derived. Further, using this law and the result in [Taki94a], recursions to compute the joint queue length distributions in FIFO single-server queues with/without vacations were constructed [Taki01a, Taki01b].

He and Alfa [He98] and Takine [Taki02] studied single-server queues, where customer arrivals follow a MMAP (see Example 1.4) and they are served on a Last-Come-First-Served (LCFS) preemptive basis. In [He98], phase-type service time distributions were assumed. As a result, the queue length process forms a G/M/1-type Markov chain with a tree structure (see Remark 1.1), and hence the stationary joint queue distribution has a matrix product-form solution. On the other hands, in [Taki02], general service time distributions were assumed and the matrix product solution for the stationary joint queue length distribution was also obtained by tracking the dynamics of the string representing classes and remaining service times of respective customers in the system.

As mentioned above, some studies have already been done for single server queues with multiple non-Poissonian arrival streams having different service time distributions. As far as we know, however, no researchers have studied the multi-server case. As for infinite-server queues with non-i.i.d. service times, a few studies have been done in [Smit72, Mach86, Liu91].

1.5 Overview of the Thesis

The main purpose of this thesis is to establish numerical algorithms to compute performance measures (e.g., the queue length distribution and the sojourn time distribution) in queues with batch Markovian arrival streams having different service time distributions.

In chapter 2, we consider a FIFO single-server queue. It had been recognized that the joint queue length distribution for a single-server queue with multiple non-Poissonian arrival streams having different service time distributions is hard to analyze. Recently, for the single arrival case (i.e., no batch arrivals happen), Takine developed recursions to compute the joint queue length distribution by a new approach [Taki01a, Taki01b, Taki01c], which is based on the distributional form of Little's law [Taki01a]. By a very similar approach, we can also establish recursions to compute the joint queue length distribution in the batch arrival case. However, in their implementation, we have to determine several truncations and stopping criteria that are associated with batch arrivals. In addition, the straightforward implementation of those recursions is much more computation-intensive, compared with the single arrival case. The algorithm proposed in chapter 2 solves these problems by efficiently discarding components that make a small contribution to the final results.

Chapter 3 considers a FIFO single-server queue with service interruptions, where the marginal on-off process of the server is a phase-type Markov renewal process, the marginal arrival process is a MMAP (see Example 1.4) with batch arrivals and these two processes are possibly correlated. This model is an extension of one in [Taki97] and it allows batch arrivals and non-i.i.d. service times. To analyze the joint queue length, we first consider a censored process obtained by observing the queue length process only when the server is *on*. This censored process is probabilistically similar to a FIFO single-server queue studied in chapter 2. This fact enables us to establish a numerical algorithm to compute the joint queue length distribution in a very similar way to in chapter 2 [Masu03a].

In chapter 4, we consider an infinite-server queue. Since Takács' work [Taka62], infinite-server queues have been studied by many researchers for forty years. To the best of our knowledge, the infinite-server queue studied by Liu et al. [Liu91] is most general among those studied in the past. In their infinite-server queue, arrival times, batch sizes of arrivals and service time distributions of arriving customers are governed by a Markov renewal process. Although their model is very general, their result is not necessarily suitable for computing performance measures of interest, e.g., the time-dependent mean and variance of the number of customers in the system. Compared with the model studied by Liu et al. [Liu91], the infinite-server queue studied in chapter 4 is more general in the sense that service times in the same batch may be not i.i.d. even though the underlying Markov renewal process is of phase-type. For this infinite-server queue, we derive explicit and numerically feasible formulas to compute the time-dependent and limiting joint binomial moments of the number of customers in the system by considering the time-reversed process of the original queueing process.

Chapter 5 considers processor-sharing queues. In chapter 5, we impose two assumptions on the input process introduced in subsection 1.1.2. First, there exists only one arrival stream. This assumption means that the arrival process is a BMAP (see subsection 1.1.1). Second, service times are i.i.d. according to an exponential distribution. Thus, the processor-sharing queue studied in chapter 5 is denoted by a BMAP/M/1-PS. It is well-known that the stationary queue length distribution in a BMAP/M/1-PS queue are identical to that in a BMAP/M/1-FIFO queue. As for the sojourn time distribution in a BMAP/M/1-PS, however, no studies have been done. In chapter 5, we derive a recursive formula to compute the sojourn time distribution by noting that the sojourn time of a customer is equivalent to the first passage time to an absorbing state in an absorbing Markov chain with a special structure. Further, for a MAP/M/1-PS queue, we develop a numerically feasible algorithm to compute the sojourn time distribution, using the uniformization technique [Tijm94]. The main advantage of the algorithm is that the numerical precision of the result can be given in advance.

Finally, the conclusion and suggestions for future research topics are provided in chapter 6.

The results discussed in chapter 2 is mainly based on [Masu03a], chapter 3 on [Masu03c], chapter 4 on [Masu02] and chapter 5 on [Masu03b].

Chapter 2

FIFO Single-Server Queue

2.1 Introduction

In this chapter, we study the joint queue length distribution in a stationary work-conserving FIFO single-server queue fed by batch Markovian arrival streams. A particular feature of this queue is that service time distributions of customers may be different for different arrival streams.

Most of previous works on FIFO single-server queues with Markovian arrival streams assume that service times of all customers are i.i.d. according to a common distribution function. As a result, the bivariate process of the total number of customers and the state of the Markov chain that governs the arrival process immediately after departures forms a Markov chain of M/G/1 type and the steady-state solution can be computed by well-known M/G/1 paradigm [Neut89].

On the other hand, if service time distributions of customers from respective arrival streams are different from one another, the bivariate process does not have the Markov property [Taki01b], except for queues with a superposition of independent Poisson streams. Thus the queue length analysis of such a queue is not straightforward. Note, however, that the virtual waiting time process in such a queue is characterized by a bivariate Markov process [Asmu91, Taki94a, Taki96, Zhu91], and algorithmic solution methods are known in the literature [Taki94a, Taki96].

Recently, a new approach was developed to characterize the joint queue length distribution in FIFO queues with MMAP (see Example 1.4 in chapter 1) having different service time distributions [Taki01a, Taki01b]. In [Taki01a], the invariant relationship of the joint queue length distributions at a random point in time and at departures was obtained and from this, the distributional form of Little's law was established in [Taki01a]. Further, based on the latter, an algorithmic solution method was developed [Taki01a, Taki01b]. Related works are found in [Mach99, Regt86]. See [Taki01c] for a survey of those developments.

The results in this chapter are considered as an extension of those in [Taki01b], allowing batch arrivals in each arrival stream. Note here that the distributional form of Little's law does not hold for FIFO queues with batch arrivals. Therefore our starting point in analyzing the time-average joint queue length distribution is the invariant relationship of the joint queue length distributions at a random point in time and at departures in [Taki01a]. By doing so, the problem is reduced to find the joint queue length distributions at departures of customers from respective arrival streams.

As you will see, the joint queue length distribution at departures in the FIFO queue is closely related to the virtual waiting time distribution that is readily obtained with the known results. Using these facts, we derive a general formula for the stationary joint queue length distribution at departures in terms of the sojourn time distribution. Further, assuming discrete phase-type batch size distributions, we derive recursions to compute the stationary joint queue length distribution at a random point in time.

The above outline is similar to the single arrival case in [Taki01b]. However, the implementation of some of those recursions is not trivial, because we have to determine several truncation and stopping criteria, which are due to batch arrivals, and their straightforward implementation would require very huge memory space and time-consuming. In this chapter, assuming discrete phase-type batch size distributions, we propose a numerically feasible procedure to compute those recursions, while ensuring the numerical accuracy in the final result. This is the main contribution of this chapter. Note that the procedure is also applicable to FIFO BMAP/G/1 queues with i.i.d. services, when batch sizes follow a discrete phase-type distribution.

The rest of this chapter is divided into six sections. In section 2.2, an input process considered in this chapter is described. In section 2.3, we briefly discuss the virtual and actual waiting time distributions. In section 2.4, we first derive a general formula for the joint queue length distribution, and assuming discrete phase-type batch sizes, we show recursive formulas to compute the joint queue length distribution. In section 2.5, the implementation of the recursions is discussed. In section 2.6, we discuss the efficiency of our algorithm and the qualitative behavior of the queue length through some numerical examples. Finally, concluding remarks are provided in section 2.7.

2.2 Input Process

This chapter imposes the following assumption on the input process introduced in subsection 1.1.2:

Assumption 2.1 *No arrivals from two or more arrival streams happen simultaneously, i.e.,*

$$\begin{aligned} &\boldsymbol{D}(\boldsymbol{n}) \geq \boldsymbol{O}, \quad \text{if } \boldsymbol{n} = n_k \boldsymbol{e}_k,\ n_k = 1, 2, \ldots,\ k \in \mathcal{K}, \\ &\boldsymbol{D}(\boldsymbol{n}) = \boldsymbol{O}, \quad \text{otherwise}, \end{aligned}$$

where $\boldsymbol{e}_k$ is defined in (1.4).

We define $\boldsymbol{D}_k(n)$ $(n = 1, 2, \ldots, k \in \mathcal{K})$ and $\boldsymbol{D}_k$ $(k \in \mathcal{K})$ as

$$\boldsymbol{D}_k(n) = \boldsymbol{D}(n\boldsymbol{e}_k), \qquad \boldsymbol{D}_k = \sum_{n=1}^{\infty} \boldsymbol{D}_k(n),$$

respectively. Then, the counting process of arrivals is characterized by $\boldsymbol{C}$ and $\boldsymbol{D}_k(n)$ $(k \in \mathcal{K}, n = 1, 2, \ldots)$. This arrival process is an extension of MMAP (see Example 1.4 in chapter 1) to allows batch arrivals, i.e., a batch MMAP. Roughly speaking, customers arrive in the following way. When a state transition driven by $\boldsymbol{D}_k(n)$ occurs, n customers of class k arrive simultaneously, and their service times are i.i.d. according to a distribution function $H_k(x)$ with finite mean h_k. On the other

hand, when a state transition driven by $\boldsymbol{C}$ occurs, no customers arrive. Therefore, the input process is characterized by $(\boldsymbol{C}, \boldsymbol{D}_k(n), H_k(x); k \in \mathcal{K})$. Note that the arrival rate λ_k of class k customers is given by

$$\lambda_k = \sum_{n=1}^{\infty} n\boldsymbol{\pi}\boldsymbol{D}_k(n)\boldsymbol{e}.$$

In the remainder of this chapter, we assume that utilization factor $\rho = \sum_{k\in\mathcal{K}} \lambda_k h_k < 1$, which ensures that all customers arriving to the system are eventually served [Loyn62].

2.3 Waiting Time Distribution

In this section, we consider the actual waiting time of class k customers in steady state. We define $W_{q,k}(n;m)$ as a generic random variable representing the actual waiting time of a randomly chosen class k customer who is a member of a batch of size n and the mth served customer among members of the same batch. Let $S^{(\mathrm{A}_k)}(n)$ denote a generic random variable representing the state of the underlying Markov chain immediately after class k batches of size n arrive. With those, we define $\boldsymbol{w}_{q,k}(x\,|\,n;m)$ as a $1 \times M$ vector whose jth element represents $\Pr[W_{q,k}(n;m) \le x, S^{(\mathrm{A}_k)}(n) = j]$. Note that $\boldsymbol{w}_{q,k}(x\,|\,n;1)$ $(n \ge 1)$ is given by [Taki94a]

$$\boldsymbol{w}_{q,k}(x\,|\,n;1) = \frac{\boldsymbol{v}(x)\boldsymbol{D}_k(n)}{\boldsymbol{\pi}\boldsymbol{D}_k(n)\boldsymbol{e}}, \tag{2.1}$$

where $\boldsymbol{v}(x)$ denotes a $1 \times M$ vector whose jth $(j \in \mathcal{M})$ element represents the stationary joint probability that the total unfinished work is not greater than x and the underlying Markov chain is in state j. Further, for $n \ge 2$,

$$\boldsymbol{w}_{q,k}(x\,|\,n;m) = \int_0^x \boldsymbol{w}_{q,k}(x-y\,|\,n;1) dH_k^{(m-1)}(y), \qquad m = 2,3,\ldots,n. \tag{2.2}$$

Let $W_{q,k}$ and $S^{(\mathrm{A}_k)}$ denote generic random variables representing the actual waiting time of class k customers and the state of the underlying Markov chain immediately after arrivals of class k batches, respectively. We then define $\boldsymbol{w}_{q,k}(x)$ as a $1 \times M$ vector whose jth element represents $\Pr[W_{q,k} \le x, S^{(\mathrm{A}_k)} = j]$. Because a randomly chosen customer of class k is a member of a batch of size n with probability $n\boldsymbol{\pi}\boldsymbol{D}_k(n)\boldsymbol{e}/\lambda_k$, we have

$$\boldsymbol{w}_{q,k}(x) = \sum_{n=1}^{\infty} \frac{n\boldsymbol{\pi}\boldsymbol{D}_k(n)\boldsymbol{e}}{\lambda_k} \cdot \frac{1}{n} \sum_{m=1}^{n} \boldsymbol{w}_{q,k}(x\,|\,n;m). \tag{2.3}$$

Let $\boldsymbol{w}^*_{q,k}(s)$ denote the LST of $\boldsymbol{w}_{q,k}(x)$. From (2.1)–(2.3), we have

$$\boldsymbol{w}^*_{q,k}(s) = \sum_{n=1}^{\infty} \frac{\boldsymbol{v}^*(s)\boldsymbol{D}_k(n)}{\lambda_k} \sum_{m=1}^{n} \{H_k^*(s)\}^{m-1}, \qquad \mathrm{Re}(s) > 0, \tag{2.4}$$

where $\boldsymbol{v}^*(s)$ and $H_k^*(s)$ denote the LSTs of $\boldsymbol{v}(x)$ and $H_k(x)$, respectively. Note here that $\boldsymbol{v}^*(s)$ satisfies (1.23), where

$$\overline{\boldsymbol{D}}^*(s) = \sum_{k\in\mathcal{K}} \sum_{n=1}^{\infty} \boldsymbol{D}_k(n)\{H_k^*(s)\}^n. \tag{2.5}$$

From (2.4), we can readily obtain the following result.

Theorem 2.1 *$\boldsymbol{w}^*_{q,k}(s)$ ($k \in \mathcal{K}$) is given by*

$$\boldsymbol{w}^*_{q,k}(s) = \frac{\boldsymbol{v}^*(s)\,(\boldsymbol{D}_k - \boldsymbol{D}^*_k(H^*_k(s)))}{\lambda_k\,(1 - H^*_k(s))}, \qquad \mathrm{Re}(s) > 0,$$

where

$$\boldsymbol{D}^*_k(z_k) = \sum_{n=1}^{\infty} z_k^n \boldsymbol{D}_k(n). \tag{2.6}$$

2.4 Joint Queue Length Distribution

This section considers the joint queue length distribution. In subsection 2.4.1, we apply a general relationship between the time-average queue length distribution and the queue length distributions at departures of customers of respective classes [Taki01a] to our specific queue. Then the problem is reduced to characterize the joint queue length distributions at departures of respective classes, which is discussed in subsection 2.4.2. Finally in subsection 2.4.3, assuming discrete phase-type batch size distributions, we derive recursions for some quantities required in computing the joint queue length distribution.

2.4.1 Relationship in the joint queue length distributions

Let N_k ($k \in \mathcal{K}$) and S denote generic random variables representing the number of class k customers and the state of the underlying Markov chain, respectively, in steady state. We define $\boldsymbol{p}(n_1, \ldots, n_K)$ as a $1 \times M$ vector whose jth element represents $\Pr[N_1 = n_1, \ldots, N_K = n_K, S = j]$. For simplicity, let $\boldsymbol{n}$ and $\boldsymbol{z}$ denote a $1 \times K$ nonnegative integer vector $(n_1, \ldots, n_K)$ and a $1 \times K$ complex vector $(z_1, \ldots, z_K)$, respectively. Further we define $\mathcal{Z}$ as

$$\mathcal{Z} = \{(n_1, \ldots, n_K)\,;\, n_k = 0, 1, \ldots, \text{ for all } k \in \mathcal{K}\}.$$

We then define $\boldsymbol{p}^*(\boldsymbol{z})$ as

$$\boldsymbol{p}^*(\boldsymbol{z}) = \sum_{\boldsymbol{n} \in \mathcal{Z}} z_1^{n_1} \cdots z_K^{n_K} \boldsymbol{p}(\boldsymbol{n}), \qquad |z_k| \le 1 \text{ for all } k \in \mathcal{K}.$$

Note that $\boldsymbol{p}^*(\boldsymbol{z})$ denotes the vector generating function of the joint queue length distribution in steady state.

Let $N_\nu^{(\mathrm{D}_k)}$ and $S^{(\mathrm{D}_k)}$ ($k, \nu \in \mathcal{K}$) denote generic random variables representing the number of class ν customers and the state of the underlying Markov chain, respectively, immediately after departures of class k customers in steady state. We then define $\boldsymbol{q}_k(\boldsymbol{n})$ ($k \in \mathcal{K}$, $\boldsymbol{n} \in \mathcal{Z}$) as a $1 \times M$ vector whose jth element represents $\Pr[N_1^{(\mathrm{D}_k)} = n_1, \ldots, N_K^{(\mathrm{D}_k)} = n_K, S^{(\mathrm{D}_k)} = j]$. Further we define $\boldsymbol{q}^*_k(\boldsymbol{z})$ ($k \in \mathcal{K}$) as

$$\boldsymbol{q}^*_k(\boldsymbol{z}) = \sum_{\boldsymbol{n} \in \mathcal{Z}} z_1^{n_1} \cdots z_K^{n_K} \boldsymbol{q}_k(\boldsymbol{n}), \qquad |z_k| \le 1 \text{ for all } k \in \mathcal{K}.$$

Note that $\boldsymbol{q}^*_k(\boldsymbol{z})$ denotes the vector generating function of the joint queue length distribution immediately after departures of class k customers. Thus, applying Theorem 1 in [Taki01a] to our model, we have the following theorem.

Theorem 2.2 ([Taki01a]) *$\boldsymbol{p}^*(\boldsymbol{z})$ and $\boldsymbol{q}_k^*(\boldsymbol{z})$ are related by*

$$\boldsymbol{p}^*(\boldsymbol{z})\left[\boldsymbol{C}+\sum_{k\in\mathcal{K}}\boldsymbol{D}_k^*(z_k)\right]=\sum_{k\in\mathcal{K}}\lambda_k(z_k-1)\boldsymbol{q}_k^*(\boldsymbol{z}),\quad k\in\mathcal{K}, \tag{2.7}$$

where $\boldsymbol{D}_k^*(z_k)$ is given in (2.6).

Further, comparing the coefficient vectors of $z_1^{n_1}\cdots z_K^{n_K}$ on both sides of (2.7), we obtain the following result.

Corollary 2.1 *The $\boldsymbol{p}(\boldsymbol{n})$ $(\boldsymbol{n}\in\mathcal{Z})$ is recursively determined by the following way.*

$$\boldsymbol{p}(\boldsymbol{0})=\sum_{k\in\mathcal{K}}\lambda_k\boldsymbol{q}_k(\boldsymbol{0})(-\boldsymbol{C})^{-1}, \tag{2.8}$$

and for $\mathcal{Z}^+=\mathcal{Z}-\{\boldsymbol{0}\}$,

$$\boldsymbol{p}(\boldsymbol{n})=\sum_{k\in\mathcal{K}}\left[\lambda_k\left(\boldsymbol{q}_k(\boldsymbol{n})-\boldsymbol{q}_k(\boldsymbol{n}-\boldsymbol{e}_k)\right)+\sum_{m_k=1}^{n_k}\boldsymbol{p}(\boldsymbol{n}-m_k\boldsymbol{e}_k)\boldsymbol{D}_k(m_k)\right](-\boldsymbol{C})^{-1}, \tag{2.9}$$

where $\boldsymbol{q}_k(\boldsymbol{n})=\boldsymbol{0}$ for $\boldsymbol{n}\notin\mathcal{Z}$ and $\boldsymbol{e}_k$ $(k\in\mathcal{K})$ is given in (1.4).

This recursion is an extension of the recursion (B.3) and (B.4) for the BMAP/GI/1 queue.

Remark 2.1 *If the service discipline is independent of the state of the underlying Markov chain, Theorem 2.2 holds for stationary queues fed by Markovian input process $(\boldsymbol{C},\boldsymbol{D}_k(n),H_k(x);k\in\mathcal{K})$, despite the number of servers, the service discipline and the service time distributions of respective classes. The details are referred to [Taki01a].*

2.4.2 Joint queue length distribution immediately after departures

In this subsection, we consider the vector generating function of the joint queue length distribution immediately after departures of each class. We denote, by $\mathrm{C}_k(n;m)$ $(k\in\mathcal{K},\ n=1,2,\ldots,\ m=1,2,\ldots,n)$, a randomly chosen class k customer who is a member of a batch of size n and the mth served customer among members of the same batch. Let $N_\nu^{(\mathrm{D}_k)}(n;m)$ and $S^{(\mathrm{D}_k)}(n;m)$ $(k,\nu\in\mathcal{K},\ n=1,2,\ldots,\ m=1,2,\ldots,n)$ denote generic random variables representing the number of class ν customers and the state of the underlying Markov chain, respectively, immediately after the departure of customer $\mathrm{C}_k(n;m)$ in steady state. We then define $\boldsymbol{q}_k^*(\boldsymbol{z}\,|\,n;m)$ $(k\in\mathcal{K},\ n=1,2,\ldots,\ m=1,2,\ldots,n)$ as a $1\times M$ vector whose jth element represents

$$\mathrm{E}\left[\prod_{\nu\in\mathcal{K}}z_\nu^{N_\nu^{(\mathrm{D}_k)}(n;m)}1\{S^{(\mathrm{D}_k)}(n;m)=j\}\right],$$

where $1\{\chi\}$ denotes an indicator function of event χ. Because a randomly chosen customer of class k is a member of a batch of size n with probability $n\boldsymbol{\pi}\boldsymbol{D}_k(n)\boldsymbol{e}/\lambda_k$, we have

$$\boldsymbol{q}_k^*(\boldsymbol{z})=\sum_{n=1}^{\infty}\frac{n\boldsymbol{\pi}\boldsymbol{D}_k(n)\boldsymbol{e}}{\lambda_k}\frac{1}{n}\sum_{m=1}^{n}\boldsymbol{q}_k^*(\boldsymbol{z}\,|\,n;m). \tag{2.10}$$

In what follows, we consider $\boldsymbol{q}_k^*(\boldsymbol{z} \,|\, n; m)$.

We define $W_k(n;m)$ ($k \in \mathcal{K}$, $n = 1, 2, \ldots$, $m = 1, 2, \ldots, n$) as a generic random variable representing the sojourn time of customer $\mathrm{C}_k(n;m)$. Note here that

$$W_k(n;m) = W_{q,k}(n;1) + H_{k,1} + \cdots + H_{k,m},$$

where $W_{q,k}(n;1)$ denotes the actual waiting time of customer $\mathrm{C}_k(n;1)$, and $H_{k,l}$ ($l = 1, 2, \ldots, m$) denotes the service time of customer $\mathrm{C}_k(n;l)$. By definition, $W_{q,k}(n;1)$ depends only on the past history up to the arrival instant of a batch including customer $\mathrm{C}_k(n;m)$. On the other hand, the number of customers in the system immediately after the departure of customer $\mathrm{C}_k(n;m)$ is equal to the sum of the $n-m$ customers in the same batch and customers who arrived during the sojourn time of customer $\mathrm{C}_k(n;m)$. Note here that the latter is conditionally independent of the past history given the length of the sojourn time and the state of the underlying Markov chain immediately after the arrival of the batch. Thus we have

$$\boldsymbol{q}_k^*(\boldsymbol{z} \,|\, n; m) = z_k^{n-m} \int_0^\infty d\boldsymbol{w}_{q,k}(x \,|\, n; 1) \boldsymbol{N}^*(x, \boldsymbol{z}) \left[\int_0^\infty dH_k(y) \boldsymbol{N}^*(y, \boldsymbol{z}) \right]^m, \tag{2.11}$$

where

$$\boldsymbol{N}^*(x, \boldsymbol{z}) = \exp\left[\left(\boldsymbol{C} + \sum_{k \in \mathcal{K}} \boldsymbol{D}_k^*(z_k) \right) x \right]. \tag{2.12}$$

Theorem 2.3 *The vector generating function $\boldsymbol{q}_k^*(\boldsymbol{z})$ ($k \in \mathcal{K}$) of the joint queue length distribution immediately after departures of class k customers in the steady state is given by*

$$\boldsymbol{q}_k^*(\boldsymbol{z}) = \frac{1}{\lambda_k} \sum_{m=1}^\infty \sum_{l=0}^\infty z_k^l \int_0^\infty d\boldsymbol{v}(x) \boldsymbol{D}_k(m+l) \boldsymbol{N}^*(x, \boldsymbol{z}) \left[\int_0^\infty dH_k(y) \boldsymbol{N}^*(y, \boldsymbol{z}) \right]^m. \tag{2.13}$$

Proof. Using (2.1), (2.10) and (2.11), we have

$$\begin{aligned} \boldsymbol{q}_k^*(\boldsymbol{z}) &= \sum_{n=1}^\infty \frac{\boldsymbol{\pi} \boldsymbol{D}_k(n) \boldsymbol{e}}{\lambda_k} \sum_{m=1}^n z_k^{n-m} \int_0^\infty \frac{d\boldsymbol{v}(x) \boldsymbol{D}_k(n)}{\boldsymbol{\pi} \boldsymbol{D}_k(n) \boldsymbol{e}} \boldsymbol{N}^*(x, \boldsymbol{z}) \left[\int_0^\infty dH_k(y) \boldsymbol{N}^*(y, \boldsymbol{z}) \right]^m \\ &= \frac{1}{\lambda_k} \sum_{n=1}^\infty \sum_{m=1}^n z_k^{n-m} \int_0^\infty d\boldsymbol{v}(x) \boldsymbol{D}_k(n) \boldsymbol{N}^*(x, \boldsymbol{z}) \left[\int_0^\infty dH_k(y) \boldsymbol{N}^*(y, \boldsymbol{z}) \right]^m, \end{aligned}$$

from which (2.13) follows. ■

2.4.3 Recursions for discrete phase-type batch sizes

In this subsection, we develop a recursive formula to compute the vector mass function $\boldsymbol{q}_k(\boldsymbol{n})$ of the joint queue length immediately after departures of each class under the following assumption.

Assumption 2.2 *Let $\boldsymbol{\alpha}_k$ denote a $1 \times M_k$ probability vector and $\boldsymbol{P}_k$ denote an $M_k \times M_k$ substochastic matrix. Then, batch sizes of class k are i.i.d. according to a discrete phase-type distribution with representation $(\boldsymbol{\alpha}_k, \boldsymbol{P}_k)$, i.e.,*

$$\boldsymbol{D}_k(n) = g_k(n) \boldsymbol{D}_k, \tag{2.14}$$

$$g_k(n) = \boldsymbol{\alpha}_k \boldsymbol{P}_k^{n-1} (\boldsymbol{I} - \boldsymbol{P}_k) \boldsymbol{e}, \qquad n = 1, 2, \ldots, \tag{2.15}$$

where $\boldsymbol{I}(m)$ denotes an $m \times m$ identity matrix (the size m is suppressed when it is clear from the context).

We now define $\boldsymbol{I}(m)$ as an $m \times m$ identity matrix. When the size of an identity matrix is clear from the context, we suppress (m).

Lemma 2.1 *Under Assumption 2.2, $\boldsymbol{q}_k^*(\boldsymbol{z})$ $(k \in \mathcal{K})$ is given by*

$$\begin{aligned}\boldsymbol{q}_k^*(\boldsymbol{z}) &= \frac{1}{\lambda_k}\int_0^\infty d\boldsymbol{v}(x)\boldsymbol{D}_k\boldsymbol{N}^*(x,\boldsymbol{z})\left(\boldsymbol{\alpha}_k \otimes \int_0^\infty dH_k(y)\boldsymbol{N}^*(y,\boldsymbol{z})\right)\\ &\quad\cdot\left[\boldsymbol{I}-\boldsymbol{P}_k\otimes\int_0^\infty dH_k(y)\boldsymbol{N}^*(y,\boldsymbol{z})\right]^{-1}\left[\left\{(\boldsymbol{I}-z_k\boldsymbol{P}_k)^{-1}(\boldsymbol{I}-\boldsymbol{P}_k)\boldsymbol{e}\right\}\otimes\boldsymbol{I}(M)\right]. \quad (2.16)\end{aligned}$$

Proof. Substituting (2.14) and (2.15) into (2.13) and using properties of Kronecker product:

$a\boldsymbol{X} = a\otimes\boldsymbol{X}$ for any scalar a, and,

$(\boldsymbol{X}_1\cdots\boldsymbol{X}_n)\otimes(\boldsymbol{Y}_1\cdots\boldsymbol{Y}_n)$

$\quad = (\boldsymbol{X}_1\otimes\boldsymbol{Y}_1)\cdots(\boldsymbol{X}_n\otimes\boldsymbol{Y}_n)$ for any $n = 1,2,\ldots,$

we obtain

$$\begin{aligned}\boldsymbol{q}_k^*(\boldsymbol{z}) &= \frac{1}{\lambda_k}\int_0^\infty d\boldsymbol{v}(x)\boldsymbol{D}_k\boldsymbol{N}^*(x,\boldsymbol{z})\cdot\left\{\boldsymbol{\alpha}_k\sum_{m=1}^\infty\boldsymbol{P}_k^{m-1}\sum_{l=0}^\infty z_k^l\boldsymbol{P}_k^l(\boldsymbol{I}-\boldsymbol{P}_k)\boldsymbol{e}\right\}\\ &\qquad\otimes\left\{\left(\int_0^\infty dH_k(y)\boldsymbol{N}^*(y,\boldsymbol{z})\right)^m\right\}\\ &= \frac{1}{\lambda_k}\int_0^\infty d\boldsymbol{v}(x)\boldsymbol{D}_k\boldsymbol{N}^*(x,\boldsymbol{z})\cdot\left\{\boldsymbol{\alpha}_k\sum_{m=1}^\infty\boldsymbol{P}_k^{m-1}(\boldsymbol{I}-z_k\boldsymbol{P}_k)^{-1}(\boldsymbol{I}-\boldsymbol{P}_k)\boldsymbol{e}\right\}\\ &\qquad\otimes\left\{\left(\int_0^\infty dH_k(y)\boldsymbol{N}^*(y,\boldsymbol{z})\right)^m\right\}\\ &= \frac{1}{\lambda_k}\int_0^\infty d\boldsymbol{v}(x)\boldsymbol{D}_k\boldsymbol{N}^*(x,\boldsymbol{z})\cdot\left(\boldsymbol{\alpha}_k\otimes\int_0^\infty dH_k(y)\boldsymbol{N}^*(y,\boldsymbol{z})\right)\\ &\quad\cdot\sum_{m=1}^\infty\left(\boldsymbol{P}_k\otimes\int_0^\infty dH_k(y)\boldsymbol{N}^*(y,\boldsymbol{z})\right)^{m-1}\left[\left\{(\boldsymbol{I}-z_k\boldsymbol{P}_k)^{-1}(\boldsymbol{I}-\boldsymbol{P}_k)\boldsymbol{e}\right\}\otimes\boldsymbol{I}(M)\right]\\ &= \frac{1}{\lambda_k}\int_0^\infty d\boldsymbol{v}(x)\boldsymbol{D}_k\boldsymbol{N}^*(x,\boldsymbol{z})\cdot\left(\boldsymbol{\alpha}_k\otimes\int_0^\infty dH_k(y)\boldsymbol{N}^*(y,\boldsymbol{z})\right)\\ &\quad\cdot\left[\boldsymbol{I}-\boldsymbol{P}_k\otimes\int_0^\infty dH_k(y)\boldsymbol{N}^*(y,\boldsymbol{z})\right]^{-1}\left[\left\{(\boldsymbol{I}-z_k\boldsymbol{P}_k)^{-1}(\boldsymbol{I}-\boldsymbol{P}_k)\boldsymbol{e}\right\}\otimes\boldsymbol{I}(M)\right],\end{aligned}$$

which completes the proof. ∎

We define $\boldsymbol{v}_k(\boldsymbol{n})$ $(k\in\mathcal{K},\ \boldsymbol{n}\in\mathcal{Z})$ as a $1\times M$ vector satisfying

$$\sum_{\boldsymbol{n}\in\mathcal{Z}} z_1^{n_1}\cdots z_K^{n_K}\boldsymbol{v}_k(\boldsymbol{n}) = \int_0^\infty d\boldsymbol{v}(x)\boldsymbol{D}_k\boldsymbol{N}^*(x,\boldsymbol{z}). \quad (2.17)$$

We also define $\boldsymbol{A}_k(\boldsymbol{n})$ and $\boldsymbol{\Gamma}_k(\boldsymbol{n})$ $(k\in\mathcal{K},\ \boldsymbol{n}\in\mathcal{Z})$ as $M\times M$ and $MM_k\times MM_k$ matrices satisfying

$$\begin{aligned}\sum_{\boldsymbol{n}\in\mathcal{Z}} z_1^{n_1}\cdots z_K^{n_K}\boldsymbol{A}_k(\boldsymbol{n}) &= \int_0^\infty dH_k(y)\boldsymbol{N}^*(y,\boldsymbol{z}), \quad (2.18)\\ \sum_{\boldsymbol{n}\in\mathcal{Z}} z_1^{n_1}\cdots z_K^{n_K}\boldsymbol{\Gamma}_k(\boldsymbol{n}) &= \left[\boldsymbol{I}-\boldsymbol{P}_k\otimes\int_0^\infty dH_k(y)\boldsymbol{N}^*(y,\boldsymbol{z})\right]^{-1}, \quad (2.19)\end{aligned}$$

respectively. Note here that

$$\left\{(\boldsymbol{I}-z_k\boldsymbol{P}_k)^{-1}(\boldsymbol{I}-\boldsymbol{P}_k)\boldsymbol{e}\right\}\otimes\boldsymbol{I}(M)=\sum_{m=0}^{\infty}z_k^m\left\{\boldsymbol{P}_k^m(\boldsymbol{I}-\boldsymbol{P}_k)\boldsymbol{e}\right\}\otimes\boldsymbol{I}(M).$$

Thus (2.16) is rewritten to be

$$\begin{aligned}\boldsymbol{q}_k^*(\boldsymbol{z}) &= \frac{1}{\lambda_k}\sum_{\boldsymbol{n}\in\mathcal{Z}}z_1^{n_1}\cdots z_K^{n_K}\sum_{m=0}^{n_k}\sum_{\substack{\boldsymbol{n}_1+\boldsymbol{n}_2+\boldsymbol{n}_3\\=\boldsymbol{n}-m\boldsymbol{e}_k}}\boldsymbol{v}_k(\boldsymbol{n}_1)\\ &\quad\cdot[\boldsymbol{\alpha}_k\otimes\boldsymbol{A}_k(\boldsymbol{n}_2)]\boldsymbol{\Gamma}_k(\boldsymbol{n}_3)\left[\{\boldsymbol{P}_k^m(\boldsymbol{I}-\boldsymbol{P}_k)\boldsymbol{e}\}\otimes\boldsymbol{I}(M)\right],\end{aligned}\tag{2.20}$$

where $\boldsymbol{n}_j\in\mathcal{Z}$ for $j=1,2,3$. Comparing coefficient vectors of $z_1^{n_1}\cdots z_K^{n_K}$ on both sizes of (2.20), we obtain the following result.

Theorem 2.4 *Under Assumption 2.2, $\boldsymbol{q}_k(\boldsymbol{n})$ ($k\in\mathcal{K}$, $\boldsymbol{n}\in\mathcal{Z}$) is given by*

$$\boldsymbol{q}_k(\boldsymbol{n}) = \frac{1}{\lambda_k}\sum_{m=0}^{n_k}\sum_{\substack{\boldsymbol{n}_1+\boldsymbol{n}_2+\boldsymbol{n}_3\\=\boldsymbol{n}-m\boldsymbol{e}_k}}\boldsymbol{v}_k(\boldsymbol{n}_1)[\boldsymbol{\alpha}_k\otimes\boldsymbol{A}_k(\boldsymbol{n}_2)]\boldsymbol{\Gamma}_k(\boldsymbol{n}_3)\left[\boldsymbol{P}_k^m(\boldsymbol{I}-\boldsymbol{P}_k)\boldsymbol{e}\otimes\boldsymbol{I}(M)\right],$$

where $\boldsymbol{n}_j\in\mathcal{Z}$ for $j=1,2,3$.

Theorem 2.4 implies that the computation of $\boldsymbol{q}_k(\boldsymbol{n})$ is reduced to those of $\boldsymbol{v}_k(\boldsymbol{n})$, $\boldsymbol{A}_k(\boldsymbol{n})$ and $\boldsymbol{\Gamma}_k(\boldsymbol{n})$, which are discussed in the rest of this subsection.

We first consider the $\boldsymbol{A}_k(\boldsymbol{n})$. Let θ denote the maximum absolute value of diagonal elements of $\boldsymbol{C}$. We define $\boldsymbol{F}_m(\boldsymbol{n})$ ($m=0,1,\ldots,\ \boldsymbol{n}\in\mathcal{Z}$) as an $M\times M$ matrix that satisfies

$$\sum_{\boldsymbol{n}\in\mathcal{Z}}z_1^{n_1}\cdots z_K^{n_K}\boldsymbol{F}_m(\boldsymbol{n})=\left[\boldsymbol{I}+\theta^{-1}\left(\boldsymbol{C}+\sum_{k\in\mathcal{K}}\boldsymbol{D}_k^*(z_k)\right)\right]^m.\tag{2.21}$$

Lemma 2.2 *$\boldsymbol{A}_k(\boldsymbol{n})$ is given by*

$$\boldsymbol{A}_k(\boldsymbol{n})=\sum_{m=0}^{\infty}\gamma_k^{(m)}(\theta)\boldsymbol{F}_m(\boldsymbol{n}),\qquad k\in\mathcal{K},\ \boldsymbol{n}\in\mathcal{Z},\tag{2.22}$$

where

$$\gamma_k^{(m)}(\theta)=\int_0^{\infty}e^{-\theta y}\frac{(\theta y)^m}{m!}dH_k(y),\ \ k\in\mathcal{K},\ m=0,1,\ldots,\tag{2.23}$$

and $\boldsymbol{F}_m(\boldsymbol{n})$'s are recursively determined by

$$\boldsymbol{F}_0(\boldsymbol{n})=\begin{cases}\boldsymbol{I}, & \text{if } \boldsymbol{n}=\boldsymbol{0},\\ \boldsymbol{O}, & \text{otherwise,}\end{cases}\tag{2.24}$$

and for $m=0,1,\ldots$,

$$\boldsymbol{F}_{m+1}(\boldsymbol{n}) = \boldsymbol{F}_m(\boldsymbol{n})(\boldsymbol{I}+\theta^{-1}\boldsymbol{C})+\theta^{-1}\sum_{k\in\mathcal{K}}\sum_{l_k=1}^{n_k}\boldsymbol{F}_m(\boldsymbol{n}-l_k\boldsymbol{e}_k)\boldsymbol{D}_k(l_k),\quad \boldsymbol{n}\in\mathcal{Z}.\tag{2.25}$$

Proof. From (2.12), (2.18) and (2.23), we obtain

$$\begin{aligned}
&\sum_{\boldsymbol{n}\in\mathcal{Z}} z_1^{n_1}\cdots z_K^{n_K}\boldsymbol{A}_k(\boldsymbol{n}) \\
&\quad= \int_0^\infty dH_k(y)\exp\left[\left(\boldsymbol{C}+\sum_{k\in\mathcal{K}}\boldsymbol{D}_k^*(z_k)\right)y\right] \\
&\quad= \sum_{m=0}^\infty\int_0^\infty e^{-\theta y}\frac{(\theta y)^m}{m!}dH_k(y)\left[\boldsymbol{I}+\theta^{-1}\left(\boldsymbol{C}+\sum_{k\in\mathcal{K}}\boldsymbol{D}_k^*(z_k)\right)\right]^m \\
&\quad= \sum_{m=0}^\infty \gamma_k^{(m)}(\theta)\left[\boldsymbol{I}+\theta^{-1}\left(\boldsymbol{C}+\sum_{k\in\mathcal{K}}\boldsymbol{D}_k^*(z_k)\right)\right]^m. \qquad (2.26)
\end{aligned}$$

Substituting (2.21) into (2.26) and changing the order of summations, we have

$$\sum_{\boldsymbol{n}\in\mathcal{Z}} z_1^{n_1}\cdots z_K^{n_K}\boldsymbol{A}_k(\boldsymbol{n}) = \sum_{\boldsymbol{n}\in\mathcal{Z}} z_1^{n_1}\cdots z_K^{n_K}\sum_{m=0}^\infty \gamma_k^{(m)}(\theta)\boldsymbol{F}_m(\boldsymbol{n}). \qquad (2.27)$$

Comparing the coefficient matrices of $z_1^{n_1}\cdots z_K^{n_K}$ on both sizes of (2.27), we obtain (2.22).

The remaining is to show (2.24) and (2.25). From (2.6) and (2.21), we have for $m=0,1,\ldots,$

$$\begin{aligned}
&\sum_{\boldsymbol{n}\in\mathcal{Z}} z_1^{n_1}\cdots z_K^{n_K}\boldsymbol{F}_{m+1}(\boldsymbol{n}) \\
&\quad= \sum_{\boldsymbol{n}\in\mathcal{Z}} z_1^{n_1}\cdots z_K^{n_K}\boldsymbol{F}_m(\boldsymbol{n})\left[\boldsymbol{I}+\theta^{-1}\left(\boldsymbol{C}+\sum_{k\in\mathcal{K}}\boldsymbol{D}_k^*(z_k)\right)\right] \\
&\quad= \sum_{\boldsymbol{n}\in\mathcal{Z}} z_1^{n_1}\cdots z_K^{n_K}\boldsymbol{F}_m(\boldsymbol{n})\left(\boldsymbol{I}+\theta^{-1}\boldsymbol{C}\right) \\
&\qquad+\sum_{\boldsymbol{n}\in\mathcal{Z}}\sum_{k\in\mathcal{K}}\left[\sum_{l_k=1}^\infty z_1^{n_1}\cdots z_{k-1}^{n_{k-1}}z_k^{n_k+l_k}z_{k+1}^{n_{k+1}}\cdots z_K^{n_K}\boldsymbol{F}_m(\boldsymbol{n})\cdot\theta^{-1}\boldsymbol{D}_k(l_k)\right] \\
&\quad= \sum_{\boldsymbol{n}\in\mathcal{Z}} z_1^{n_1}\cdots z_K^{n_K}\boldsymbol{F}_m(\boldsymbol{n})\left(\boldsymbol{I}+\theta^{-1}\boldsymbol{C}\right) \\
&\qquad+\sum_{\boldsymbol{n}\in\mathcal{Z}^+} z_1^{n_1}\cdots z_K^{n_K}\theta^{-1}\sum_{k\in\mathcal{K}}\sum_{l_k=1}^{n_k}\boldsymbol{F}_m(\boldsymbol{n}-l_k\boldsymbol{e}_k)\boldsymbol{D}_k(l_k).
\end{aligned}$$

Comparing the coefficient vectors of $z_1^{n_1}\cdots z_K^{n_K}$ on both sides of the above equation, we obtain (2.25). Finally, (2.24) is clear from the definition. ■

Next we consider the $\boldsymbol{\Gamma}_k(\boldsymbol{n})$ in (2.19).

Lemma 2.3 *$\boldsymbol{\Gamma}_k(\boldsymbol{n})$ ($k\in\mathcal{K}$, $\boldsymbol{n}\in\mathcal{Z}$) is determined by the following recursion:*

$$\begin{aligned}
\boldsymbol{\Gamma}_k(\boldsymbol{0}) &= \left[\boldsymbol{I}-\boldsymbol{P}_k\otimes\boldsymbol{A}_k(\boldsymbol{0})\right]^{-1}, \\
\boldsymbol{\Gamma}_k(\boldsymbol{n}) &= \sum_{\substack{\boldsymbol{0}\le\boldsymbol{l}\le\boldsymbol{n}\\ \boldsymbol{l}\neq\boldsymbol{0}}}\boldsymbol{\Gamma}_k(\boldsymbol{n}-\boldsymbol{l})\left[\boldsymbol{P}_k\otimes\boldsymbol{A}_k(\boldsymbol{l})\right]\boldsymbol{\Gamma}_k(\boldsymbol{0}), \qquad \boldsymbol{n}\in\mathcal{Z}^+.
\end{aligned}$$

Proof. Note first that (2.19) is equivalent to

$$\sum_{\boldsymbol{n}\in\mathcal{Z}} z_1^{n_1}\cdots z_K^{n_K}\boldsymbol{\Gamma}_k(\boldsymbol{n})\left[\boldsymbol{I}-\boldsymbol{P}_k\otimes\int_0^\infty dH_k(y)\boldsymbol{N}^*(y,\boldsymbol{z})\right]=\boldsymbol{I}.$$

Substituting (2.18) into the above equation, we have

$$\sum_{\boldsymbol{n}\in\mathcal{Z}} z_1^{n_1}\cdots z_K^{n_K}\boldsymbol{\Gamma}_k(\boldsymbol{n})\left[\boldsymbol{I}-\boldsymbol{P}_k\otimes\sum_{\boldsymbol{l}\in\mathcal{Z}} z_1^{l_1}\cdots z_K^{l_K}\boldsymbol{A}_k(\boldsymbol{l})\right]=\boldsymbol{I},$$

from which it follows that

$$\sum_{\boldsymbol{n}\in\mathcal{Z}} z_1^{n_1}\cdots z_K^{n_K}\boldsymbol{\Gamma}_k(\boldsymbol{n})-\sum_{\boldsymbol{n}\in\mathcal{Z}} z_1^{n_1}\cdots z_K^{n_K}\sum_{\boldsymbol{0}\le\boldsymbol{l}\le\boldsymbol{n}}\boldsymbol{\Gamma}_k(\boldsymbol{n}-\boldsymbol{l})\left[\boldsymbol{P}_k\otimes\boldsymbol{A}_k(\boldsymbol{l})\right]=\boldsymbol{I}.$$

Comparing the coefficient matrices of $z_1^{n_1}\cdots z_K^{n_K}$ on both sizes of the above equation, we have

$$\begin{aligned}\boldsymbol{\Gamma}_k(\boldsymbol{0})-\boldsymbol{\Gamma}_k(\boldsymbol{0})\left[\boldsymbol{P}_k\otimes\boldsymbol{A}_k(\boldsymbol{0})\right] &= \boldsymbol{I},\\ \boldsymbol{\Gamma}_k(\boldsymbol{n})-\sum_{\boldsymbol{0}\le\boldsymbol{l}\le\boldsymbol{n}}\boldsymbol{\Gamma}_k(\boldsymbol{n}-\boldsymbol{l})\left[\boldsymbol{P}_k\otimes\boldsymbol{A}_k(\boldsymbol{l})\right] &= \boldsymbol{O},\qquad \boldsymbol{n}\in\mathcal{Z}^+,\end{aligned}$$

or equivalently,

$$\boldsymbol{\Gamma}_k(\boldsymbol{0})=\left[\boldsymbol{I}-\boldsymbol{P}_k\otimes\boldsymbol{A}_k(\boldsymbol{0})\right]^{-1},$$

and for $\boldsymbol{n}\in\mathcal{Z}^+$,

$$\boldsymbol{\Gamma}_k(\boldsymbol{n})=\sum_{\substack{\boldsymbol{0}\le\boldsymbol{l}\le\boldsymbol{n}\\ \boldsymbol{l}\ne\boldsymbol{0}}}\boldsymbol{\Gamma}_k(\boldsymbol{n}-\boldsymbol{l})\left[\boldsymbol{P}_k\otimes\boldsymbol{A}_k(\boldsymbol{l})\right]\left[\boldsymbol{I}-\boldsymbol{P}_k\otimes\boldsymbol{A}_k(\boldsymbol{0})\right]^{-1},$$

from which Lemma 2.3 follows. ■

Finally, we consider the $\boldsymbol{v}_k(\boldsymbol{n})$ in (2.17). In a very similar way to derive (2.22), we obtain the following lemma.

Lemma 2.4 *$\boldsymbol{v}_k(\boldsymbol{n})$ ($k\in\mathcal{K}$, $\boldsymbol{n}\in\mathcal{Z}$) is given by*

$$\boldsymbol{v}_k(\boldsymbol{n})=\sum_{m=0}^{\infty}\boldsymbol{v}^{(m)}(\theta)\boldsymbol{D}_k\boldsymbol{F}_m(\boldsymbol{n}), \tag{2.28}$$

where $\boldsymbol{F}_m(\boldsymbol{n})$ is given in (2.24) and (2.25), and

$$\boldsymbol{v}^{(m)}(\theta)=\int_0^\infty e^{-\theta x}\frac{(\theta x)^m}{m!}d\boldsymbol{v}(x),\qquad m=0,1,\ldots.$$

Thus $\boldsymbol{v}_k(\boldsymbol{n})$ is given in terms of the $\boldsymbol{v}^{(m)}(\theta)$. Because the computation of the $\boldsymbol{v}^{(m)}(\theta)$ has already been studied in [Taki01b], we summarize the result below. As for the details, readers are referred to Lemma 3 in [Taki01b].

Note first that

$$\sum_{m=0}^{\infty} z^m\boldsymbol{v}^{(m)}(\theta)=\boldsymbol{v}^*(\theta-\theta z), \tag{2.29}$$

where $\boldsymbol{v}^*(s)$ is given in (1.23). Thus, substituting $\theta - \theta z$ for s in (1.23) and using (2.29) yield

$$\sum_{m=0}^{\infty} z^m \boldsymbol{v}^{(m)}(\theta) \left[(\theta - \theta z)\boldsymbol{I} + \boldsymbol{C} + \sum_{m=0}^{\infty} z^m \boldsymbol{D}^{(m)}(\theta) \right] = (\theta - \theta z)(1-\rho)\boldsymbol{\kappa}, \tag{2.30}$$

where $\boldsymbol{D}^{(m)}(\theta)$ denotes

$$\boldsymbol{D}^{(m)}(\theta) = \int_0^{\infty} e^{-\theta x} \frac{(\theta x)^m}{m!} d\overline{\boldsymbol{D}}(x).$$

Comparing the coefficient vectors of z^m $(m = 0, 1, \ldots)$ on both sides of (2.30), we can show that the $\boldsymbol{v}^{(m)}(\theta)$ is identical to the stationary distribution of a Markov chain of M/G/1 type whose transition probability matrix is given by [Taki01b]

$$\begin{bmatrix} \widetilde{\boldsymbol{B}}_0 + \widetilde{\boldsymbol{B}}_1 & \widetilde{\boldsymbol{B}}_2 & \widetilde{\boldsymbol{B}}_3 & \widetilde{\boldsymbol{B}}_4 & \cdots \\ \widetilde{\boldsymbol{B}}_0 & \widetilde{\boldsymbol{B}}_1 & \widetilde{\boldsymbol{B}}_2 & \widetilde{\boldsymbol{B}}_3 & \cdots \\ \boldsymbol{O} & \widetilde{\boldsymbol{B}}_0 & \widetilde{\boldsymbol{B}}_1 & \widetilde{\boldsymbol{B}}_2 & \cdots \\ \boldsymbol{O} & \boldsymbol{O} & \widetilde{\boldsymbol{B}}_0 & \widetilde{\boldsymbol{B}}_1 & \cdots \\ \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} & \widetilde{\boldsymbol{B}}_0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix},$$

where

$$\widetilde{\boldsymbol{B}}_0 = \boldsymbol{I} + \theta^{-1}(\boldsymbol{C} + \boldsymbol{D}^{(0)}(\theta)), \quad \widetilde{\boldsymbol{B}}_m = \theta^{-1}\boldsymbol{D}^{(m)}(\theta), \quad m \geq 1.$$

Thus applying the general theory of Markov chains of M/G/1 type [Neut89], we can compute the $\boldsymbol{v}^{(m)}(\theta)$. As for the truncation and stopping criteria in computing the steady-state solution of Markov chains of M/G/1 type, readers are referred to [Neut89, Taki94b].

Let $\boldsymbol{d}_k^{(m)}(\theta)$ $(k \in \mathcal{K},\ m = 0, 1, \ldots)$ denote a $1 \times M_k$ vector which satisfies

$$\sum_{m=0}^{\infty} z^m \boldsymbol{d}_k^{(m)}(\theta) = H_k^*(\theta - \theta z)\boldsymbol{\alpha}_k(\boldsymbol{I} - \boldsymbol{P}_k)[\boldsymbol{I} - H_k^*(\theta - \theta z)\boldsymbol{P}_k]^{-1}. \tag{2.31}$$

Lemma 2.5 *Under Assumption 2.2, $\boldsymbol{D}^{(m)}(\theta)$ is given by*

$$\boldsymbol{D}^{(m)}(\theta) = \sum_{k \in \mathcal{K}} \boldsymbol{d}_k^{(m)}(\theta) \boldsymbol{e} \boldsymbol{D}_k, \qquad m = 0, 1, \ldots,$$

where $\boldsymbol{d}_k^{(m)}(\theta)$'s $(k \in \mathcal{K})$ are recursively determined by

$$\boldsymbol{d}_k^{(0)}(\theta) = \gamma_k^{(0)}(\theta)\boldsymbol{\alpha}_k \left(\boldsymbol{I} - \boldsymbol{P}_k\right) \left[\boldsymbol{I} - \gamma_k^{(0)}(\theta)\boldsymbol{P}_k\right]^{-1},$$

and for $m = 1, 2, \ldots$,

$$\boldsymbol{d}_k^{(m)}(\theta) = \frac{\gamma_k^{(m)}(\theta)}{\gamma_k^{(0)}(\theta)} \boldsymbol{d}_k^{(0)}(\theta) + \left[\sum_{l=1}^{m} \gamma_k^{(l)}(\theta) \boldsymbol{d}_k^{(m-l)}(\theta) \right] \boldsymbol{P}_k \left[\boldsymbol{I} - \gamma_k^{(0)}(\theta)\boldsymbol{P}_k\right]^{-1}.$$

Proof. Note first that

$$\sum_{m=0}^{\infty} z^m \boldsymbol{D}^{(m)}(\theta) = \overline{\boldsymbol{D}}^*(\theta - \theta z),$$

where $\overline{\boldsymbol{D}}^*(s)$ is given in (2.5). Thus, substituting $\theta - \theta z$ for s in (2.5) and using (2.14) and (2.15), we have

$$\begin{aligned}\sum_{m=0}^{\infty} z^m \boldsymbol{D}^{(m)}(\theta) &= \sum_{k\in\mathcal{K}} \sum_{n=1}^{\infty} \boldsymbol{\alpha}_k \boldsymbol{P}_k^{n-1}(\boldsymbol{I} - \boldsymbol{P}_k)\boldsymbol{e}\{H_k^*(\theta - \theta z)\}^n \boldsymbol{D}_k \\ &= \sum_{k\in\mathcal{K}} H_k^*(\theta - \theta z)\boldsymbol{\alpha}_k(\boldsymbol{I} - \boldsymbol{P}_k)[\boldsymbol{I} - H_k^*(\theta - \theta z)\boldsymbol{P}_k]^{-1}\boldsymbol{e}\boldsymbol{D}_k. \end{aligned} \tag{2.32}$$

It then follows from (2.31) and (2.32) that

$$\sum_{m=0}^{\infty} z^m \boldsymbol{D}^{(m)}(\theta) = \sum_{k\in\mathcal{K}} \sum_{m=0}^{\infty} z^m \boldsymbol{d}_k^{(m)}(\theta)\boldsymbol{e}\boldsymbol{D}_k = \sum_{m=0}^{\infty} z^m \sum_{k\in\mathcal{K}} \boldsymbol{d}_k^{(m)}(\theta)\boldsymbol{e}\boldsymbol{D}_k. \tag{2.33}$$

Note here that

$$H_k^*(\theta - \theta z) = \sum_{m=0}^{\infty} z^m \gamma_k^{(m)}(\theta), \qquad k \in \mathcal{K}. \tag{2.34}$$

Thus from (2.31) and (2.34), we have

$$\sum_{m=0}^{\infty} z^m \boldsymbol{d}_k^{(m)}(\theta) \left[\boldsymbol{I} - \sum_{l=0}^{\infty} z^l \gamma_k^{(l)}(\theta)\boldsymbol{P}_k\right] = \sum_{m=0}^{\infty} z^m \gamma_k^{(m)}(\theta)\boldsymbol{\alpha}_k \left[\boldsymbol{I} - \boldsymbol{P}_k\right],$$

or equivalently,

$$\sum_{m=0}^{\infty} z^m \boldsymbol{d}_k^{(m)}(\theta) - \sum_{m=0}^{\infty} z^m \sum_{l=0}^{m} \boldsymbol{d}_k^{(m-l)}(\theta)\gamma_k^{(l)}(\theta)\boldsymbol{P}_k = \sum_{m=0}^{\infty} z^m \gamma_k^{(m)}(\theta)\boldsymbol{\alpha}_k \left[\boldsymbol{I} - \boldsymbol{P}_k\right].$$

Comparing the coefficient vectors of z^m $(m = 0, 1, \ldots)$ on both sides of the above equation, we have

$$\boldsymbol{d}_k^{(0)}(\theta)\left[\boldsymbol{I} - \gamma_k^{(0)}(\theta)\boldsymbol{P}_k\right] = \gamma_k^{(0)}(\theta)\boldsymbol{\alpha}_k\left[\boldsymbol{I} - \boldsymbol{P}_k\right], \tag{2.35}$$

and for $m = 1, 2, \ldots$,

$$\boldsymbol{d}_k^{(m)}(\theta)\left[\boldsymbol{I} - \gamma_k^{(0)}(\theta)\boldsymbol{P}_k\right] - \sum_{l=1}^{m} \boldsymbol{d}_k^{(m-l)}(\theta)\gamma_k^{(l)}(\theta)\boldsymbol{P}_k = \gamma_k^{(m)}(\theta)\boldsymbol{\alpha}_k\left[\boldsymbol{I} - \boldsymbol{P}_k\right]. \tag{2.36}$$

Lemma 2.5 now follows from (2.33), (2.35) and (2.36). ■

2.5 Implementations of Recursions

In this section, we consider the implementation of recursions for $\boldsymbol{A}_k(\boldsymbol{n})$, $\boldsymbol{v}_k(\boldsymbol{n})$ and $\boldsymbol{\Gamma}_k(\boldsymbol{n})$, derived in the preceding section. At a glance, they would seem to be easy to implement. Contrary to the single arrival case [Taki01a, Taki01b], however, the computation of the $\boldsymbol{F}_m(\boldsymbol{n})$ appeared in $\boldsymbol{A}_k(\boldsymbol{n})$ and $\boldsymbol{v}_k(\boldsymbol{n})$ is not straightforward, because the direct implementation of the recursion requires very huge memory space and time-consuming. In what follows, we construct a numerically feasible procedure

to compute the approximate sequences of $\boldsymbol{A}_k(\boldsymbol{n})$ and $\boldsymbol{v}_k(\boldsymbol{n})$, avoiding the computation of $\boldsymbol{F}_m(\boldsymbol{n})$'s whose contributions to $\boldsymbol{A}_k(\boldsymbol{n})$ and $\boldsymbol{v}_k(\boldsymbol{n})$ are negligible, and establish the truncation/stopping criteria and error bounds. Further, we propose a computational procedure for the $\boldsymbol{\Gamma}_k(\boldsymbol{n})$ and establish the error bound.

We start with $\boldsymbol{A}_k(\boldsymbol{n})$ and $\boldsymbol{v}_k(\boldsymbol{n})$. Note first that for $k \in \mathcal{K}$,

$$\sum_{\boldsymbol{n}\in\mathcal{Z}} \boldsymbol{A}_k(\boldsymbol{n})\boldsymbol{e} = \boldsymbol{e}, \qquad \sum_{\boldsymbol{n}\in\mathcal{Z}} \boldsymbol{v}_k(\boldsymbol{n})\boldsymbol{e} = \lambda_k^{(\mathrm{B})},$$

where

$$\lambda_k^{(\mathrm{B})} = \boldsymbol{\pi}\boldsymbol{D}_k\boldsymbol{e}.$$

In numerical computation, we have to stop the computation of those sequences. Thus we develop a numerical procedure to obtain approximations $\breve{\boldsymbol{A}}_k(\boldsymbol{n})$ and $\breve{\boldsymbol{v}}_k(\boldsymbol{n})$ to $\boldsymbol{A}_k(\boldsymbol{n})$ and $\boldsymbol{v}_k(\boldsymbol{n})$, respectively, while ensuring the following error bounds: For a given ε $(0 < \varepsilon < 1)$, there exist $n_A(k)$ and $n_v(k)$ such that

$$\sum_{\substack{\boldsymbol{n}\in\mathcal{Z}\\ |\boldsymbol{n}|\le n_A(k)}} \breve{\boldsymbol{A}}_k(\boldsymbol{n})\boldsymbol{e} \;>\; (1-\varepsilon)\boldsymbol{e}, \tag{2.37}$$

$$\sum_{\substack{\boldsymbol{n}\in\mathcal{Z}\\ |\boldsymbol{n}|\le n_v(k)}} \breve{\boldsymbol{v}}_k(\boldsymbol{n})\boldsymbol{e} \;>\; (1-\varepsilon)\lambda_k^{(\mathrm{B})}, \tag{2.38}$$

where $|\boldsymbol{n}| = \sum_{k\in\mathcal{K}} n_k$ for $\boldsymbol{n} \in \mathcal{Z}$. In what follows, we first show our proposed algorithm and then show that the above error bounds are satisfied.

Numerical algorithm for $\boldsymbol{A}_k(\boldsymbol{n})$ and $\boldsymbol{v}_k(\boldsymbol{n})$

Input.

Stopping criterion : ε $(0 < \varepsilon < 1)$,
Underlying Markov chain : $\boldsymbol{C}$, $\boldsymbol{D}_k$ $(k \in \mathcal{K})$,
Batch size distribution : $\boldsymbol{\alpha}_k$, $\boldsymbol{P}_k$ $(k \in \mathcal{K})$,
Service time distribution : $H_k(x)$ $(k \in \mathcal{K})$.

Step 1. Choose ε_F $(0 < \varepsilon_F < 1)$ such that

$$\frac{\varepsilon_F}{\varepsilon} < \min_{k\in\mathcal{K}} \min\left(\frac{1}{\theta h_k}, \frac{\lambda_k^{(\mathrm{B})}}{\theta\overline{\boldsymbol{v}}^{(1)}\boldsymbol{D}_k\boldsymbol{e}}\right), \tag{2.39}$$

where $\overline{\boldsymbol{v}}^{(1)} = -\lim_{s\to 0+} d\boldsymbol{v}^*(s)/ds$, whose computational procedure can be found in [Taki94a] (see Proposition 1.9). Then compute the $\gamma_k^{(m)}(\theta)$ and the $\boldsymbol{v}^{(m)}(\theta)$ until they satisfy

$$\sum_{m=0}^{m_\gamma(k)} \gamma_k^{(m)}(\theta)(1-\varepsilon_F)^m \;>\; 1-\varepsilon, \quad k \in \mathcal{K}, \tag{2.40}$$

$$\sum_{m=0}^{m_v(k)} \boldsymbol{v}^{(m)}(\theta)\boldsymbol{D}_k\boldsymbol{e}(1-\varepsilon_F)^m \;>\; (1-\varepsilon)\lambda_k^{(\mathrm{B})}, \quad k \in \mathcal{K}, \tag{2.41}$$

for some $m_\gamma(k)$ and $m_v(k)$, respectively. Define $m_{\max}$ as

$$m_{\max} = \max_{k\in\mathcal{K}} \max(m_\gamma(k), m_v(k)).$$

Step 2. Choose ε_g such that $0 < \varepsilon_g < \varepsilon_F$. Then compute $g_k(n)$ $(n = 1, 2, \ldots)$ by (2.15) until the $g_k(n)$ satisfies

$$\theta^{-1} \sum_{n=1}^{n_g(k)} g_k(n) \boldsymbol{D}_k \boldsymbol{e} > \theta^{-1} \boldsymbol{D}_k \boldsymbol{e} - \frac{\varepsilon_g}{K} \boldsymbol{e}, \quad k \in \mathcal{K}, \tag{2.42}$$

for some $n_g(k)$.

Step 3. Compute $\breve{\boldsymbol{A}}_k(\boldsymbol{n})$ and $\breve{\boldsymbol{v}}_k(\boldsymbol{n})$ by the following procedure, where the initial values of $\breve{\boldsymbol{A}}_k(\boldsymbol{n})$ and $\breve{\boldsymbol{v}}_k(\boldsymbol{n})$ $(\boldsymbol{n} \in \mathcal{Z})$ are assumed to be $\boldsymbol{O}$ and $\boldsymbol{0}$, respectively.

Step (3–a). Set $\breve{\boldsymbol{F}}_0(\boldsymbol{0}) = \boldsymbol{I}$ and $n_F^{(0)} = 0$. Also set $\breve{\boldsymbol{A}}_k(\boldsymbol{0}) = \gamma_k^{(0)}(\theta)\boldsymbol{I}$ and $\breve{\boldsymbol{v}}_k(\boldsymbol{0}) = \boldsymbol{v}^{(0)}(\theta)\boldsymbol{D}_k$ for all $k \in \mathcal{K}$.

Step (3–b). Set $n_F^{(1)} = \max_{k\in\mathcal{K}} n_g(k)$ and $m = 1$, and compute $\breve{\boldsymbol{F}}_1(\boldsymbol{n})$'s $(|\boldsymbol{n}| \le n_F^{(1)})$ by

$$\breve{\boldsymbol{F}}_1(\boldsymbol{n}) = \begin{cases} \boldsymbol{I} + \theta^{-1}\boldsymbol{C}, & \text{if } \boldsymbol{n} = \boldsymbol{0}, \\ \theta^{-1} g_k(n_k)\boldsymbol{D}_k, & \text{if } \boldsymbol{n} \in \mathcal{Z}_k(F_1),\ k \in \mathcal{K}, \\ \boldsymbol{O}, & \text{otherwise}, \end{cases} \tag{2.43}$$

where

$$\mathcal{Z}_k(F_1) = \{\boldsymbol{n}; \boldsymbol{n} = n_k \boldsymbol{e}_k, n_k = 1, 2, \ldots, n_g(k)\}, \qquad k \in \mathcal{K}.$$

Step (3–c). For each $k \in \mathcal{K}$, if $m \le m_\gamma(k)$, add $\gamma_k^{(m)}(\theta)\breve{\boldsymbol{F}}_m(\boldsymbol{n})$ to $\breve{\boldsymbol{A}}_k(\boldsymbol{n})$ for all $\boldsymbol{n}$ $(|\boldsymbol{n}| \le n_F^{(m)})$. Also, for each $k \in \mathcal{K}$, if $m \le m_v(k)$, add $\boldsymbol{v}^{(m)}(\theta)\boldsymbol{D}_k\breve{\boldsymbol{F}}_m(\boldsymbol{n})$ to $\breve{\boldsymbol{v}}_k(\boldsymbol{n})$ for all $\boldsymbol{n}$ $(|\boldsymbol{n}| \le n_F^{(m)})$.

Step (3–d). If $m \ge m_{\max}$, stop computing, and otherwise, add one to m and go to Step (3–e).

Step (3–e). For each $n = 0, 1, \ldots$, compute $\breve{\boldsymbol{F}}_m(\boldsymbol{n})$'s $(|\boldsymbol{n}| = n)$ by

$$\begin{aligned} \breve{\boldsymbol{F}}_m(\boldsymbol{n}) = {} & U\left(n_F^{(m-1)} - |\boldsymbol{n}|\right) \breve{\boldsymbol{F}}_{m-1}(\boldsymbol{n})(\boldsymbol{I} + \theta^{-1}\boldsymbol{C}) \\ & + \theta^{-1} \sum_{k\in\mathcal{K}} \sum_{l_k=1}^{\min(n_k,\ n_g(k))} U\left(n_F^{(m-1)} - |\boldsymbol{n} - l_k\boldsymbol{e}_k|\right) \\ & \cdot \breve{\boldsymbol{F}}_{m-1}(\boldsymbol{n} - l_k\boldsymbol{e}_k) g_k(l_k) \boldsymbol{D}_k, \end{aligned} \tag{2.44}$$

until $\breve{\boldsymbol{F}}_m(\boldsymbol{n})$'s satisfy $\sum_{|\boldsymbol{n}|\le n^*} \breve{\boldsymbol{F}}_m(\boldsymbol{n})\boldsymbol{e} > (1 - \varepsilon_F)^m \boldsymbol{e}$ for some n^*, where $U(x)$ denotes a unit step function:

$$U(x) = \begin{cases} 1, & x \ge 0, \\ 0, & x < 0. \end{cases}$$

Let $n_F^{(m)} = n^*$ and go to Step (3–c).

Remark 2.2 *Note that* $\breve{\boldsymbol{A}}_k(\boldsymbol{n})$ *(*$|\boldsymbol{n}| \le n_A(k)$*) and* $\breve{\boldsymbol{v}}_k(\boldsymbol{n})$ *(*$|\boldsymbol{n}| \le n_v(k)$*) obtained by the above algorithm satisfy*

$$\begin{aligned}\breve{\boldsymbol{A}}_k(\boldsymbol{n}) &= \sum_{m=0}^{m_\gamma(k)} U\left(n_F^{(m)} - |\boldsymbol{n}|\right)\gamma_k^{(m)}(\theta)\breve{\boldsymbol{F}}_m(\boldsymbol{n}),\\ \breve{\boldsymbol{v}}_k(\boldsymbol{n}) &= \sum_{m=0}^{m_v(k)} U\left(n_F^{(m)} - |\boldsymbol{n}|\right)\boldsymbol{v}^{(m)}(\theta)\boldsymbol{D}_k\breve{\boldsymbol{F}}_m(\boldsymbol{n}),\end{aligned}$$

respectively, where

$$\begin{aligned}n_A(k) &= \max\left(n_F^{(m)}; m = 0, 1, \ldots, m_\gamma(k)\right),\\ n_v(k) &= \max\left(n_F^{(m)}; m = 0, 1, \ldots, m_v(k)\right).\end{aligned} \tag{2.45}$$

Remark 2.3 *If we are interested only in the* $\boldsymbol{p}(\boldsymbol{n})$ *(*$|\boldsymbol{n}| \le N_p$*) for some* N_p*, we do not need to compute* $\breve{\boldsymbol{F}}(\boldsymbol{n})$ *for* $\boldsymbol{n}$ *such that* $|\boldsymbol{n}| > N_p$*. Thus, in this case,* $n_g(k)$ *is redefined as* $\min(n_g(k), N_p)$ *and Step (3–e) is replaced by*

> **Step (3–e').** *For each* n *(*$n = 0, 1, \ldots$*), compute* $\breve{\boldsymbol{F}}_m(\boldsymbol{n})$*'s (*$|\boldsymbol{n}| = n$*,* $\boldsymbol{n} \in \mathcal{Z}$*) by (2.44) until* $\breve{\boldsymbol{F}}_m(\boldsymbol{n})$*'s satisfy*
>
> $$\sum_{|\boldsymbol{n}| \le n^*} \breve{\boldsymbol{F}}_m(\boldsymbol{n})\boldsymbol{e} > (1 - \varepsilon_F)^m \boldsymbol{e},$$
>
> *for some* n^**, or* $n = N_p$*, whichever occurs first. Let* $n_F^{(m)} = \min(n^*, N_p)$ *and go to Step (3–c).*

This procedure can save the computational cost, while maintaining the accuracy of the results.

We now provide two lemmas that ensure the above procedure eventually stops.

Lemma 2.6 *There exist integers* $m_\gamma(k)$ *and* $m_v(k)$ *satisfying (2.40) and (2.41), respectively.*

Proof. Substituting $1 - \varepsilon_F$ for z in (2.34), we have

$$\sum_{m=0}^{\infty} \gamma_k^{(m)}(\theta)(1 - \varepsilon_F)^m = H_k^*(\theta\varepsilon_F), \tag{2.46}$$

where $H_k^*(s)$ denotes the LST of $H_k(x)$. Similarly, from (2.29), we have

$$\sum_{m=0}^{\infty} \boldsymbol{v}^{(m)}(\theta)\boldsymbol{D}_k\boldsymbol{e}(1 - \varepsilon_F)^m = \boldsymbol{v}^*(\theta\varepsilon_F)\boldsymbol{D}_k\boldsymbol{e}. \tag{2.47}$$

Note here that

$$H_k^*(\theta\varepsilon_F) > 1 - h_k \cdot (\theta\varepsilon_F), \qquad k \in \mathcal{K}, \tag{2.48}$$

$$\boldsymbol{v}^*(\theta\varepsilon_F)\boldsymbol{D}_k\boldsymbol{e} > \lambda_k^{(\mathrm{B})} - \overline{\boldsymbol{v}}^{(1)}\boldsymbol{D}_k\boldsymbol{e} \cdot (\theta\varepsilon_F), \qquad k \in \mathcal{K}, \tag{2.49}$$

because $H_k^*(s)$ and each element of $\boldsymbol{v}^*(s)$ are convex functions of s. Note also that (2.39) is equivalent to

$$1 - h_k\theta\varepsilon_F > 1-\varepsilon, \qquad k \in \mathcal{K}, \tag{2.50}$$

$$\lambda_k^{(\mathrm{B})} - \overline{\boldsymbol{v}}^{(1)}\boldsymbol{D}_k\boldsymbol{e}\theta\varepsilon_F > (1-\varepsilon)\lambda_k^{(\mathrm{B})}, \qquad k \in \mathcal{K}. \tag{2.51}$$

It then follows from (2.46)–(2.51) that

$$\sum_{m=0}^{\infty} \gamma_k^{(m)}(\theta)(1-\varepsilon_F)^m > 1 - h_k\theta\varepsilon_F > 1-\varepsilon,$$

$$\sum_{m=0}^{\infty} \boldsymbol{v}^{(m)}(\theta)\boldsymbol{D}_k\boldsymbol{e}(1-\varepsilon_F)^m > \lambda_k^{(\mathrm{B})} - \overline{\boldsymbol{v}}^{(1)}\boldsymbol{D}_k\boldsymbol{e}\theta\varepsilon_F > (1-\varepsilon)\lambda_k^{(\mathrm{B})},$$

which complete the proof. ■

Lemma 2.7 *There exists an integer $n_F^{(m)}$ such that*

$$\sum_{\substack{\boldsymbol{n}\in\mathcal{Z}\\ |\boldsymbol{n}|\le n_F^{(m)}}} \breve{\boldsymbol{F}}_m(\boldsymbol{n})\boldsymbol{e} > (1-\varepsilon_F)^m\boldsymbol{e}, \qquad \forall m = 1,2,\ldots,m_{\max}. \tag{2.52}$$

Proof. We first consider the case $m = 1$. It follows from (2.42) and (2.43) that

$$\begin{aligned}\sum_{\substack{\boldsymbol{n}\in\mathcal{Z}\\ |\boldsymbol{n}|\le n_F^{(1)}}} \breve{\boldsymbol{F}}_1(\boldsymbol{n})\boldsymbol{e} &= \left[\boldsymbol{I} + \theta^{-1}\boldsymbol{C} + \theta^{-1}\sum_{k\in\mathcal{K}}\sum_{n_k=1}^{n_g(k)} g_k(n_k)\boldsymbol{D}_k\right]\boldsymbol{e}\\ &> \boldsymbol{e} + \theta^{-1}\boldsymbol{C}\boldsymbol{e} + \sum_{k\in\mathcal{K}}\left[\theta^{-1}\boldsymbol{D}_k\boldsymbol{e} - \frac{\varepsilon_g}{K}\boldsymbol{e}\right]\\ &= (1-\varepsilon_g)\boldsymbol{e} > (1-\varepsilon_F)\boldsymbol{e},\end{aligned} \tag{2.53}$$

where we use $(\boldsymbol{C}+\boldsymbol{D})\boldsymbol{e} = \boldsymbol{0}$ and $\varepsilon_g < \varepsilon_F$.

Suppose that for some m $(1 \le m \le m_{\max}-1)$, there exists an integer $n_F^{(m)}$ such that

$$\sum_{\substack{\boldsymbol{n}\in\mathcal{Z}\\ |\boldsymbol{n}|\le n_F^{(m)}}} \breve{\boldsymbol{F}}_m(\boldsymbol{n})\boldsymbol{e} > (1-\varepsilon_F)^m\boldsymbol{e}. \tag{2.54}$$

Using (2.43) and (2.44), we have for $\boldsymbol{n} \in \mathcal{Z}$,

$$\begin{aligned}\breve{\boldsymbol{F}}_{m+1}(\boldsymbol{n}) &= U\left(n_F^{(m)} - |\boldsymbol{n}|\right)\breve{\boldsymbol{F}}_m(\boldsymbol{n})\breve{\boldsymbol{F}}_1(\boldsymbol{0})\\ &\quad + \sum_{k\in\mathcal{K}}\sum_{l_k=1}^{\min(n_k,n_g(k))} U\left(n_F^{(m)} - |\boldsymbol{n} - l_k\boldsymbol{e}_k|\right)\breve{\boldsymbol{F}}_m(\boldsymbol{n} - l_k\boldsymbol{e}_k)\breve{\boldsymbol{F}}_1(l_k\boldsymbol{e}_k).\end{aligned} \tag{2.55}$$

It then follows from (2.53), (2.54) and (2.55) that

$$\sum_{\substack{\boldsymbol{n}\in\mathcal{Z}\\ |\boldsymbol{n}|\le n_F^{(m)}+n_F^{(1)}}} \breve{\boldsymbol{F}}_{m+1}(\boldsymbol{n})\boldsymbol{e} = \sum_{\substack{\boldsymbol{n}\in\mathcal{Z}\\ |\boldsymbol{n}|\le n_F^{(m)}}} \breve{\boldsymbol{F}}_m(\boldsymbol{n}) \sum_{\substack{\boldsymbol{n}\in\mathcal{Z}\\ |\boldsymbol{n}|\le n_F^{(1)}}} \breve{\boldsymbol{F}}_1(\boldsymbol{n})\boldsymbol{e} > (1-\varepsilon_F)^{m+1}\boldsymbol{e}. \tag{2.56}$$

Thus we can choose $n_F^{(m+1)}$ in such a way that

$$n_F^{(m+1)} \le n_F^{(m)} + n_F^{(1)}, \quad m = 1,2,\ldots,m_{\max}-1,$$

which completes the proof. ■

Theorem 2.5 *For $0 < \varepsilon < 1$, the $\breve{\boldsymbol{A}}_k(\boldsymbol{n})$ and the $\breve{\boldsymbol{v}}_k(\boldsymbol{n})$, computed by Step 3, satisfy error bounds (2.37) and (2.38), respectively.*

Proof. Using Lemma 2.6, Lemma 2.7 and (2.45), we obtain

$$\begin{aligned}
\sum_{\substack{\boldsymbol{n}\in\mathcal{Z}\\ |\boldsymbol{n}|\le n_A(k)}} \breve{\boldsymbol{A}}_k(\boldsymbol{n})\boldsymbol{e} &= \sum_{\substack{\boldsymbol{n}\in\mathcal{Z}\\ |\boldsymbol{n}|\le n_A(k)}} \sum_{m=0}^{m_\gamma(k)} U\left(n_F^{(m)} - |\boldsymbol{n}|\right)\gamma_k^{(m)}(\theta)\breve{\boldsymbol{F}}_m(\boldsymbol{n})\boldsymbol{e} \\
&= \sum_{m=0}^{m_\gamma(k)} \gamma_k^{(m)}(\theta) \sum_{\substack{\boldsymbol{n}\in\mathcal{Z}\\ |\boldsymbol{n}|\le n_F^{(m)}}} \breve{\boldsymbol{F}}_m(\boldsymbol{n})\boldsymbol{e} \\
&> \sum_{m=0}^{m_\gamma(k)} \gamma_k^{(m)}(\theta)(1-\varepsilon_F)^m\boldsymbol{e} > (1-\varepsilon)\boldsymbol{e}, \qquad k\in\mathcal{K}.
\end{aligned}$$

In the same way, we can obtain (2.38), so that the proof for the $\breve{\boldsymbol{v}}_k(\boldsymbol{n})$ is omitted. ■

Finally, we consider the $\boldsymbol{\Gamma}_k(\boldsymbol{n})$. Note here that

$$\sum_{\boldsymbol{n}\in\mathcal{Z}} \boldsymbol{\Gamma}_k(\boldsymbol{n})\boldsymbol{e} = \left\{(\boldsymbol{I}-\boldsymbol{P}_k)^{-1}\boldsymbol{e}(M_k)\right\}\otimes\boldsymbol{e}(M),$$

where $\boldsymbol{e}(m)$ denotes an $m\times 1$ vector whose elements are all equal to one. Keeping the above equation in mind, we propose to compute an approximation $\breve{\boldsymbol{\Gamma}}_k(\boldsymbol{n})$ to $\boldsymbol{\Gamma}_k(\boldsymbol{n})$ in the following way.

Step 4. For each $k\in\mathcal{K}$, compute $\breve{\boldsymbol{\Gamma}}_k(\boldsymbol{n})$'s ($|\boldsymbol{n}| = n$) for $n = 0, 1, \ldots$ by

$$\breve{\boldsymbol{\Gamma}}_k(\boldsymbol{0}) = \left[\boldsymbol{I}-\boldsymbol{P}_k\otimes\breve{\boldsymbol{A}}_k(\boldsymbol{0})\right]^{-1}, \tag{2.57}$$

$$\breve{\boldsymbol{\Gamma}}_k(\boldsymbol{n}) = \sum_{\substack{\boldsymbol{0}\le\boldsymbol{l}\le\boldsymbol{n}\\ \boldsymbol{l}\ne\boldsymbol{0}}} U\left(n_A(k)-|\boldsymbol{l}|\right)\breve{\boldsymbol{\Gamma}}_k(\boldsymbol{n}-\boldsymbol{l})\left[\boldsymbol{P}_k\otimes\breve{\boldsymbol{A}}_k(\boldsymbol{l})\right]\breve{\boldsymbol{\Gamma}}_k(\boldsymbol{0}), \quad \boldsymbol{n}\in\mathcal{Z}^+, \tag{2.58}$$

until $\breve{\boldsymbol{\Gamma}}_k(\boldsymbol{n})$'s satisfy

$$\begin{aligned}
\sum_{\substack{\boldsymbol{n}\in\mathcal{Z}\\ |\boldsymbol{n}|\le n_\Gamma(k)}} \breve{\boldsymbol{\Gamma}}_k(\boldsymbol{n})\boldsymbol{e} &> \left\{(\boldsymbol{I}-\boldsymbol{P}_k)^{-1}\boldsymbol{e}(M_k)\right\}\otimes\boldsymbol{e}(M) \\
&\quad -\varepsilon\left\{(\boldsymbol{I}-\boldsymbol{P}_k)^{-2}\boldsymbol{P}_k\boldsymbol{e}(M_k)\right\}\otimes\boldsymbol{e}(M),
\end{aligned} \tag{2.59}$$

for some integer $n_\Gamma(k)$.

Remark 2.4 *Let G_k ($k\in\mathcal{K}$) denote a generic random variable representing a batch size of class k. We then have*

$$(\boldsymbol{\alpha}_k\otimes\boldsymbol{\pi})\sum_{\boldsymbol{n}\in\mathcal{Z}}\boldsymbol{\Gamma}_k(\boldsymbol{n})\boldsymbol{e} = \mathrm{E}[G_k],$$

and if (2.59) satisfies for some $n_\Gamma(k)$,

$$(\boldsymbol{\alpha}_k\otimes\boldsymbol{\pi})\sum_{\substack{\boldsymbol{n}\in\mathcal{Z}\\ |\boldsymbol{n}|\le n_\Gamma(k)}}\breve{\boldsymbol{\Gamma}}_k(\boldsymbol{n})\boldsymbol{e} > \mathrm{E}[G_k] - \frac{1}{2}\mathrm{E}[G_k(G_k-1)]\varepsilon.$$

Lemma 2.8 *Suppose $\breve{\boldsymbol{A}}_k(\boldsymbol{n})$ satisfies (2.37). Then there exists $n_\Gamma(k)$ satisfying (2.59).*

Proof. From (2.57) and (2.58), it can be seen that $\breve{\boldsymbol{A}}_k(\boldsymbol{n})$ ($|\boldsymbol{n}| \le n_A(k)$) and $\breve{\boldsymbol{\Gamma}}_k(\boldsymbol{n})$ ($\boldsymbol{n} \in \mathcal{Z}$) are related by

$$\sum_{\boldsymbol{n}\in\mathcal{Z}} \breve{\boldsymbol{\Gamma}}_k(\boldsymbol{n}) = \sum_{m=0}^{\infty} \left(\boldsymbol{P}_k \otimes \sum_{\substack{\boldsymbol{l}\in\mathcal{Z} \\ |\boldsymbol{l}|\le n_A(k)}} \breve{\boldsymbol{A}}_k(\boldsymbol{l}) \right)^m .$$

Post-multiplying both sides of the above equation by $\boldsymbol{e} = \boldsymbol{e}(M_k) \otimes \boldsymbol{e}(M)$, we have

$$\begin{aligned} \sum_{\boldsymbol{n}\in\mathcal{Z}} \breve{\boldsymbol{\Gamma}}_k(\boldsymbol{n})\boldsymbol{e} &= \sum_{m=0}^{\infty} \left(\boldsymbol{P}_k \otimes \sum_{\substack{\boldsymbol{l}\in\mathcal{Z} \\ |\boldsymbol{l}|\le n_A(k)}} \breve{\boldsymbol{A}}_k(\boldsymbol{l}) \right)^m \cdot [\boldsymbol{e}(M_k) \otimes \boldsymbol{e}(M)] \\ &= \sum_{m=0}^{\infty} [\boldsymbol{P}_k^m \boldsymbol{e}(M_k)] \otimes \left[\left(\sum_{\substack{\boldsymbol{l}\in\mathcal{Z} \\ |\boldsymbol{l}|\le n_A(k)}} \breve{\boldsymbol{A}}_k(\boldsymbol{l}) \right)^m \boldsymbol{e}(M) \right]. \end{aligned}$$

Further, using (2.37), we obtain

$$\begin{aligned} \sum_{\boldsymbol{n}\in\mathcal{Z}} \breve{\boldsymbol{\Gamma}}_k(\boldsymbol{n})\boldsymbol{e} &> \sum_{m=0}^{\infty} (1-\varepsilon)^m [\boldsymbol{P}_k^m \boldsymbol{e}(M_k)] \otimes \boldsymbol{e}(M) \\ &> \sum_{m=0}^{\infty} (1-m\varepsilon) [\boldsymbol{P}_k^m \boldsymbol{e}(M_k)] \otimes \boldsymbol{e}(M) \\ &= \left\{ (\boldsymbol{I}-\boldsymbol{P}_k)^{-1} \boldsymbol{e}(M_k) \right\} \otimes \boldsymbol{e}(M) \\ &\qquad - \varepsilon \left\{ (\boldsymbol{I}-\boldsymbol{P}_k)^{-2} \boldsymbol{P}_k \boldsymbol{e}(M_k) \right\} \otimes \boldsymbol{e}(M), \end{aligned}$$

which completes the proof. ■

2.6 Numerical Examples

In this section, we show some numerical examples for queues with two arrival streams. Even though the algorithmic analysis has already been done for the single arrival cases [Taki01a, Taki01b], no numerical examples were shown there. Thus the numerical result provided below is the first report in the literature, as for FIFO queues with multiple Markovian arrival streams having different service time distributions.

In all numerical examples, the counting process of class k ($k = 1, 2$) arrivals follows a batch interrupted Poisson process with geometrically distributed batch sizes with mean g. Namely, the counting process of class k ($k = 1, 2$) is characterized by $(\widetilde{\boldsymbol{C}}_k, \widetilde{\boldsymbol{D}}_k(n))$, where

$$\begin{aligned} \widetilde{\boldsymbol{C}}_k &= \begin{bmatrix} -2\lambda_k g^{-1} - 0.1 & 0.1 \\ 0.1 & -0.1 \end{bmatrix}, \\ \widetilde{\boldsymbol{D}}_k(n) &= (1-p)p^{n-1} \begin{bmatrix} 2\lambda_k g^{-1} & 0 \\ 0 & 0 \end{bmatrix}, \quad n = 1, 2, \ldots, \end{aligned}$$

where $p = 1 - 1/g$. Note that the arrival rate of class k is fixed to be λ_k regardless of the mean batch size g.

We now consider three types of the superposition of these two streams.

[Case P]

$$\boldsymbol{C} = \begin{bmatrix} -2(\lambda_1+\lambda_2)g^{-1} - 0.1 & 0.1 \\ 0.1 & -0.1 \end{bmatrix},$$

and for $n = 1, 2, \ldots$,

$$\begin{aligned} \boldsymbol{D}_1(n) &= (1-p)p^{n-1} \begin{bmatrix} 2\lambda_1 g^{-1} & 0 \\ 0 & 0 \end{bmatrix}, \\ \boldsymbol{D}_2(n) &= (1-p)p^{n-1} \begin{bmatrix} 2\lambda_2 g^{-1} & 0 \\ 0 & 0 \end{bmatrix}. \end{aligned}$$

[Case I]

$$\boldsymbol{C} = \widetilde{\boldsymbol{C}}_1 \oplus \widetilde{\boldsymbol{C}}_2,$$

and for $n = 1, 2, \ldots$,

$$\boldsymbol{D}_1(n) = \widetilde{\boldsymbol{D}}_1(n) \otimes \boldsymbol{I}(2), \quad \boldsymbol{D}_2(n) = \boldsymbol{I}(2) \otimes \widetilde{\boldsymbol{D}}_2(n),$$

where $\oplus$ denotes the Kronecker sum, and

[Case N]

$$\boldsymbol{C} = \begin{bmatrix} -2\lambda_1 g^{-1} - 0.1 & 0.1 \\ 0.1 & -2\lambda_2 g^{-1} - 0.1 \end{bmatrix},$$

and for $n = 1, 2, \ldots$,

$$\begin{aligned} \boldsymbol{D}_1(n) &= (1-p)p^{n-1} \begin{bmatrix} 2\lambda_1 g^{-1} & 0 \\ 0 & 0 \end{bmatrix}, \\ \boldsymbol{D}_2(n) &= (1-p)p^{n-1} \begin{bmatrix} 0 & 0 \\ 0 & 2\lambda_2 g^{-1} \end{bmatrix}. \end{aligned}$$

Note that in Case P, two arrival streams are positively correlated, in Case I, they are independent each other and in Case N, they are negatively correlated. As for the service time distributions, we consider two cases, Case GD (class-dependent service times) and Case GI (i.i.d. service times):

[Case GD]

$$H_1 = 1, \text{ with prob. } 1, \quad H_2 = 4, \text{ with prob. } 1,$$

[Case GI]

$$H_k = \begin{cases} 1, & \text{with prob. } \lambda_1/(\lambda_1+\lambda_2), \\ 4, & \text{with prob. } \lambda_2/(\lambda_1+\lambda_2), \end{cases} \qquad k = 1, 2,$$

where H_k $(k = 1, 2)$ denotes a generic random variable for a service time of a class k customer. Note that the overall service time distributions are identical in both cases. We denote the queueing model with Case i (i = P, I, N) arrivals and Case j (j = GD, GI) services by Case (i, j).

In what follows, we consider two examples, Examples 1 and 2, within the above settings. In Example 1, we set $\lambda_1 = \lambda_2 = 0.15$, so that $\rho_1 = 0.15$ and $\rho_2 = 0.6$ in Case (i, GD) (i = P, I, N), and $\rho_1 = \rho_2 = 0.375$ in Case (i, GI) (i = P, I, N). On the other hand, in Example 2, we set $\lambda_1 = 0.4$ and $\lambda_2 = 0.1$, so that $\rho_1 = \rho_2 = 0.4$ in Case (i, GD) (i = P, I, N) and that $\rho_1 = 0.64$ and $\rho_2 = 0.16$ in Case (i, GI) (i = P, I, N).

2.6.1 Efficiency of the algorithm

Before showing the quantitative behavior of the queue length distribution, we discuss the efficiency of our numerical algorithm for the $\boldsymbol{F}_m(\boldsymbol{n})$. It follows from (2.21) that for $m = 0, 1, \ldots,$

$$\sum_{\boldsymbol{n}\in\mathcal{Z}} \boldsymbol{F}_m(\boldsymbol{n}) = \left[\boldsymbol{I} + \theta^{-1}\left(\boldsymbol{C} + \sum_{k\in\mathcal{K}}\sum_{n_k=1}^{\infty} \boldsymbol{D}_k(n_k)\right)\right]^m, \tag{2.60}$$

where $\boldsymbol{I} + \theta^{-1}[\boldsymbol{C} + \sum_{k\in\mathcal{K}}\sum_{n_k=1}^{\infty}\boldsymbol{D}_k(n_k)]$ is a stochastic matrix. Thus a straightforward implementation of the recursion for the $\boldsymbol{F}_m(\boldsymbol{n})$ in (2.24) and (2.25) would be the following. We first truncate the $\boldsymbol{D}_k(n_k)$ at $n_k = n'_g(k)$ in such a way that

$$\theta^{-1}\sum_{n_k=1}^{n'_g(k)} \boldsymbol{D}_k(n_k)\boldsymbol{e} > \theta^{-1}\boldsymbol{D}_k\boldsymbol{e} - \frac{\varepsilon'_g}{K}\boldsymbol{e},$$

so that

$$\left[\boldsymbol{I} + \theta^{-1}\left(\boldsymbol{C} + \sum_{k\in\mathcal{K}}\sum_{n_k=1}^{n'_g(k)} \boldsymbol{D}_k(n_k)\right)\right]\boldsymbol{e} > (1 - \varepsilon'_g)\boldsymbol{e}.$$

We then compute all terms obtained by expanding the right hand side of (2.60) with the truncated $\boldsymbol{D}_k(n_k)$ $(k \in \mathcal{K})$. Note that if $\varepsilon'_g = \varepsilon_F$, the resulting $\boldsymbol{F}_m(\boldsymbol{n})$ satisfies (2.52) in Lemma 2.7, where the summation on the left hand side of (2.52) is taken for all computed $\boldsymbol{F}_m(\boldsymbol{n})$'s.

In Table 2.1, we show the numbers of $\boldsymbol{F}_m(\boldsymbol{n})$'s computed by our algorithm and the above straightforward implementation, using Example 1, where we set $\varepsilon = 10^{-6}$, ε_F = r.h.s. of (2.39)$\times\varepsilon/2$ and $\varepsilon_g = \varepsilon_F/10$. We observe that for unbounded batch size cases (i.e., $g > 1$), the number of the computed $\boldsymbol{F}_m(\boldsymbol{n})$'s in our algorithm is less than that in the straightforward algorithm about by three order of magnitude. Thus, compared to the straightforward implementation, our algorithm is very efficient in terms of the computational time when the batch size is unbounded.

Table 2.1: Number of computed $\boldsymbol{F}_m(\boldsymbol{n})$'s in Example 1.
A : Our algorithm B : Straightforward

Case		$g=1$	$g=2$	$g=5$	$g=10$
(P, GD)	A	1.021×10^6	3.918×10^6	2.504×10^7	1.257×10^8
	B	2.453×10^6	2.107×10^9	3.696×10^{10}	4.243×10^{11}
(P, GI)	A	1.021×10^6	2.898×10^6	1.476×10^7	6.732×10^7
	B	2.453×10^6	1.438×10^9	1.812×10^{10}	1.825×10^{11}
(I, GD)	A	6.108×10^5	3.314×10^6	3.012×10^7	1.854×10^8
	B	1.993×10^6	2.919×10^9	9.286×10^{10}	1.649×10^{12}
(I, GI)	A	4.123×10^5	1.859×10^6	1.532×10^7	9.032×10^7
	B	1.253×10^6	1.292×10^9	3.741×10^{10}	6.173×10^{11}
(N, GD)	A	6.657×10^4	8.895×10^5	1.165×10^7	8.095×10^7
	B	1.378×10^5	2.620×10^8	1.066×10^{10}	1.955×10^{11}
(N, GI)	A	1.411×10^4	3.113×10^5	4.813×10^6	3.540×10^7
	B	2.743×10^4	7.158×10^7	3.278×10^9	6.442×10^{10}

We note that a very huge memory space is required to store all $\breve{\boldsymbol{F}}_m(\boldsymbol{n})$'s in some cases, even using our truncation and stopping criteria. For example, in Case (I, GD) with $g = 10$, the memory space to store all $\breve{\boldsymbol{F}}_m(\boldsymbol{n})$'s is given by $16\times1.854\times10^8\times8$ bytes ≈ 23.73 Gbytes, because each $\breve{\boldsymbol{F}}_m(\boldsymbol{n})$ is a 4×4 matrix and one element requires 8 bytes in double precision. Thus in our implementation, every time $\breve{\boldsymbol{F}}_m(\boldsymbol{n})$'s for each m are obtained, we compute the contributions of $\breve{\boldsymbol{F}}_m(\boldsymbol{n})$'s to $\breve{\boldsymbol{A}}_k(\boldsymbol{n})$ and $\breve{\boldsymbol{v}}_k(\boldsymbol{n})$ in Step (3–c), and discard all $\breve{\boldsymbol{F}}_{m-1}(\boldsymbol{n})$'s.

Table 2.2 shows the maximum number of $\breve{\boldsymbol{F}}_m(\boldsymbol{n})$'s stored temporarily in our algorithm, where the ratio of it to the total number of computed $\breve{\boldsymbol{F}}_m(\boldsymbol{n})$'s is also shown in parenthesis. We observe that in most cases, the number of temporarily stored $\breve{\boldsymbol{F}}_m(\boldsymbol{n})$'s is a few percent of the total number of computed ones. Thus our implementation is expected to save the required memory space, especially when a large number of $\breve{\boldsymbol{F}}_m(\boldsymbol{n})$'s should be computed.

2.6.2 Number of customers in Example 1

Figures 2.1–2.3 plot the complementary distributions of the total number N of customers in Case (i, GD) and Case (i, GI) (i = P, I, N), where the batch size is fixed to be one, i.e., $g = 1$. Note that the overall input processes in Case (P, GD) and Case (P, GI) are identical, so that the distributions of the total number of customers are also identical, as shown in Figure 2.1. However, as shown in Table 2.3, the joint queue length distributions in these two cases are different. Note also that in Case (P, GI), $\boldsymbol{p}(n_1, n_2)\boldsymbol{e} = \boldsymbol{p}(n_2, n_1)\boldsymbol{e}$, because the conditional joint distribution $\Pr(N_1 = n_1, N_2 = n_2 \mid N_1 + N_2 = n_1 + n_2)$ follows a binomial distribution with parameter 0.5. We also observe that $\boldsymbol{p}(n, n)\boldsymbol{e}$'s in both cases take the same value for each n. Unfortunately, we cannot provide any intuitive explanation of this phenomenon.

Table 2.2: Number of stored $\breve{\boldsymbol{F}}_m(\boldsymbol{n})$'s in Example 1.

Case	$g = 1$	$g = 2$	$g = 5$	$g = 10$
(P, GD)	27225	90601	455625	1651227
	(2.67%)	(2.31%)	(1.82%)	(1.31%)
(P, GI)	27225	75076	330051	1125723
	(2.67%)	(2.59%)	(2.24%)	(1.67%)
(I, GD)	16641	68121	399424	1548781
	(2.72%)	(2.06%)	(1.33%)	(0.84%)
(I, GI)	13110	47524	263683	997003
	(3.18%)	(2.56%)	(1.72%)	(1.10%)
(N, GD)	4970	41209	324331	1387686
	(7.47%)	(4.63%)	(2.78%)	(1.71%)
(N, GI)	1764	21171	187491	833571
	(12.51%)	(6.80%)	(3.90%)	(2.35%)

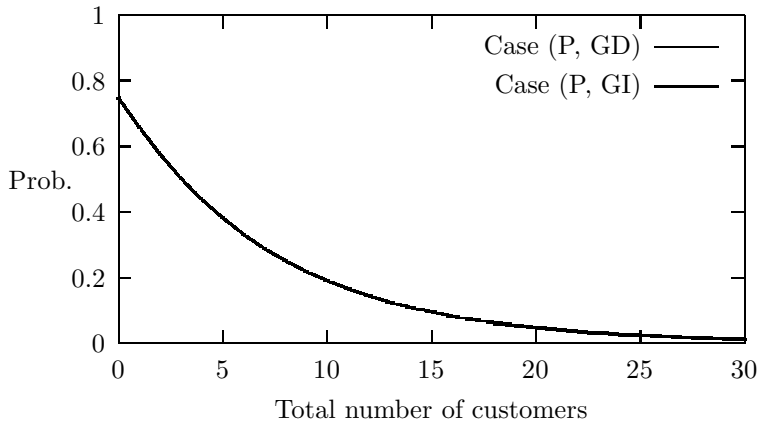

Figure 2.1: Complementary distribution of total number of customers in Example 1.

From Figures 2.2 and 2.3, we observe that class-dependent service times cause longer tails in the total queue length distributions, in these specific examples. We shall explain this phenomenon for Case N. In Case (N, GD), the conditional expected amounts of work brought into the system per unit time given the state of the underlying Markov chain are different, and they are given by 0.3 and 1.2, respectively. Thus in Case (N, GD), the system is overloaded during a half of time. On the other hand, in Case (N, GI), the conditional expected amount of work brought into the system per unit time is fixed to be 0.75, regardless of the state of the underlying Markov chain. Therefore the distribution of the total number of customers in Case (N, GD) has a longer tail than that in Case (N, GI).

Next, we consider the expected total number $\mathrm{E}[N]$ of customers as a function of the mean batch

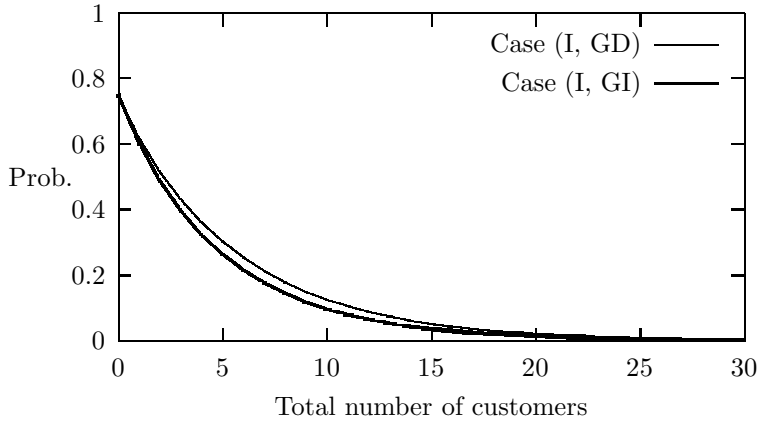

Figure 2.2: Complementary distribution of total number of customers in Example 1.

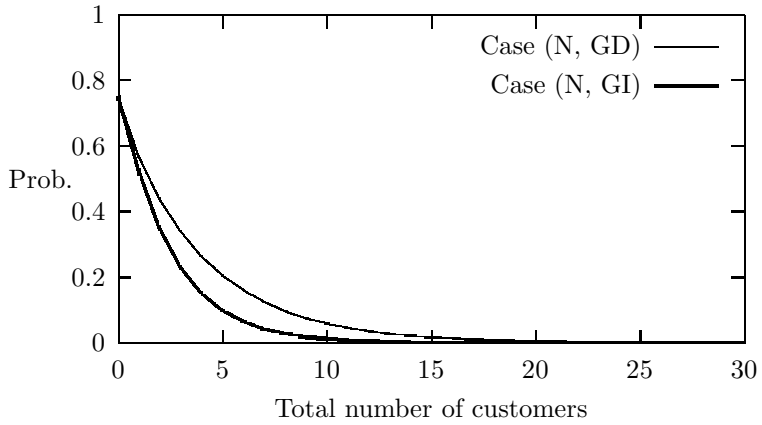

Figure 2.3: Complementary distribution of total number of customers in Example 1.

size g. Table 2.4 shows E[N] for the mean batch size $g = 1$, 2, 3, 4, 5 and 10. We observe that E[N] increases with the mean batch size g in all cases. This phenomenon comes from the fact that the deviation of the amount of work brought into the system per unit time increases with g. We also observe that for a fixed g, the positive correlation in the two streams leads to a larger E[N] in both Cases GD and GI, as expected.

2.6.3 Number of customers in Example 2

Table 2.5 shows the expected total number E[N] of customers for the mean batch size $g = 1$, 2, 3, 4, 5. and 10. We first examine the case of $g = 1$. Contrary to Example 1, we observe that the class-dependent service time (Case GD) decreases the expected total number of customers in Cases I and N. This phenomenon can be explained in a similar way to Example 1. For example, in Case (N, GD), the conditional expected amount of work brought into the system per unit time is fixed to be 0.8, regardless of the state of the underlying Markov chain. On the other hand, in Case (N, GI),

Table 2.3: Joint queue length distribution $\boldsymbol{p}(n_1, n_2)\boldsymbol{e}$.
(Upper rows for Case (P, GD) and lower rows for Case (P, GI))

n_1 / n_2	0	1	2	3
0	2.500×10^{-1}	2.472×10^{-2}	8.593×10^{-3}	3.481×10^{-3}
	2.500×10^{-1}	4.501×10^{-2}	2.054×10^{-2}	9.224×10^{-3}
1	6.530×10^{-2}	4.108×10^{-2}	2.193×10^{-2}	1.118×10^{-2}
	4.501×10^{-2}	4.108×10^{-2}	2.767×10^{-2}	1.629×10^{-2}
2	3.249×10^{-2}	3.341×10^{-2}	2.444×10^{-2}	
	2.054×10^{-2}	2.767×10^{-2}	2.444×10^{-2}	
3	1.497×10^{-2}	2.141×10^{-2}		
	9.224×10^{-3}	1.629×10^{-2}		
4	6.630×10^{-3}			
	4.073×10^{-3}			

Table 2.4: Expected total number of customers in Example 1.

Case	$g = 1$	$g = 2$	$g = 3$	$g = 4$	$g = 5$	$g = 10$
(P, GD)	5.8760	9.9815	13.9356	17.8320	21.7001	40.8865
(P, GI)	5.8760	9.1466	12.2898	15.3793	18.4408	33.5873
(I, GD)	4.5417	8.5777	12.4865	16.3524	20.1987	39.3295
(I, GI)	4.0010	7.1857	10.2714	13.3219	16.3555	31.4326
(N, GD)	3.2822	7.2033	11.0527	14.8822	18.7035	37.7739
(N, GI)	2.2800	5.2800	8.2800	11.2800	14.2800	29.2800

the conditional expected amounts of work brought into the system per unit time given the state of the underlying Markov chain are different, and they are given by 1.28 and 0.32, respectively. Thus in Case (N, GI), the system is overloaded during a half of time, so that E[N] in Case (N, GI) is greater than that in Case (N, GD).

We observe that in any case, the expected total number of customer increases with the mean batch size g, as in Example 1, and that E[N] in Case GD eventually becomes greater than E[N] in Case GI. We also observe that for a fixed g, the positive correlation in the two streams leads to a larger E[N] in both Cases GD and GI, as in Example 1.

2.7 Concluding Remarks

We developed a numerically feasible procedure to compute the joint queue length distribution in a FIFO single-server queue with multiple batch Markovian arrival streams having different service time distributions, assuming discrete phase-type batch size distributions. We established several

Table 2.5: Expected total number of customers in Example 2.

Case	$g = 1$	$g = 2$	$g = 3$	$g = 4$	$g = 5$	$g = 10$
(P, GD)	11.5019	17.7712	23.8347	29.8053	35.7261	65.0310
(P, GI)	11.5019	15.9366	20.1933	24.3657	28.4904	48.8117
(I, GD)	7.1517	13.3270	19.3007	25.2050	31.0760	60.2474
(I, GI)	8.7304	13.1052	17.3093	21.4407	25.5333	45.7640
(N, GD)	3.2168	9.0326	14.8425	20.6497	26.4551	55.4705
(N, GI)	6.0892	10.3399	14.4641	18.5407	22.5933	42.7206

truncation and stopping criteria to ensure numerical accuracy in the final result.

Note, however, that the computation of the joint queue length distribution is intensive by nature, especially when the number of classes is large. Even in such a case, the stationary total queue length distribution can be readily computed by modifying our algorithm. For the sake of completeness, we show algorithm steps for the total queue length distribution in Appendix C. Note here that the algorithm to compute $\boldsymbol{A}_k^{(\mathrm{T})}(n)$ in (C.1) for the number of arrivals in a service time can be also used for the computation of $\boldsymbol{A}_n$, which is a matrix for the number of arrivals per service in an ordinary BMAP/GI/1 queue (see Appendix B), because $\boldsymbol{A}_k^{(\mathrm{T})}(n)$ ($k \in \mathcal{K}$) in the case of $\mathcal{K} = \{1\}$ corresponds to $\boldsymbol{A}_n$ in the BMAP/GI/1 queue (see (B.2) in Appendix B and (C.2) in Appendix C). To the best of our knowledge, there is no work to consider the truncation and stopping criterion to compute $\boldsymbol{A}_n$ in the BMAP/GI/1 queue. Thus our development also contributes to the standard algorithm for the BMAP/GI/1 queue.

Chapter 3

FIFO Single-Server Queue with Service Interruptions

3.1 Introduction

This chapter considers a FIFO single-server queue with service interruptions. In such a queue, the state of the server changes *on* and *off* alternately. While in on-state, the server is available for service. On the other hand, while in off-state, the server dose not work even if customers are present in the system. Hereafter periods during which the server is in on-state (resp. off-state) are called on-periods (resp. off-periods).

Queues with service interruptions have many applications in the fields of manufacturing, computer and telecommunications systems, and many studies on those queues have been done for a few decades. A detailed survey on queues with service interruptions can be found in the introduction of the paper by Federgruen and Green [Fede86]. They mainly discussed approximation methods for an M/G/1 queue with service interruptions, where on- and off-periods are generally distributed [Fede86]. Further, assuming a phase-type on-period distribution, they established an exact algorithm to compute the steady-state queue length distribution [Fede88].

Recently, more general queues with service interruptions have been studied. Sengupta [Seng90] considered the model where on- and off-period distributions are general, customers arrive according to a Poisson process whose arrival rate depends on the server state, and service times are generally distributed, depending on the server state upon arrival. He showed that the amount of unfinished work in such a queue is closely related to the waiting time in a special GI/G/1 queue. Also, Takine and Sengupta [Taki97] considered a single-server queue with service interruptions, where both the server state and arrival processes are governed by a finite-state Markov chain. Namely, the marginal processes of the server state and customer arrivals form an alternating phase-type Markov renewal process and a MAP (see subsection 1.1.1), respectively, and they may be dependent. For this queue, they obtained the steady-state queue length distribution. The crucial assumption posed in [Taki97] was i.i.d. (independent and identically distributed) service times.

This chapter considers an extension of the results in [Taki97] to allow multiple batch Markovian arrival streams. Thus, the marginal arrival process follows a batch MMAP (see section 2.2).

Further, service times of customers can depends on both their arrival stream and the server state on arrival. As stated in [Taki97], such a queue cannot be analyzed by the conventional M/G/1 paradigm (see subsection 1.2.2). To analyze the extended model, therefore, we use a new approach developed in [Masu03a, Taki01a, Taki01b, Taki01c], which is based on the invariant relationship of the joint queue length distributions at a random point in time and at departures [Taki01a]. We then derive the vector joint generating function of the numbers of customers from respective arrival streams. Further assuming discrete phase-type batch size distributions as in subsection 2.4.3, we provide a computational algorithm for the steady-state joint queue length distribution. We also show some numerical examples and examine the impact of system parameters on the queue length distribution.

The rest of this chapter is divided into four sections. In section 2, the mathematical model is described. Section 3 briefly discusses the sojourn time distribution. In section 4, we first derive a general formula for the joint queue length distribution, and assuming discrete phase-type batch size distributions, we show recursive formulas to compute the joint queue length distribution. Finally, in section 5, we show some numerical examples.

3.2 Model

We consider a FIFO single-server queue with service interruptions. The state of the server changes on and off alternately, and while the server is being on, customers are served successively. On the other hand, services of customers stop temporally while the server is being off, and interrupted services are restarted in a preemptive-resume manner when the server becomes on again. In what follows, we call the process of the server state the on-off process.

We assume that both the on-off and arrival processes are governed by an underlying finite-state Markov chain that is assumed to be irreducible. Let $\mathcal{M} = \{1, \ldots, M\}$ denote the state space of the underlying Markov chain, where $M \geq 2$. It stays in state i ($i \in \mathcal{M}$) for an exponential interval of time with mean μ_i^{-1}, and when the sojourn time in state i has elapsed, the underlying Markov chain changes its state to state j with probability $\sigma_{i,j}$ ($j \in \mathcal{M}$), where

$$\sum_{j \in \mathcal{M}} \sigma_{i,j} = 1.$$

The on-off process of the server is defined in the following way. The state space $\mathcal{M}$ is divided into two disjoint sub-spaces, $\mathcal{M}_{\text{on}} = \{1, \ldots, M_{\text{on}}\}$ and $\mathcal{M}_{\text{off}} = \{M_{\text{on}} + 1, \ldots, M_{\text{on}} + M_{\text{off}}\}$, where $M_{\text{on}} \geq 1$, $M_{\text{off}} \geq 1$ and $M_{\text{on}} + M_{\text{off}} = M$. The server is assumed to be on (resp. off) while the underlying Markov chain is being in state $i \in \mathcal{M}_{\text{on}}$ (resp. $i \in \mathcal{M}_{\text{off}}$). Thus the on-off process forms a phase-type alternating Markov renewal process.

Next we describe the arrival process of customers. We assume that there are K ($K \geq 1$) arrival streams. Let $\mathcal{K}$ denote a set of class indices, i.e., $\mathcal{K} = \{1, \ldots, K\}$. Customers arriving from the kth ($k \in \mathcal{K}$) arrival stream are called class k customers. Given a state transition of the underlying Markov chain from state i to state j ($i, j \in \mathcal{M}$), n ($n = 1, 2, \ldots$) customers of class k ($k \in \mathcal{K}$) arrive

in batch with probability $\sigma_{k,i,j}(n)/\sigma_{i,j}$, where

$$\sum_{k\in\mathcal{K}}\sum_{n=1}^{\infty} \sigma_{k,i,j}(n) \le \sigma_{i,j},$$

for all $i, j \in \mathcal{M}$. Note that customers in the same batch belong to the same class. For later use, we define $\sigma_{i,j}(0)$ $(i, j \in \mathcal{M})$ as

$$\sigma_{i,j}(0) = \sigma_{i,j} - \sum_{k\in\mathcal{K}}\sum_{n=1}^{\infty} \sigma_{k,i,j}(n).$$

Note that $\sigma_{i,j}(0)/\sigma_{i,j}$ represents the conditional probability of no arrivals given that a state transition from state i to state j happens. Without loss of generality, we assume $\sigma_{i,i}(0) = 0$ for all i $(i \in \mathcal{M})$.

In terms of service times, class k customers are further classified into two sub-classes based on the server state on arrival. We call class k customers arriving in on-periods (resp. off-periods) class k-on (resp. k-off) customers. Service times of class k-on (resp. k-off) customers are assumed to be i.i.d. according to a distribution function $H_{k,\mathrm{on}}(x)$ (resp. $H_{k,\mathrm{off}}(x)$) with finite mean $h_{k,\mathrm{on}}$ (resp. $h_{k,\mathrm{off}}$).

In the rest of this chapter, we impose two assumptions on $\sigma_{k,i,j}(n)$. For each k $(k \in \mathcal{K})$, there exists at least one triad (i, j, n) $(i, j \in \mathcal{M},\ n = 1, 2, \ldots)$ such that $\sigma_{k,i,j}(n) > 0$. Thus arrivals of class k customers are certain. Further $\sigma_{k,i,j}(n) = 0$ $(k \in \mathcal{K})$ if $i \in \mathcal{M}_{\mathrm{on}}$ and $j \in \mathcal{M}_{\mathrm{off}}$ or if $i \in \mathcal{M}_{\mathrm{off}}$ and $j \in \mathcal{M}_{\mathrm{on}}$. Thus arrivals of customers and changes of the server state never happen simultaneously.

We now introduce some notations. Let $\boldsymbol{C}$ denote an $M \times M$ matrix whose (i, j)th $(i, j \in \mathcal{M})$ element $C_{i,j}$ is given by

$$C_{i,j} = \begin{cases} -\mu_i, & \text{if } i = j, \\ \sigma_{i,j}(0)\mu_i, & \text{otherwise.} \end{cases}$$

For each $k \in \mathcal{K}$, let $\boldsymbol{D}_k(n)$ $(n = 1, 2, \ldots)$ denote an $M \times M$ matrix whose (i, j)th $(i, j \in \mathcal{M})$ element $D_{k,i,j}(n)$ is given by

$$D_{k,i,j}(n) = \begin{cases} \sigma_{k,i,j}(n)\mu_i, & \text{if } i, j \in \mathcal{M}_{\mathrm{on}} \text{ or } i, j \in \mathcal{M}_{\mathrm{off}}, \\ 0, & \text{otherwise.} \end{cases}$$

Then the on-off and arrival processes are characterized by $\boldsymbol{C}$ and $\boldsymbol{D}_k(n)$ $(k \in \mathcal{K}, n = 1, 2, \ldots)$, and the marginal arrival process is a batch MMAP (see section 2.2). Note here that $\boldsymbol{C}$ and $\boldsymbol{D}_k(n)$ have the following structure:

$$\boldsymbol{C} = \begin{bmatrix} \boldsymbol{C}_{\mathrm{on}} & \boldsymbol{E}_{\mathrm{on,off}} \\ \boldsymbol{E}_{\mathrm{off,on}} & \boldsymbol{C}_{\mathrm{off}} \end{bmatrix}, \qquad \boldsymbol{D}_k(n) = \begin{bmatrix} \boldsymbol{D}_{k,\mathrm{on}}(n) & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{D}_{k,\mathrm{off}}(n) \end{bmatrix},$$

where $\boldsymbol{C}_{\mathrm{on}}$ and $\boldsymbol{D}_{k,\mathrm{on}}(n)$ are $M_{\mathrm{on}} \times M_{\mathrm{on}}$ matrices, $\boldsymbol{C}_{\mathrm{off}}$ and $\boldsymbol{D}_{k,\mathrm{off}}(n)$ are $M_{\mathrm{off}} \times M_{\mathrm{off}}$ matrices, and $\boldsymbol{E}_{\mathrm{on,off}}$ and $\boldsymbol{E}_{\mathrm{off,on}}$ are $M_{\mathrm{on}} \times M_{\mathrm{off}}$ and $M_{\mathrm{off}} \times M_{\mathrm{on}}$ matrices, respectively.

We define $\boldsymbol{D}_{k,\xi}$ ($k \in \mathcal{K}$, $\xi = \text{on}, \text{off}$) and $\overline{\boldsymbol{D}}_{\xi}$ ($\xi = \text{on}, \text{off}$) as

$$\boldsymbol{D}_{k,\xi} = \sum_{n=1}^{\infty} \boldsymbol{D}_{k,\xi}(n), \qquad \overline{\boldsymbol{D}}_{\xi} = \sum_{k \in \mathcal{K}} \boldsymbol{D}_{k,\xi},$$

respectively. We also define $\boldsymbol{D}$ as

$$\boldsymbol{D} = \begin{bmatrix} \overline{\boldsymbol{D}}_{\text{on}} & \boldsymbol{O} \\ \boldsymbol{O} & \overline{\boldsymbol{D}}_{\text{off}} \end{bmatrix}.$$

Note that the infinitesimal generator of the underlying Markov chain is given by $\boldsymbol{C} + \boldsymbol{D}$. Note also that $(\boldsymbol{C} + \boldsymbol{D})\boldsymbol{e} = \boldsymbol{0}$, where $\boldsymbol{e}$ denotes a column vector whose elements are all equal to one. We denote the stationary probability vector of the underlying Markov chain by $\boldsymbol{\pi}$. Because of the finite state space $\mathcal{M}$ and the irreducibility of the underlying Markov chain, $\boldsymbol{\pi}$ is uniquely determined so as to satisfy $\boldsymbol{\pi}(\boldsymbol{C} + \boldsymbol{D}) = \boldsymbol{0}$ and $\boldsymbol{\pi}\boldsymbol{e} = 1$. Let $\boldsymbol{\pi}_{\text{on}}$ (resp. $\boldsymbol{\pi}_{\text{off}}$) denote a $1 \times M_{\text{on}}$ (resp. $1 \times M_{\text{off}}$) vector representing the conditional stationary probability vector of the underlying Markov chain given that the server is on (resp. off). Note that $\boldsymbol{\pi}_{\text{on}}$ and $\boldsymbol{\pi}_{\text{off}}$ satisfy

$$\begin{aligned} \boldsymbol{\pi}_{\text{on}} \left[\boldsymbol{C}_{\text{on}} + \overline{\boldsymbol{D}}_{\text{on}} + \boldsymbol{E}_{\text{on,off}} \left[-\boldsymbol{C}_{\text{off}} - \overline{\boldsymbol{D}}_{\text{off}} \right]^{-1} \boldsymbol{E}_{\text{off,on}} \right] &= \boldsymbol{0}, \qquad \boldsymbol{\pi}_{\text{on}}\boldsymbol{e} = 1, \\ \boldsymbol{\pi}_{\text{off}} \left[\boldsymbol{C}_{\text{off}} + \overline{\boldsymbol{D}}_{\text{off}} + \boldsymbol{E}_{\text{off,on}} \left[-\boldsymbol{C}_{\text{on}} - \overline{\boldsymbol{D}}_{\text{on}} \right]^{-1} \boldsymbol{E}_{\text{on,off}} \right] &= \boldsymbol{0}, \qquad \boldsymbol{\pi}_{\text{off}}\boldsymbol{e} = 1, \end{aligned}$$

respectively. Let r_{on} and r_{off} denote fractions of time being in on- and off-periods, respectively. We then have

$$r_{\text{on}} = \frac{\overline{I}_{\text{on}}}{\overline{I}_{\text{on}} + \overline{I}_{\text{off}}}, \qquad r_{\text{off}} = \frac{\overline{I}_{\text{off}}}{\overline{I}_{\text{on}} + \overline{I}_{\text{off}}},$$

where $\overline{I}_{\text{on}}$ and $\overline{I}_{\text{off}}$ denote the mean lengths of on- and off-periods, respectively, and they are given by

$$\overline{I}_{\text{on}} = \frac{\boldsymbol{\pi}_{\text{off}}\boldsymbol{E}_{\text{off,on}}}{\boldsymbol{\pi}_{\text{off}}\boldsymbol{E}_{\text{off,on}}\boldsymbol{e}} \left[-\boldsymbol{C}_{\text{on}} - \overline{\boldsymbol{D}}_{\text{on}} \right]^{-1} \boldsymbol{e}, \qquad \overline{I}_{\text{off}} = \frac{\boldsymbol{\pi}_{\text{on}}\boldsymbol{E}_{\text{on,off}}}{\boldsymbol{\pi}_{\text{on}}\boldsymbol{E}_{\text{on,off}}\boldsymbol{e}} \left[-\boldsymbol{C}_{\text{off}} - \overline{\boldsymbol{D}}_{\text{off}} \right]^{-1} \boldsymbol{e}.$$

Note here that $\boldsymbol{\pi}$, $\boldsymbol{\pi}_{\text{on}}$ and $\boldsymbol{\pi}_{\text{off}}$ are related by

$$\boldsymbol{\pi} = (r_{\text{on}}\boldsymbol{\pi}_{\text{on}}, r_{\text{off}}\boldsymbol{\pi}_{\text{off}}).$$

We denote the mean arrival rate of class k ($k \in \mathcal{K}$) customers during on- (resp. off-) periods by $\lambda_{k,\text{on}}$ (resp. $\lambda_{k,\text{off}}$):

$$\lambda_{k,\xi} = \boldsymbol{\pi}_{\xi} \sum_{n=1}^{\infty} n\boldsymbol{D}_{k,\xi}(n)\boldsymbol{e}, \qquad \xi = \text{on, off}.$$

Let $\lambda_k = r_{\text{on}}\lambda_{k,\text{on}} + r_{\text{off}}\lambda_{k,\text{off}}$ ($k \in \mathcal{K}$) denote the mean arrival rate of class k customers. We define ρ as the offered load, i.e.,

$$\rho = r_{\text{on}} \sum_{k \in \mathcal{K}} \lambda_{k,\text{on}} h_{k,\text{on}} + r_{\text{off}} \sum_{k \in \mathcal{K}} \lambda_{k,\text{off}} h_{k,\text{off}}.$$

Further, we define ρ_{on} as the conditional utilization factor given that the server is on, which is given by

$$\rho_{\text{on}} = r_{\text{on}}^{-1}\rho.$$

In the remainder of this chapter, we assume that $\rho_{\text{on}} < 1$, and the system is in steady state.

3.3 Sojourn Time

This section considers sojourn time. Because sojourn time is closely related to the amount of unfinished work in the system, we first discuss the latter.

Let V and S denote generic random variables representing the amount of unfinished work and the state of the underlying Markov chain, respectively, in steady state. With these, we define $\boldsymbol{v}(x)$ as a $1 \times M$ vector whose jth element represents $\Pr[V \le x, S = j]$. Further we define $\boldsymbol{v}_\xi(x)$ ($\xi =$ on, off) as a $1 \times M_\xi$ vector whose jth element represents $\Pr[V \le x, S = j \mid S \in \mathcal{M}_\xi]$. Note here that $\boldsymbol{v}(x)$ is given in terms of $\boldsymbol{v}_{\rm on}(x)$ and $\boldsymbol{v}_{\rm off}(x)$:

$$\boldsymbol{v}(x) = (r_{\rm on}\boldsymbol{v}_{\rm on}(x), r_{\rm off}\boldsymbol{v}_{\rm off}(x)).$$

Thus, we derive the LST $\boldsymbol{v}^*_\xi(s)$ ($\xi =$ on, off) of $\boldsymbol{v}_\xi(x)$ below.

We first introduce some notations. Let $\boldsymbol{J}_{\rm off,on}(x)$ denote an $M_{\rm off} \times M_{\rm on}$ matrix whose (i,j)th element represents the probability that the amount of work arriving during an off-period is not greater than x and the underlying Markov chain is in state $j \in \mathcal{M}_{\rm on}$ at the beginning of the next on-period, given that the off-period starts in state $i \in \mathcal{M}_{\rm off}$. We define $\overline{\boldsymbol{D}}_\xi(x)$ ($\xi =$ on, off) as

$$\overline{\boldsymbol{D}}_\xi(x) = \sum_{k\in\mathcal{K}}\sum_{n=1}^{\infty} \boldsymbol{D}_{k,\xi}(n) H^{(n)}_{k,\xi}(x),$$

where $H^{(1)}_{k,\xi}(x) = H_{k,\xi}(x)$ and $H^{(n)}_{k,\xi}(x)$ ($n = 2, 3, \ldots$) denotes the n-fold convolution of $H_{k,\xi}(x)$ with itself. Note here that the LST $\boldsymbol{J}^*_{\rm off,on}(s)$ of $\boldsymbol{J}_{\rm off,on}(x)$ is given by

$$\boldsymbol{J}^*_{\rm off,on}(s) = \left[-\boldsymbol{C}_{\rm off} - \overline{\boldsymbol{D}}^*_{\rm off}(s)\right]^{-1} \boldsymbol{E}_{\rm off,on},$$

where $\overline{\boldsymbol{D}}^*_\xi(s)$ ($\xi =$ on, off) denotes the LST of $\overline{\boldsymbol{D}}^*_\xi(s)$. According to [Taki94a], we define an $M_{\rm on}\times M_{\rm on}$ matrix $\boldsymbol{Q}_{\rm on}$ as an infinitesimal generator of the irreducible Markov chain obtained by observing the underlying Markov chain only when the system is idle in on-periods. Note that $\boldsymbol{Q}_{\rm on}$ satisfies

$$\boldsymbol{Q}_{\rm on} = \boldsymbol{C}_{\rm on} + \int_0^\infty d\overline{\boldsymbol{D}}_{\rm on}(x) \exp(\boldsymbol{Q}_{\rm on}x) + \boldsymbol{E}_{\rm on,off}\int_0^\infty d\boldsymbol{J}_{\rm off,on}(x)\exp(\boldsymbol{Q}_{\rm on}x).$$

Because $\rho_{\rm on} < 1$, $\boldsymbol{Q}_{\rm on}$ is uniquely determined by the above equation [Taki94a]. Let $\boldsymbol{\kappa}_{\rm on}$ denote a probability vector satisfying $\boldsymbol{\kappa}_{\rm on}\boldsymbol{Q}_{\rm on} = \boldsymbol{0}$.

We now derive $\boldsymbol{v}^*_{\rm on}(s)$ in such a way as to obtain Proposition 1.8 in subsection 1.3.2. We consider the bivariate process $\{(V(t), S(t)); t \ge 0\}$, where $V(t)$ and $S(t)$ denote the virtual waiting time and the state of the underlying Markov chain, respectively, at time t. We then construct the censored process obtained by observing $\{(V(t), S(t)); t \ge 0\}$ only when $V(t) > 0$ and $S(t) \in \mathcal{M}_{\rm on}$. It is easy to see that the resulting bivariate process is equivalent to $\{(X_{\rm L}(t), J_{\rm L}(t)); t \ge 0\}$ with

$$\begin{aligned} \boldsymbol{A}_{\rm slope} &= \boldsymbol{C}_{\rm on}, \quad \boldsymbol{A}_{\rm jump}(x) = \overline{\boldsymbol{D}}_{\rm on}(x) + \boldsymbol{E}_{\rm on,off}\boldsymbol{J}_{\rm off,on}(x), \\ \boldsymbol{B}_{\rm jump}(x) &= (-\boldsymbol{C}_{\rm on})^{-1}\left[\overline{\boldsymbol{D}}_{\rm on}(x) + \boldsymbol{E}_{\rm on,off}\boldsymbol{J}_{\rm off,on}(x)\right]. \end{aligned}$$

It then follows that the LST $\boldsymbol{v}^*_{\rm on}(s)$ of $\boldsymbol{v}_{\rm on}(x)$ satisfies [Taki94a]

$$\boldsymbol{v}^*_{\rm on}(s)\left[s\boldsymbol{I} + \boldsymbol{C}_{\rm on} + \overline{\boldsymbol{D}}^*_{\rm on}(s) + \boldsymbol{E}_{\rm on,off}\left[-\boldsymbol{C}_{\rm off} - \overline{\boldsymbol{D}}^*_{\rm off}(s)\right]^{-1}\boldsymbol{E}_{\rm off,on}\right] = s(1-\rho_{\rm on})\boldsymbol{\kappa}_{\rm on}, \tag{3.1}$$

where $\boldsymbol{I}(m)$ denotes an $m \times m$ identity matrix. We suppress the size m when it is clear from the context. As for the LST $\boldsymbol{v}^*_{\rm off}(s)$ of $\boldsymbol{v}_{\rm off}(x)$, using the same approach as in [Taki97], we readily obtain

$$\boldsymbol{v}^*_{\rm off}(s) = \frac{\boldsymbol{v}^*_{\rm on}(s)\boldsymbol{E}_{\rm on,off}}{\boldsymbol{\pi}_{\rm on}\boldsymbol{E}_{\rm on,off}\boldsymbol{e}} \frac{[-\boldsymbol{C}_{\rm off} - \overline{\boldsymbol{D}}^*_{\rm off}(s)]^{-1}}{\overline{I}_{\rm off}}. \tag{3.2}$$

Next we analyze sojourn time. To do so, we first consider *completion time*, which is a time interval from the beginning of a service to its completion, including service interruptions. Let $T_c(u)$ denote a generic random variable representing the completion time of a service of u units. Note that the completion time $T_c(u)$ depends on the state of the underlying Markov chain at the beginning of the service, as well as the amount of the service. We denote the number of class k ($k \in \mathcal{K}$) customers arriving in interval $(0, t]$ by $L_k(t)$. Assuming that the service of u unit commences at time 0, we define $\boldsymbol{P}^{**}(\boldsymbol{z}, s \mid u)$ as an $M_{\rm on} \times M_{\rm on}$ matrix whose (i, j)th element represents

$$\mathrm{E}\left[\prod_{k\in\mathcal{K}} z_k^{L_k(T_c(u))} \exp(-sT_c(u))\, 1\{S(T_c(u)) = j \in \mathcal{M}_{\rm on}\} \mid S(0) = i \in \mathcal{M}_{\rm on}\right],$$

where $\boldsymbol{z}$ denotes a $1 \times K$ complex vector $(z_1, \ldots, z_K)$ and $1\{\chi\}$ denotes an indicator function of event χ. Further, we define $\boldsymbol{D}^*_{k,\xi}(z_k)$ ($k \in \mathcal{K}$, $\xi =$ on, off) as

$$\boldsymbol{D}^*_{k,\xi}(z_k) = \sum_{n=1}^{\infty} z_k^n \boldsymbol{D}_{k,\xi}(n). \tag{3.3}$$

Following an approach similar to [Taki97], we obtain $\boldsymbol{P}^{**}(\boldsymbol{z}, s \mid u) = \exp[\boldsymbol{K}(\boldsymbol{z}, s)u]$, where

$$\boldsymbol{K}(\boldsymbol{z}, s) = \boldsymbol{C}_{\rm on} + \sum_{k\in\mathcal{K}} \boldsymbol{D}^*_{k,\rm on}(z_k) - s\boldsymbol{I} + \boldsymbol{E}_{\rm on,off}\left[s\boldsymbol{I} - \boldsymbol{C}_{\rm off} - \sum_{k\in\mathcal{K}} \boldsymbol{D}^*_{k,\rm off}(z_k)\right]^{-1} \boldsymbol{E}_{\rm off,on}.$$

We are now ready to discuss sojourn time.

Let W_k ($k \in \mathcal{K}$) (resp. $W_{k,\xi}$ ($k \in \mathcal{K}$, $\xi =$ on, off)) denote a generic random variable representing the sojourn time of a class k (resp. k-ξ) customer. Also let $W_{k,\xi}(n; m)$ ($k \in \mathcal{K}$, $\xi =$ on, off, $n = 1, 2, \ldots$, $m = 1, \ldots, n$) denote a generic random variable representing the sojourn time of a randomly chosen class k-ξ customer who is a member of a batch of size n and the mth served customer among members in the same batch. For convenience, we assume that if $\lambda_{k,\xi} = 0$, $W_{k,\xi} = 0$, and if $\boldsymbol{D}_{k,\xi}(n) = \boldsymbol{O}$ for some n ($n \geq 1$), $W_{k,\xi}(n; m) = 0$ for all m ($m = 1, \ldots, n$). Further, let $w^*_k(s)$, $w^*_{k,\xi}(s)$ and $w^*_{k,\xi}(s \mid n; m)$ denote the LSTs of the distributions of W_k, $W_{k,\xi}$ and $W_{k,\xi}(n; m)$, respectively. Because a randomly chosen departing customer of class k ($k \in \mathcal{K}$) belongs to class k-ξ ($\xi =$ on, off) with probability $r_\xi \lambda_{k,\xi}/\lambda_k$, we obtain

$$w^*_k(s) = \frac{r_{\rm on}\lambda_{k,\rm on}}{\lambda_k} w^*_{k,\rm on}(s) + \frac{r_{\rm off}\lambda_{k,\rm off}}{\lambda_k} w^*_{k,\rm off}(s).$$

Note here that

$$w^*_{k,\xi}(s) \;=\; \sum_{n=1}^{\infty} \frac{n\boldsymbol{\pi}_\xi \boldsymbol{D}_{k,\xi}(n)\boldsymbol{e}}{\lambda_{k,\xi}} \cdot \frac{1}{n} \sum_{m=1}^{n} w^*_{k,\xi}(s \mid n; m), \qquad k \in \mathcal{K},\ \xi = \text{on, off}, \tag{3.4}$$

if $\lambda_{k,\xi} > 0$, and otherwise $w^*_{k,\xi}(s) = 1$. Thus, in what follows, we consider $w^*_{k,\xi}(s \mid n; m)$ ($k \in \mathcal{K}$, $\xi =$ on, off, $n = 1, 2, \ldots$, $m = 1, \ldots, n$).

Let $H_{k,\xi}(n;m)$ ($k \in \mathcal{K}$, $\xi = \text{on}, \text{off}$, $n = 1, 2, \ldots$, $m = 1, \ldots, n$) denote a generic random variable representing the service time of a randomly chosen class k-ξ customer who is a member of a batch of size n and the mth served customer among members of the same batch. Because $W_{k,\xi}(n;m) = W_{k,\xi}(n;1) + \sum_{l=2}^{m} T_c(H_{k,\xi}(n;l))$ for $\xi = \text{on}, \text{off}$, $n = 1, 2, \ldots$ and $m = 1, \ldots, n$, we obtain for $n = 1, 2, \ldots$ and $m = 1, \ldots, n$,

$$w^*_{k,\text{on}}(s \mid n;m) = \int_0^\infty \frac{d\boldsymbol{v}_{\text{on}}(x)\boldsymbol{D}_{k,\text{on}}(n)\exp[\boldsymbol{\Omega}(s)x]}{\boldsymbol{\pi}_{\text{on}}\boldsymbol{D}_{k,\text{on}}(n)\boldsymbol{e}} \left[\int_0^\infty dH_{k,\text{on}}(y)\exp[\boldsymbol{\Omega}(s)y]\right]^m \boldsymbol{e}, \tag{3.5}$$

$$w^*_{k,\text{off}}(s \mid n;m) = \int_0^\infty \frac{d\boldsymbol{v}_{\text{off}}(x)\boldsymbol{D}_{k,\text{off}}(n)\left[s\boldsymbol{I} - \boldsymbol{C}_{\text{off}} - \overline{\boldsymbol{D}}_{\text{off}}\right]^{-1}\boldsymbol{E}_{\text{off,on}}\exp[\boldsymbol{\Omega}(s)x]}{\boldsymbol{\pi}_{\text{off}}\boldsymbol{D}_{k,\text{off}}(n)\boldsymbol{e}} \cdot \left[\int_0^\infty dH_{k,\text{off}}(y)\exp[\boldsymbol{\Omega}(s)y]\right]^m \boldsymbol{e}, \tag{3.6}$$

respectively, where $\boldsymbol{\Omega}(s) = \boldsymbol{K}(1, \ldots, 1, s)$. Note here that the (i,j)th ($i, j \in \mathcal{M}_{\text{on}}$) element of $\exp[\boldsymbol{\Omega}(s)x]$ represents $\text{E}[\exp(-sT_c(x))1\{S(T_c(x)) = j\} \mid \text{a service of } x \text{ units starts at time 0 and } S(0) = i]$. Thus from (3.4) and (3.5), we obtain for $k \in \mathcal{K}$,

$$w^*_{k,\text{on}}(s) = \frac{1}{\lambda_{k,\text{on}}}\sum_{n=1}^{\infty}\int_0^\infty d\boldsymbol{v}_{\text{on}}(x)\boldsymbol{D}_{k,\text{on}}(n)\exp[\boldsymbol{\Omega}(s)x]\sum_{m=1}^{n}\left[\int_0^\infty dH_{k,\text{on}}(y)\exp[\boldsymbol{\Omega}(s)y]\right]^m \boldsymbol{e},$$

if $\lambda_{k,\text{on}} > 0$, and otherwise $w^*_{k,\text{on}}(s) = 1$. Similarly, from (3.4) and (3.6), we obtain for $k \in \mathcal{K}$,

$$w^*_{k,\text{off}}(s) = \frac{1}{\lambda_{k,\text{off}}}\sum_{n=1}^{\infty}\int_0^\infty d\boldsymbol{v}_{\text{off}}(x)\boldsymbol{D}_{k,\text{off}}(n)\left[s\boldsymbol{I} - \boldsymbol{C}_{\text{off}} - \overline{\boldsymbol{D}}_{\text{off}}\right]^{-1}\boldsymbol{E}_{\text{off,on}}\exp[\boldsymbol{\Omega}(s)x] \cdot \sum_{m=1}^{n}\left[\int_0^\infty dH_{k,\text{off}}(y)\exp[\boldsymbol{\Omega}(s)y]\right]^m \boldsymbol{e},$$

if $\lambda_{k,\text{off}} > 0$, and otherwise $w^*_{k,\text{off}}(s) = 1$.

3.4 Joint Queue Length Distribution

This section considers the joint queue length distribution. Let N_k ($k \in \mathcal{K}$) denote a generic random variable representing the number of class k customers in the stationary system. We then define $\boldsymbol{p}(\boldsymbol{n})$ ($\boldsymbol{n} \in \mathcal{Z}$) as a $1 \times M$ vector whose jth element represents $\Pr[N_1 = n_1, \ldots, N_K = n_K, S = j]$, where $\boldsymbol{n}$ denotes a $1 \times K$ nonnegative integer vector $(n_1, \ldots, n_K)$ and $\mathcal{Z} = \{(n_1, \ldots, n_K)\,;\, n_k = 0, 1, \ldots \text{ for all } k \in \mathcal{K}\}$. Further, let $N_\nu^{(\text{D}_k)}$ ($k, \nu \in \mathcal{K}$) and $S^{(\text{D}_k)}$ ($k \in \mathcal{K}$) denote generic random variables that represent the number of class ν customers in the system and the state of the underlying Markov chain, respectively, immediately after departures of class k customers in steady state. We then define $\boldsymbol{q}_k(\boldsymbol{n})$ ($k \in \mathcal{K}$, $\boldsymbol{n} \in \mathcal{Z}$) as a $1 \times M$ vector whose jth element represents $\Pr[N_1^{(\text{D}_k)} = n_1, \ldots, N_K^{(\text{D}_k)} = n_K, S^{(\text{D}_k)} = j]$.

Note here that the $\boldsymbol{p}(\boldsymbol{n})$ is given in terms of the $\boldsymbol{q}_k(\boldsymbol{n})$ (see Corollary 2.1 in chapter 2). Thus, we consider the $\boldsymbol{q}_k(\boldsymbol{n})$ in section 4.1. Further, in section 4.2, assuming discrete phase-type batch size distributions, we derive numerically feasible recursions for some quantities required in computing the $\boldsymbol{q}_k(\boldsymbol{n})$.

3.4.1 Joint queue length distribution immediately after departures

We define $\boldsymbol{q}_k^*(\boldsymbol{z})$ ($k \in \mathcal{K}$) as the vector generating function of the joint queue length distribution immediately after departures of class k customers.

$$\boldsymbol{q}_k^*(\boldsymbol{z}) = \sum_{\boldsymbol{n}\in\mathcal{Z}} z_1^{n_1}\cdots z_K^{n_K}\boldsymbol{q}_k(\boldsymbol{n}), \qquad |z_k| \le 1 \text{ for all } k \in \mathcal{K}.$$

Let $N_\nu^{(\mathrm{D}_{k,\xi})}$ and $S^{(\mathrm{D}_{k,\xi})}$ ($k,\nu \in \mathcal{K}$, $\xi =$ on, off) denote generic random variables that represent the number of class ν customers in the system and the state of the underlying Markov chain, respectively, immediately after departures of class k-ξ customers in steady state. We define $\boldsymbol{q}_{k,\xi}(\boldsymbol{n})$ ($k \in \mathcal{K}$, $\xi =$ on, off) as a $1\times M_{\mathrm{on}}$ vector whose jth element represents $\Pr[N_1^{(\mathrm{D}_{k,\xi})} = n_1, \ldots, N_K^{(\mathrm{D}_{k,\xi})} = n_K, S^{(\mathrm{D}_{k,\xi})} = j]$. We also define $\boldsymbol{q}_{k,\xi}^*(\boldsymbol{z})$ ($k \in \mathcal{K}$, $\xi =$ on, off) as

$$\boldsymbol{q}_{k,\xi}^*(\boldsymbol{z}) = \sum_{\boldsymbol{n}\in\mathcal{Z}} z_1^{n_1}\cdots z_K^{n_K}\boldsymbol{q}_{k,\xi}(\boldsymbol{n}).$$

We then have

$$\boldsymbol{q}_k^*(\boldsymbol{z}) = \left(\frac{r_{\mathrm{on}}\lambda_{k,\mathrm{on}}}{\lambda_k}\boldsymbol{q}_{k,\mathrm{on}}^*(\boldsymbol{z}) + \frac{r_{\mathrm{off}}\lambda_{k,\mathrm{off}}}{\lambda_k}\boldsymbol{q}_{k,\mathrm{off}}^*(\boldsymbol{z}), 0, \ldots, 0\right), \tag{3.7}$$

because all departures always occur in on-periods.

In the rest of this subsection, we derive $\boldsymbol{q}_{k,\mathrm{on}}^*(\boldsymbol{z})$ and $\boldsymbol{q}_{k,\mathrm{off}}^*(\boldsymbol{z})$. We call a randomly chosen class k-ξ ($k \in \mathcal{K}$, $\xi =$ on, off) customer who is a member of a batch of size n and the mth served customer among members in the same batch *the tagged customer*. Further we call the batch to which the tagged customer belongs *the tagged batch*. Let $N_\nu^{(\mathrm{D}_{k,\xi})}(n;m)$ and $S^{(\mathrm{D}_{k,\xi})}(n;m)$ ($k,\nu \in \mathcal{K}$, $\xi =$ on, off, $n = 1, 2\ldots$, $m = 1, \ldots, n$) denote generic random variables that represent the number of class ν customers in the system and the state of the underlying Markov chain, respectively, immediately after the departure of the tagged customer. We then define $\boldsymbol{q}_{k,\xi}^*(\boldsymbol{z} \mid n;m)$ ($k \in \mathcal{K}$, $\xi =$ on, off, $n = 1, 2\ldots$, $m = 1, \ldots, n$) as a $1 \times M_{\mathrm{on}}$ vector whose jth element represents $\mathrm{E}\Big[\prod_{\nu\in\mathcal{K}} z_\nu^{N_\nu^{(\mathrm{D}_{k,\xi})}(n;m)} 1\{S^{(\mathrm{D}_{k,\xi})}(n;m) = j\}\Big]$. It is easy to see that $\boldsymbol{q}_{k,\xi}^*(\boldsymbol{z})$ can be written in terms of $\boldsymbol{q}_{k,\xi}^*(\boldsymbol{z} \mid n;m)$:

$$\boldsymbol{q}_{k,\xi}^*(\boldsymbol{z}) = \sum_{n=1}^{\infty} \frac{n\boldsymbol{\pi}_\xi \boldsymbol{D}_{k,\xi}(n)\boldsymbol{e}}{\lambda_{k,\xi}} \frac{1}{n}\sum_{m=1}^{n} \boldsymbol{q}_{k,\xi}^*(\boldsymbol{z} \mid n;m), \qquad k \in \mathcal{K},\ \xi = \text{on, off}, \tag{3.8}$$

if $\lambda_{k,\xi} > 0$, and otherwise $\boldsymbol{q}_{k,\xi}^*(\boldsymbol{z}) = \boldsymbol{0}$.

Note here that customers who contribute to $\boldsymbol{q}_{k,\mathrm{on}}^*(\boldsymbol{z} \mid n;m)$ can be divided into three types: (i) customers arriving during the completion time of the total unfinished work immediately before the arrival of the tagged batch, (ii) customers arriving during an interval from the beginning of the first service of a member in the tagged batch to the completion of the service of the tagged customer, and (iii) $n-m$ customers who belong to the tagged batch and receive their services after the tagged customer. It then follows that for $k \in \mathcal{K}$, $n = 1, 2, \ldots$ and $m = 1, \ldots, n$,

$$\boldsymbol{q}_{k,\mathrm{on}}^*(\boldsymbol{z} \mid n;m) = z_k^{n-m}\int_0^\infty \frac{d\boldsymbol{v}_{\mathrm{on}}(x)\boldsymbol{D}_{k,\mathrm{on}}(n)\boldsymbol{N}^*(\boldsymbol{z} \mid x)}{\boldsymbol{\pi}_{\mathrm{on}}\boldsymbol{D}_{k,\mathrm{on}}(n)\boldsymbol{e}} \left[\int_0^\infty dH_{k,\mathrm{on}}(y)\boldsymbol{N}^*(\boldsymbol{z} \mid y)\right]^m, \tag{3.9}$$

where $\boldsymbol{N}^*(\boldsymbol{z} \mid x) = \boldsymbol{P}^{**}(\boldsymbol{z}, 0 \mid x)$, i.e.,

$$\begin{aligned}\boldsymbol{N}^*(\boldsymbol{z} \mid x) &= \exp\left[\left\{\boldsymbol{C}_{\rm on} + \sum_{k\in\mathcal{K}} \boldsymbol{D}^*_{k,{\rm on}}(z_k)\right.\right. \\ &\qquad \left.\left. + \boldsymbol{E}_{\rm on,off}\left(-\boldsymbol{C}_{\rm off} - \sum_{k\in\mathcal{K}} \boldsymbol{D}^*_{k,{\rm off}}(z_k)\right)^{-1} \boldsymbol{E}_{\rm off,on}\right\} x\right]. \qquad (3.10)\end{aligned}$$

Note that $\boldsymbol{N}^*(\boldsymbol{z} \mid x)$ denotes the matrix joint generating function for the numbers of arrivals in respective classes during the completion time of a service of x units. Similarly, we have for $k \in \mathcal{K}$, $n = 1, 2, \ldots$ and $m = 1, \ldots, n$,

$$\begin{aligned}&\boldsymbol{q}^*_{k,{\rm off}}(\boldsymbol{z} \mid n; m) \\ &\quad = z_k^{n-m} \int_0^\infty \frac{d\boldsymbol{v}_{\rm off}(x)\boldsymbol{D}_{k,{\rm off}}(n)}{\boldsymbol{\pi}_{\rm off}\boldsymbol{D}_{k,{\rm off}}(n)\boldsymbol{e}} \left[-\boldsymbol{C}_{\rm off} - \sum_{k\in\mathcal{K}} \boldsymbol{D}^*_{k,{\rm off}}(z_k)\right]^{-1} \boldsymbol{E}_{\rm off,on}\boldsymbol{N}^*(\boldsymbol{z} \mid x) \\ &\qquad \cdot \left[\int_0^\infty dH_{k,{\rm off}}(y)\boldsymbol{N}^*(\boldsymbol{z} \mid y)\right]^m. \qquad (3.11)\end{aligned}$$

Thus, from (3.8), (3.9) and (3.11), we obtain the following theorem.

Theorem 3.1 *$\boldsymbol{q}^*_{k,{\rm on}}(\boldsymbol{z})$ ($k \in \mathcal{K}$) is given by*

$$\boldsymbol{q}^*_{k,{\rm on}}(\boldsymbol{z}) = \frac{1}{\lambda_{k,{\rm on}}} \sum_{m=1}^\infty \sum_{l=0}^\infty z_k^l \int_0^\infty d\boldsymbol{v}_{\rm on}(x)\boldsymbol{D}_{k,{\rm on}}(m+l)\boldsymbol{N}^*(\boldsymbol{z} \mid x) \left[\int_0^\infty dH_{k,{\rm on}}(y)\boldsymbol{N}^*(\boldsymbol{z} \mid y)\right]^m,$$

if $\lambda_{k,{\rm on}} > 0$, and otherwise $\boldsymbol{q}^*_{k,{\rm on}}(\boldsymbol{z}) = \boldsymbol{0}$. On the other hand, $\boldsymbol{q}^*_{k,{\rm off}}(\boldsymbol{z})$ ($k \in \mathcal{K}$) is given by

$$\begin{aligned}\boldsymbol{q}^*_{k,{\rm off}}(\boldsymbol{z}) &= \frac{1}{\lambda_{k,{\rm off}}} \sum_{m=1}^\infty \sum_{l=0}^\infty z_k^l \int_0^\infty d\boldsymbol{v}_{\rm off}(x)\boldsymbol{D}_{k,{\rm off}}(m+l) \left[-\boldsymbol{C}_{\rm off} - \sum_{k\in\mathcal{K}} \boldsymbol{D}^*_{k,{\rm off}}(z_k)\right]^{-1} \boldsymbol{E}_{\rm off,on} \\ &\qquad \cdot \boldsymbol{N}^*(\boldsymbol{z} \mid x) \left[\int_0^\infty dH_{k,{\rm off}}(y)\boldsymbol{N}^*(\boldsymbol{z} \mid y)\right]^m,\end{aligned}$$

if $\lambda_{k,{\rm off}} > 0$, and otherwise $\boldsymbol{q}^*_{k,{\rm off}}(\boldsymbol{z}) = \boldsymbol{0}$.

3.4.2 Recursions for discrete phase-type batch sizes

In this subsection, we develop recursive formulas to compute the joint queue length distribution $\boldsymbol{q}_k(\boldsymbol{n})$ immediately after departures of class k customers under the following assumption.

Assumption 3.1 *Batch sizes of class k-ξ ($k \in \mathcal{K}$, ξ = on, off) are i.i.d. according to a discrete phase-type distribution with representation $(\boldsymbol{\alpha}_{k,\xi}, \boldsymbol{P}_{k,\xi})$, where $\boldsymbol{\alpha}_{k,\xi}$ denotes a $1 \times M_{k,\xi}$ probability vector and $\boldsymbol{P}_{k,\xi}$ denotes an $M_{k,\xi} \times M_{k,\xi}$ substochastic matrix.*

Under Assumption 3.1, $\boldsymbol{D}_{k,\xi}(n)$ ($k \in \mathcal{K}$, ξ = on, off) is given by

$$\boldsymbol{D}_{k,\xi}(n) = g_{k,\xi}(n)\boldsymbol{D}_{k,\xi}, \qquad n = 1, 2, \ldots,$$

where $g_{k,\xi}(n)$ denotes the probability mass function of the batch size of class k-ξ:

$$g_{k,\xi}(n) = \boldsymbol{\alpha}_{k,\xi}\boldsymbol{P}_{k,\xi}^{n-1}(\boldsymbol{I}-\boldsymbol{P}_{k,\xi})\boldsymbol{e}, \qquad n = 1, 2, \ldots.$$

Thus, Theorem 3.1 is reduced to:

Corollary 3.1 *Under Assumption 3.1, $\boldsymbol{q}^*_{k,\mathrm{on}}(\boldsymbol{z})$ $(k \in \mathcal{K})$ is given by*

$$\begin{aligned}\boldsymbol{q}^*_{k,\mathrm{on}}(\boldsymbol{z}) = \frac{1}{\lambda_{k,\mathrm{on}}}\int_0^\infty d\boldsymbol{v}_{\mathrm{on}}(x)\boldsymbol{D}_{k,\mathrm{on}}\boldsymbol{N}^*(\boldsymbol{z}\mid x)\\ \cdot\left(\boldsymbol{\alpha}_{k,\mathrm{on}}\otimes\int_0^\infty dH_{k,\mathrm{on}}(y)\boldsymbol{N}^*(\boldsymbol{z}\mid y)\right)\left[\boldsymbol{I}-\boldsymbol{P}_{k,\mathrm{on}}\otimes\int_0^\infty dH_{k,\mathrm{on}}(y)\boldsymbol{N}^*(\boldsymbol{z}\mid y)\right]^{-1}\\ \cdot\left[\left\{(\boldsymbol{I}-z_k\boldsymbol{P}_{k,\mathrm{on}})^{-1}(\boldsymbol{I}-\boldsymbol{P}_{k,\mathrm{on}})\boldsymbol{e}\right\}\otimes\boldsymbol{I}(M_{\mathrm{on}})\right], \qquad (3.12)\end{aligned}$$

if $\lambda_{k,\mathrm{on}} > 0$, and otherwise $\boldsymbol{q}^*_{k,\mathrm{on}}(\boldsymbol{z}) = \boldsymbol{0}$. Similarly, $\boldsymbol{q}^*_{k,\mathrm{off}}(\boldsymbol{z})$ $(k \in \mathcal{K})$ is given by

$$\begin{aligned}\boldsymbol{q}^*_{k,\mathrm{off}}(\boldsymbol{z}) = \frac{1}{\lambda_{k,\mathrm{off}}}\int_0^\infty d\boldsymbol{v}_{\mathrm{off}}(x)\boldsymbol{D}_{k,\mathrm{off}}\left[-\boldsymbol{C}_{\mathrm{off}}-\sum_{k\in\mathcal{K}}\boldsymbol{D}^*_{k,\mathrm{off}}(z_k)\right]^{-1}\boldsymbol{E}_{\mathrm{off,on}}\boldsymbol{N}^*(\boldsymbol{z}\mid x)\\ \cdot\left(\boldsymbol{\alpha}_{k,\mathrm{off}}\otimes\int_0^\infty dH_{k,\mathrm{off}}(y)\boldsymbol{N}^*(\boldsymbol{z}\mid y)\right)\left[\boldsymbol{I}-\boldsymbol{P}_{k,\mathrm{off}}\otimes\int_0^\infty dH_{k,\mathrm{off}}(y)\boldsymbol{N}^*(\boldsymbol{z}\mid y)\right]^{-1}\\ \cdot\left[\left\{(\boldsymbol{I}-z_k\boldsymbol{P}_{k,\mathrm{off}})^{-1}(\boldsymbol{I}-\boldsymbol{P}_{k,\mathrm{off}})\boldsymbol{e}\right\}\otimes\boldsymbol{I}(M_{\mathrm{on}})\right], \qquad (3.13)\end{aligned}$$

if $\lambda_{k,\mathrm{off}} > 0$, and otherwise $\boldsymbol{q}^*_{k,\mathrm{off}}(\boldsymbol{z}) = \boldsymbol{0}$.

This corollary can be obtained in the same way as Lemma 2.1 in chapter 2, and therefore we omit the proof.

We define $\boldsymbol{v}_{k,\mathrm{on}}(\boldsymbol{n})$ and $\boldsymbol{v}_{k,\mathrm{off}}(\boldsymbol{n})$ $(k \in \mathcal{K}, \boldsymbol{n} \in \mathcal{Z})$ as $1 \times M_{\mathrm{on}}$ vectors satisfying

$$\begin{aligned}\sum_{\boldsymbol{n}\in\mathcal{Z}} z_1^{n_1}\cdots z_K^{n_K}\boldsymbol{v}_{k,\mathrm{on}}(\boldsymbol{n}) &= \int_0^\infty d\boldsymbol{v}_{\mathrm{on}}(x)\boldsymbol{D}_{k,\mathrm{on}}\boldsymbol{N}^*(\boldsymbol{z}\mid x), \qquad (3.14)\\ \sum_{\boldsymbol{n}\in\mathcal{Z}} z_1^{n_1}\cdots z_K^{n_K}\boldsymbol{v}_{k,\mathrm{off}}(\boldsymbol{n}) &= \int_0^\infty d\boldsymbol{v}_{\mathrm{off}}(x)\boldsymbol{D}_{k,\mathrm{off}}\left[-\boldsymbol{C}_{\mathrm{off}}-\sum_{k\in\mathcal{K}}\boldsymbol{D}^*_{k,\mathrm{off}}(z_k)\right]^{-1}\boldsymbol{E}_{\mathrm{off,on}}\boldsymbol{N}^*(\boldsymbol{z}\mid x), \qquad (3.15)\end{aligned}$$

respectively. We also define $\boldsymbol{A}_{k,\xi}(\boldsymbol{n})$ and $\boldsymbol{\Gamma}_{k,\xi}(\boldsymbol{n})$ $(k \in \mathcal{K},\ \xi = \mathrm{on}, \mathrm{off},\ \boldsymbol{n} \in \mathcal{Z})$ as $M_{\mathrm{on}} \times M_{\mathrm{on}}$ and $M_{k,\xi}M_{\mathrm{on}} \times M_{k,\xi}M_{\mathrm{on}}$ matrices satisfying

$$\begin{aligned}\sum_{\boldsymbol{n}\in\mathcal{Z}} z_1^{n_1}\cdots z_K^{n_K}\boldsymbol{A}_{k,\xi}(\boldsymbol{n}) &= \int_0^\infty dH_{k,\xi}(y)\boldsymbol{N}^*(\boldsymbol{z}\mid y), \qquad (3.16)\\ \sum_{\boldsymbol{n}\in\mathcal{Z}} z_1^{n_1}\cdots z_K^{n_K}\boldsymbol{\Gamma}_{k,\xi}(\boldsymbol{n}) &= \left[\boldsymbol{I}-\boldsymbol{P}_{k,\xi}\otimes\int_0^\infty dH_{k,\xi}(y)\boldsymbol{N}^*(\boldsymbol{z}\mid y)\right]^{-1},\end{aligned}$$

respectively. Then $\boldsymbol{q}^*_{k,\xi}(\boldsymbol{z})$ $(k \in \mathcal{K}, \xi = \mathrm{on}, \mathrm{off})$ in (3.12) and (3.13) are rewritten to be

$$\begin{aligned}\boldsymbol{q}^*_{k,\xi}(\boldsymbol{z}) = \frac{1}{\lambda_{k,\xi}}\sum_{\boldsymbol{n}\in\mathcal{Z}} z_1^{n_1}\cdots z_K^{n_K}\sum_{m=0}^{n_k}\sum_{\substack{\boldsymbol{n}_1,\boldsymbol{n}_2,\boldsymbol{n}_3\in\mathcal{Z}\\ \boldsymbol{n}_1+\boldsymbol{n}_2+\boldsymbol{n}_3\\ =\boldsymbol{n}-m\boldsymbol{e}_k}}\boldsymbol{v}_{k,\xi}(\boldsymbol{n}_1)[\boldsymbol{\alpha}_{k,\xi}\otimes\boldsymbol{A}_{k,\xi}(\boldsymbol{n}_2)]\boldsymbol{\Gamma}_{k,\xi}(\boldsymbol{n}_3)\\ \cdot\left[\left\{\boldsymbol{P}^m_{k,\xi}(\boldsymbol{I}-\boldsymbol{P}_{k,\xi})\boldsymbol{e}\right\}\otimes\boldsymbol{I}(M_{\mathrm{on}})\right], \qquad (3.17)\end{aligned}$$

if $\lambda_{k,\xi} > 0$, and otherwise $\boldsymbol{q}^*_{k,\xi}(\boldsymbol{z}) = \boldsymbol{0}$. Comparing coefficient vectors of $z_1 \cdots z_K$ on both sides of (3.7) and (3.17), respectively, we obtain the following result.

Theorem 3.2 *Under Assumption 3.1, the $\boldsymbol{q}_k(\boldsymbol{n})$ is given by*

$$\boldsymbol{q}_k(\boldsymbol{n}) = \left(\frac{r_{\rm on}\lambda_{k,\rm on}}{\lambda_k}\boldsymbol{q}_{k,\rm on}(\boldsymbol{n}) + \frac{r_{\rm off}\lambda_{k,\rm off}}{\lambda_k}\boldsymbol{q}_{k,\rm off}(\boldsymbol{n}), 0, \ldots, 0 \right), \qquad k \in \mathcal{K},\ \boldsymbol{n} \in \mathcal{Z},$$

where the $\boldsymbol{q}_{k,\xi}(\boldsymbol{n})$ ($k \in \mathcal{K}$, ξ = on, off, $\boldsymbol{n} \in \mathcal{Z}$) is given by

$$\begin{aligned} \boldsymbol{q}_{k,\xi}(\boldsymbol{n}) &= \frac{1}{\lambda_{k,\xi}} \sum_{m=0}^{n_k} \sum_{\substack{\boldsymbol{n}_1,\boldsymbol{n}_2,\boldsymbol{n}_3 \in \mathcal{Z} \\ \boldsymbol{n}_1+\boldsymbol{n}_2+\boldsymbol{n}_3 \\ = \boldsymbol{n}-m\boldsymbol{e}_k}} \boldsymbol{v}_{k,\xi}(\boldsymbol{n}_1)[\boldsymbol{\alpha}_{k,\xi} \otimes \boldsymbol{A}_{k,\xi}(\boldsymbol{n}_2)]\boldsymbol{\Gamma}_{k,\xi}(\boldsymbol{n}_3) \\ &\qquad \cdot \left[\left\{ \boldsymbol{P}^m_{k,\xi}(\boldsymbol{I} - \boldsymbol{P}_{k,\xi})\boldsymbol{e} \right\} \otimes \boldsymbol{I}(M_{\rm on}) \right], \end{aligned}$$

if $\lambda_{k,\xi} > 0$, and otherwise $\boldsymbol{q}_{k,\xi}(\boldsymbol{n}) = \boldsymbol{0}$ for all $\boldsymbol{n} \in \mathcal{Z}$.

Theorem 3.2 implies that the computation of the $\boldsymbol{q}_k(\boldsymbol{n})$ is reduced to those of the $\boldsymbol{\Gamma}_{k,\xi}(\boldsymbol{n})$, $\boldsymbol{A}_{k,\xi}(\boldsymbol{n})$ and $\boldsymbol{v}_{k,\xi}(\boldsymbol{n})$ (ξ = on, off). Note here that the $\boldsymbol{\Gamma}_{k,\xi}(\boldsymbol{n})$ is given in terms of the $\boldsymbol{A}_{k,\xi}(\boldsymbol{n})$.

Lemma 3.1 *The $\boldsymbol{\Gamma}_{k,\xi}(\boldsymbol{n})$ ($k \in \mathcal{K}$, ξ = on, off, $\boldsymbol{n} \in \mathcal{Z}$) is determined by the following recursion:*

$$\begin{aligned} \boldsymbol{\Gamma}_{k,\xi}(\boldsymbol{0}) &= [\boldsymbol{I} - \boldsymbol{P}_{k,\xi} \otimes \boldsymbol{A}_{k,\xi}(\boldsymbol{0})]^{-1}, \\ \boldsymbol{\Gamma}_{k,\xi}(\boldsymbol{n}) &= \sum_{\substack{\boldsymbol{0} \le \boldsymbol{l} \le \boldsymbol{n} \\ \boldsymbol{l} \ne \boldsymbol{0}}} \boldsymbol{\Gamma}_{k,\xi}(\boldsymbol{n}-\boldsymbol{l}) \left[\boldsymbol{P}_{k,\xi} \otimes \boldsymbol{A}_{k,\xi}(\boldsymbol{l}) \right] \boldsymbol{\Gamma}_{k,\xi}(\boldsymbol{0}), \qquad \boldsymbol{n} \in \mathcal{Z}^+. \end{aligned}$$

Proof. We can prove this lemma in the same way as Lemma 2.3. ∎

The rest of this subsection therefore discusses the computations of the $\boldsymbol{A}_{k,\xi}(\boldsymbol{n})$ and the $\boldsymbol{v}_{k,\xi}(\boldsymbol{n})$. We first consider the $\boldsymbol{A}_{k,\xi}(\boldsymbol{n})$. Let $\overline{\boldsymbol{F}}_m(\boldsymbol{n})$ ($m = 0, 1, \ldots,\ \boldsymbol{n} \in \mathcal{Z}$) denote an $M_{\rm on} \times M_{\rm on}$ matrix that satisfies

$$\begin{aligned} \sum_{\boldsymbol{n} \in \mathcal{Z}} z_1^{n_1} \cdots z_K^{n_K} \overline{\boldsymbol{F}}_m(\boldsymbol{n}) &= \left[\boldsymbol{I} + \theta_{\rm on}^{-1} \left\{ \boldsymbol{C}_{\rm on} + \sum_{k \in \mathcal{K}} \boldsymbol{D}^*_{k,\rm on}(z_k) \right. \right. \\ &\qquad \left. \left. + \boldsymbol{E}_{\rm on,off} \left(-\boldsymbol{C}_{\rm off} - \sum_{k \in \mathcal{K}} \boldsymbol{D}^*_{k,\rm off}(z_k) \right)^{-1} \boldsymbol{E}_{\rm off,on} \right\} \right]^m, \end{aligned} \tag{3.18}$$

where $\theta_{\rm on}$ denotes the maximum absolute value of diagonal elements of $\boldsymbol{C}_{\rm on}$.

Lemma 3.2 *The $\overline{\boldsymbol{F}}_m(\boldsymbol{n})$ is recursively determined by*

$$\overline{\boldsymbol{F}}_0(\boldsymbol{n}) = \begin{cases} \boldsymbol{I}, & \text{if } \boldsymbol{n} = \boldsymbol{0}, \\ \boldsymbol{O}, & \text{otherwise}, \end{cases} \tag{3.19}$$

and for $m = 1, 2, \ldots,$

$$\begin{aligned} \overline{\boldsymbol{F}}_m(\boldsymbol{n}) &= \overline{\boldsymbol{F}}_{m-1}(\boldsymbol{n})(\boldsymbol{I} + \theta_{\rm on}^{-1}\boldsymbol{C}_{\rm on}) + \theta_{\rm on}^{-1} \sum_{k \in \mathcal{K}} \sum_{l_k=1}^{n_k} \overline{\boldsymbol{F}}_{m-1}(\boldsymbol{n} - l_k\boldsymbol{e}_k)\boldsymbol{D}_{k,\rm on}(l_k) \\ &\qquad + \theta_{\rm on}^{-1} \left[\sum_{\boldsymbol{0} \le \boldsymbol{l} \le \boldsymbol{n}} \overline{\boldsymbol{F}}_{m-1}(\boldsymbol{n}-\boldsymbol{l})\boldsymbol{E}_{\rm on,off}\boldsymbol{N}_{\rm off}(\boldsymbol{l}) \right] \boldsymbol{E}_{\rm off,on}, \qquad \boldsymbol{n} \in \mathcal{Z}, \end{aligned} \tag{3.20}$$

where $M_{\rm off} \times M_{\rm off}$ matrices $\boldsymbol{N}_{\rm off}(\boldsymbol{n})$'s are determined by the following recursion:

$$\boldsymbol{N}_{\rm off}(\boldsymbol{0}) = (-\boldsymbol{C}_{\rm off})^{-1}, \tag{3.21}$$

$$\boldsymbol{N}_{\rm off}(\boldsymbol{n}) = \left[\sum_{k\in\mathcal{K}}\sum_{l_k=1}^{n_k} \boldsymbol{N}_{\rm off}(\boldsymbol{n}-l_k\boldsymbol{e}_k)\boldsymbol{D}_{k,\rm off}(l_k)\right]\boldsymbol{N}_{\rm off}(\boldsymbol{0}), \qquad \boldsymbol{n}\in\mathcal{Z}^+. \tag{3.22}$$

The proof of Lemma 3.2 is given in Appendix D.

The $\boldsymbol{A}_{k,\xi}(\boldsymbol{n})$ is given in terms of the $\overline{\boldsymbol{F}}_m(\boldsymbol{n})$ in the following way. It follows from (3.10), (3.16) and (3.18) that

$$\begin{aligned}
&\sum_{\boldsymbol{n}\in\mathcal{Z}} z_1^{n_1}\cdots z_K^{n_K}\boldsymbol{A}_{k,\xi}(\boldsymbol{n}) \\
&\quad= \sum_{m=0}^{\infty}\int_0^{\infty} dH_{k,\xi}(y)\frac{(\theta_{\rm on}y)^m}{m!}e^{-\theta_{\rm on}y}\left[\boldsymbol{I}+\theta_{\rm on}^{-1}\left\{\boldsymbol{C}_{\rm on}+\sum_{k\in\mathcal{K}}\boldsymbol{D}^*_{k,\rm on}(z_k)\right.\right. \\
&\qquad\qquad\left.\left.+\boldsymbol{E}_{\rm on,off}\left(-\boldsymbol{C}_{\rm off}-\sum_{k\in\mathcal{K}}\boldsymbol{D}^*_{k,\rm off}(z_k)\right)^{-1}\boldsymbol{E}_{\rm off,on}\right\}\right]^m \\
&\quad= \sum_{\boldsymbol{n}\in\mathcal{Z}} z_1^{n_1}\cdots z_K^{n_K}\sum_{m=0}^{\infty}\gamma^{(m)}_{k,\xi}(\theta_{\rm on})\overline{\boldsymbol{F}}_m(\boldsymbol{n}),
\end{aligned} \tag{3.23}$$

where

$$\gamma^{(m)}_{k,\xi}(\theta_{\rm on}) = \int_0^{\infty} dH_{k,\xi}(y)\frac{(\theta_{\rm on}y)^m}{m!}e^{-\theta_{\rm on}y}, \qquad k\in\mathcal{K},\ \xi=\text{on, off},\ m=0,1,\ldots.$$

Comparing coefficient vectors of $z_1^{n_1}\cdots z_K^{n_K}$ on both sides in (3.23), we obtain the following theorem.

Theorem 3.3 *The $\boldsymbol{A}_{k,\xi}(\boldsymbol{n})$ is given by*

$$\boldsymbol{A}_{k,\xi}(\boldsymbol{n}) = \sum_{m=0}^{\infty}\gamma^{(m)}_{k,\xi}(\theta_{\rm on})\overline{\boldsymbol{F}}_m(\boldsymbol{n}), \qquad k\in\mathcal{K},\ \xi=\textit{on, off},\ \boldsymbol{n}\in\mathcal{Z},$$

where the $\overline{\boldsymbol{F}}_m(\boldsymbol{n})$ is given in Lemma 3.2.

Next we consider the $\boldsymbol{v}_{k,\rm on}(\boldsymbol{n})$ in (3.14) and the $\boldsymbol{v}_{k,\rm off}(\boldsymbol{n})$ in (3.15). Expanding $\boldsymbol{N}^*(\boldsymbol{z}\mid x)$ in (3.14) and (3.15), and comparing coefficient vectors of $z_1^{n_1}\cdots z_K^{n_K}$ on both sizes of each equation, we obtain the following theorem.

Theorem 3.4 *The $\boldsymbol{v}_{k,\rm on}(\boldsymbol{n})$ $(k\in\mathcal{K})$ and the $\boldsymbol{v}_{k,\rm off}(\boldsymbol{n})$ $(k\in\mathcal{K})$ are given by*

$$\boldsymbol{v}_{k,\rm on}(\boldsymbol{n}) = \sum_{m=0}^{\infty}\boldsymbol{v}^{(m)}_{\rm on}(\theta_{\rm on})\boldsymbol{D}_{k,\rm on}\overline{\boldsymbol{F}}_m(\boldsymbol{n}), \qquad \boldsymbol{n}\in\mathcal{Z},$$

$$\boldsymbol{v}_{k,\rm off}(\boldsymbol{n}) = \sum_{m=0}^{\infty}\boldsymbol{v}^{(m)}_{\rm off}(\theta_{\rm on})\boldsymbol{D}_{k,\rm off}\sum_{\boldsymbol{0}\le\boldsymbol{l}\le\boldsymbol{n}}\boldsymbol{N}_{\rm off}(\boldsymbol{n}-\boldsymbol{l})\boldsymbol{E}_{\rm off,on}\overline{\boldsymbol{F}}_m(\boldsymbol{l}), \qquad \boldsymbol{n}\in\mathcal{Z},$$

respectively, where the $\overline{\boldsymbol{F}}_m(\boldsymbol{n})$ and the $\boldsymbol{N}_{\rm off}(\boldsymbol{n})$ are given in Lemma 3.2, and

$$\boldsymbol{v}^{(m)}_{\xi}(\theta_{\rm on}) = \int_0^{\infty} d\boldsymbol{v}_{\xi}(x)\frac{(\theta_{\rm on}x)^m}{m!}e^{-\theta_{\rm on}x}, \qquad \xi=\text{on, off},\ m=0,1,\ldots.$$

Thus the $\boldsymbol{v}_{k,\xi}(\boldsymbol{n})$ ($\xi = \text{on}, \text{off}$) is given in terms of the $\boldsymbol{v}_{\xi}^{(m)}(\theta_{\text{on}})$ whose computation has already been studied in [Taki01b]. In what follows, we summarize the result. The perceptive reader will have noticed that the computational procedure for $\boldsymbol{v}_{\text{on}}^{(m)}(\theta_{\text{on}})$ is very similar to that for $\boldsymbol{v}^{(m)}(\theta)$ in (2.28). Note first that

$$\boldsymbol{v}_{\xi}^{*}(\theta_{\text{on}} - \theta_{\text{on}} z) = \sum_{m=0}^{\infty} z^m \boldsymbol{v}_{\xi}^{(m)}(\theta_{\text{on}}), \qquad \xi = \text{on}, \text{off}. \tag{3.24}$$

Substituting $\theta_{\text{on}} - \theta_{\text{on}} z$ for s in (3.1) and using (3.24), we have

$$\begin{aligned} \sum_{m=0}^{\infty} z^m \boldsymbol{v}_{\text{on}}^{(m)}(\theta_{\text{on}}) \Bigg[& (\theta_{\text{on}} - \theta_{\text{on}} z)\boldsymbol{I} + \boldsymbol{C}_{\text{on}} + \sum_{m=0}^{\infty} z^m \boldsymbol{D}_{\text{on}}^{(m)}(\theta_{\text{on}}) \\ & + \boldsymbol{E}_{\text{on,off}} \sum_{m=0}^{\infty} z^m \boldsymbol{J}_{\text{off}}^{(m)}(\theta_{\text{on}}) \boldsymbol{E}_{\text{off,on}} \Bigg] = (\theta_{\text{on}} - \theta_{\text{on}} z)(1 - \rho_{\text{on}}) \boldsymbol{\kappa}_{\text{on}}, \end{aligned} \tag{3.25}$$

where $\boldsymbol{D}_{\xi}^{(m)}(\theta_{\text{on}})$ ($\xi = \text{on}, \text{off}$) and $\boldsymbol{J}_{\text{off}}^{(m)}(\theta_{\text{on}})$ are matrices satisfying

$$\begin{aligned} \sum_{m=0}^{\infty} z^m \boldsymbol{D}_{\xi}^{(m)}(\theta_{\text{on}}) &= \overline{\boldsymbol{D}}_{\xi}^{*}(\theta_{\text{on}} - \theta_{\text{on}} z), \\ \sum_{m=0}^{\infty} z^m \boldsymbol{J}_{\text{off}}^{(m)}(\theta_{\text{on}}) &= \left[-\boldsymbol{C}_{\text{off}} - \overline{\boldsymbol{D}}_{\text{off}}^{*}(\theta_{\text{on}} - \theta_{\text{on}} z) \right]^{-1}, \end{aligned} \tag{3.26}$$

respectively. The computation of the $\boldsymbol{D}_{\xi}^{(m)}(\theta_{\text{on}})$ ($\xi = \text{on}, \text{off}$) has already been studied (see Lemma 2.5 in chapter 2).

Lemma 3.3 *Under Assumption 3.1, the $\boldsymbol{D}_{\xi}^{(m)}(\theta_{\text{on}})$ (ξ = on, off) is given by*

$$\boldsymbol{D}_{\xi}^{(m)}(\theta_{\text{on}}) = \sum_{k \in \mathcal{K}} \boldsymbol{d}_{k,\xi}^{(m)}(\theta_{\text{on}}) \boldsymbol{e} \boldsymbol{D}_{k,\xi}, \qquad m = 0, 1, \ldots,$$

where the $\boldsymbol{d}_{k,\xi}^{(m)}(\theta_{\text{on}})$ ($k \in \mathcal{K}$, ξ = on, off) is given by the following recursion:

$$\begin{aligned} \boldsymbol{d}_{k,\xi}^{(0)}(\theta_{\text{on}}) &= \gamma_{k,\xi}^{(0)}(\theta_{\text{on}}) \boldsymbol{\alpha}_{k,\xi} (\boldsymbol{I} - \boldsymbol{P}_{k,\xi}) \left[\boldsymbol{I} - \gamma_{k,\xi}^{(0)}(\theta_{\text{on}}) \boldsymbol{P}_{k,\xi} \right]^{-1}, \\ \boldsymbol{d}_{k,\xi}^{(m)}(\theta_{\text{on}}) &= \frac{\gamma_{k,\xi}^{(m)}(\theta_{\text{on}})}{\gamma_{k,\xi}^{(0)}(\theta_{\text{on}})} \boldsymbol{d}_{k,\xi}^{(0)}(\theta_{\text{on}}) + \left[\sum_{l=1}^{m} \gamma_{k,\xi}^{(l)}(\theta_{\text{on}}) \boldsymbol{d}_{k,\xi}^{(m-l)}(\theta_{\text{on}}) \right] \\ & \quad \cdot \boldsymbol{P}_{k,\xi} \left[\boldsymbol{I} - \gamma_{k,\xi}^{(0)}(\theta_{\text{on}}) \boldsymbol{P}_{k,\xi} \right]^{-1}, \qquad m = 1, 2, \ldots. \end{aligned}$$

Proof. We can prove this lemma in the same way as Lemma 2.5. ∎

The recursion for the $\boldsymbol{J}_{\text{off}}^{(m)}(\theta_{\text{on}})$ can be readily obtained from

$$\sum_{m=0}^{\infty} z^m \boldsymbol{J}_{\text{off}}^{(m)}(\theta_{\text{on}}) \left[-\boldsymbol{C}_{\text{off}} - \sum_{m=0}^{\infty} z^m \boldsymbol{D}_{\text{off}}^{(m)}(\theta_{\text{on}}) \right] = \boldsymbol{I}.$$

Lemma 3.4 *The $\boldsymbol{J}_{\rm off}^{(m)}(\theta_{\rm on})$ is recursively determined by the following recursion:*

$$\begin{aligned}
\boldsymbol{J}_{\rm off}^{(0)}(\theta_{\rm on}) &= \left[-\boldsymbol{C}_{\rm off} - \boldsymbol{D}_{\rm off}^{(0)}(\theta_{\rm on})\right]^{-1}, \\
\boldsymbol{J}_{\rm off}^{(m)}(\theta_{\rm on}) &= \left[\sum_{l=0}^{m-1} \boldsymbol{J}_{\rm off}^{(l)}(\theta_{\rm on})\boldsymbol{D}_{\rm off}^{(m-l)}(\theta_{\rm on})\right] \boldsymbol{J}_{\rm off}^{(0)}(\theta_{\rm on}), \qquad m = 1, 2, \ldots.
\end{aligned}$$

The $\boldsymbol{v}_{\rm on}^{(m)}(\theta_{\rm on})$ is computed as follows. Comparing the coefficient vectors of z^m $(m = 0, 1, \ldots)$ on both sides of (3.25), we can show that the $\boldsymbol{v}_{\rm on}^{(m)}(\theta_{\rm on})$ is identical to the stationary distribution of a Markov chain of M/G/1 type whose transition probability matrix is given by [Taki01b]

$$\begin{bmatrix}
\boldsymbol{B}_0 + \boldsymbol{B}_1 & \boldsymbol{B}_2 & \boldsymbol{B}_3 & \boldsymbol{B}_4 & \cdots \\
\boldsymbol{B}_0 & \boldsymbol{B}_1 & \boldsymbol{B}_2 & \boldsymbol{B}_3 & \cdots \\
\boldsymbol{O} & \boldsymbol{B}_0 & \boldsymbol{B}_1 & \boldsymbol{B}_2 & \cdots \\
\boldsymbol{O} & \boldsymbol{O} & \boldsymbol{B}_0 & \boldsymbol{B}_1 & \cdots \\
\boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{B}_0 & \cdots \\
\vdots & \vdots & \vdots & \vdots & \ddots
\end{bmatrix},$$

where

$$\begin{aligned}
\boldsymbol{B}_0 &= \boldsymbol{I} + \theta_{\rm on}^{-1}\left[\boldsymbol{C}_{\rm on} + \boldsymbol{D}_{\rm on}^{(0)}(\theta_{\rm on}) + \boldsymbol{E}_{\rm on,off}\boldsymbol{J}_{\rm off}^{(0)}(\theta_{\rm on})\boldsymbol{E}_{\rm off,on}\right], \\
\boldsymbol{B}_m &= \theta_{\rm on}^{-1}\left[\boldsymbol{D}_{\rm on}^{(m)}(\theta_{\rm on}) + \boldsymbol{E}_{\rm on,off}\boldsymbol{J}_{\rm off}^{(m)}(\theta_{\rm on})\boldsymbol{E}_{\rm off,on}\right], \quad m = 1, 2, \ldots.
\end{aligned}$$

Thus applying the general theory of Markov chains of M/G/1 type (see subsection 1.2.2), we can compute the $\boldsymbol{v}_{\rm on}^{(m)}(\theta_{\rm on})$.

On the other hand, the $\boldsymbol{v}_{\rm off}^{(m)}(\theta_{\rm on})$ can be computed by the following theorem whose proof is given in Appendix E.

Theorem 3.5 *The $\boldsymbol{v}_{\rm off}^{(m)}(\theta_{\rm on})$ is determined by the following recursion:*

$$\boldsymbol{v}_{\rm off}^{(0)}(\theta_{\rm on}) = \frac{\boldsymbol{v}_{\rm on}^{(0)}(\theta_{\rm on})\boldsymbol{E}_{\rm on,off}}{\boldsymbol{\pi}_{\rm on}\boldsymbol{E}_{\rm on,off}\boldsymbol{e}} \frac{\left[-\boldsymbol{C}_{\rm off} - \boldsymbol{D}_{\rm off}^{(0)}(\theta_{\rm on})\right]^{-1}}{\overline{I}_{\rm off}}, \tag{3.27}$$

$$\begin{aligned}
\boldsymbol{v}_{\rm off}^{(m)}(\theta_{\rm on}) &= \left[\frac{\boldsymbol{v}_{\rm on}^{(m)}(\theta_{\rm on})\boldsymbol{E}_{\rm on,off}}{\boldsymbol{\pi}_{\rm on}\boldsymbol{E}_{\rm on,off}\boldsymbol{e}}\frac{1}{\overline{I}_{\rm off}} + \sum_{l=0}^{m-1} \boldsymbol{v}_{\rm off}^{(l)}(\theta_{\rm on})\boldsymbol{D}_{\rm off}^{(m-l)}(\theta_{\rm on})\right] \\
&\quad \cdot \left[-\boldsymbol{C}_{\rm off} - \boldsymbol{D}_{\rm off}^{(0)}(\theta_{\rm on})\right]^{-1}, \qquad m = 1, 2, \ldots,
\end{aligned} \tag{3.28}$$

where the $\boldsymbol{D}_{\rm off}^{(m)}(\theta_{\rm on})$ is given in Lemma 3.3.

Among the recursions required in computing the joint queue length distribution, Lemma 3.2 for the $\overline{\boldsymbol{F}}_m(\boldsymbol{n})$ is the most extensive. In fact, its straightforward implementation will require very huge memory space. Note that an efficient implementation scheme for it is proposed in section 2.5. All other recursions can be readily implemented as they are. See [Taki94b] and chapter 2 in this thesis for details. Note also that from the results in this subsection, we can readily obtain recursions to compute the total queue length distribution, which are much less extensive than those for the joint queue length distribution. The results for the total queue length distribution are summarized in Appendix F.

3.5 Numerical Examples

In this section, we provide some numerical examples using two-class models. Throughout this section, we assume that the marginal arrival process follows a two-state batch MMAP ($\widetilde{\boldsymbol{C}}, \widetilde{\boldsymbol{D}}_1(n)$, $\widetilde{\boldsymbol{D}}_2(n)$):

$$\widetilde{\boldsymbol{C}} = \begin{bmatrix} -\lambda g^{-1} - a & a \\ a & -\lambda g^{-1} - a \end{bmatrix}, \tag{3.29}$$

$$\widetilde{\boldsymbol{D}}_1(n) = g(n) \begin{bmatrix} \lambda g^{-1} & 0 \\ 0 & 0 \end{bmatrix}, \quad \widetilde{\boldsymbol{D}}_2(n) = g(n) \begin{bmatrix} 0 & 0 \\ 0 & \lambda g^{-1} \end{bmatrix}, \qquad n = 1, 2, \ldots, \tag{3.30}$$

where $\lambda, g, a > 0$ and $g(n) = 1$ if $n = g$, and otherwise $g(n) = 0$.

We also assume that the marginal on-off process follows an alternating Markov renewal process whose infinitesimal generator is given by

$$\begin{bmatrix} \boldsymbol{S}_{\rm on} & \boldsymbol{T}_{\rm on,off} \\ \boldsymbol{T}_{\rm off,on} & \boldsymbol{S}_{\rm off} \end{bmatrix},$$

where $\boldsymbol{S}_{\rm on}$ (resp. $\boldsymbol{S}_{\rm off}$) denotes an $m_{\rm on} \times m_{\rm on}$ (resp. $m_{\rm off} \times m_{\rm off}$) matrix representing an infinitesimal generator that governs transitions in on-periods (resp. off-periods), and $\boldsymbol{T}_{\rm on,off}$ (resp. $\boldsymbol{T}_{\rm off,on}$) denotes a transition rate matrix from on-states (resp. off-states) to off-states (resp. on-states). When the on-off and arrival processes are independent of each other, the model is characterized as follows:

$$\begin{aligned}
&\boldsymbol{C}_{\rm on} = \boldsymbol{S}_{\rm on} \oplus \widetilde{\boldsymbol{C}}, \quad \boldsymbol{C}_{\rm off} = \boldsymbol{S}_{\rm off} \oplus \widetilde{\boldsymbol{C}}, \quad \boldsymbol{E}_{\rm on,off} = \boldsymbol{T}_{\rm on,off} \otimes \boldsymbol{I}(2), \quad \boldsymbol{E}_{\rm off,on} = \boldsymbol{T}_{\rm off,on} \otimes \boldsymbol{I}(2), \\
&\boldsymbol{D}_{1,\rm on}(n) = \boldsymbol{I}(m_{\rm on}) \otimes \widetilde{\boldsymbol{D}}_1(n), \quad \boldsymbol{D}_{1,\rm off}(n) = \boldsymbol{I}(m_{\rm off}) \otimes \widetilde{\boldsymbol{D}}_1(n), \\
&\boldsymbol{D}_{2,\rm on}(n) = \boldsymbol{I}(m_{\rm on}) \otimes \widetilde{\boldsymbol{D}}_2(n), \quad \boldsymbol{D}_{2,\rm off}(n) = \boldsymbol{I}(m_{\rm off}) \otimes \widetilde{\boldsymbol{D}}_2(n).
\end{aligned}$$

3.5.1 Impact of service time dependency

In this subsection, we discuss the impact of the service time dependency on the queue length. We assume that the on-off and arrival processes are mutually independent. Let $\boldsymbol{S}_{\rm on} = \boldsymbol{S}_{\rm off} = -\alpha$ and $\boldsymbol{T}_{\rm on,off} = \boldsymbol{T}_{\rm off,on} = \alpha$, where $\alpha > 0$. Also let $a = 0.1$ and $\lambda = 0.125$ in (3.29) and (3.30). As for the service time, we consider two cases, Case GD (class-dependent service times) and Case GI (i.i.d. service times):

$$\begin{aligned}
&[\textbf{Case GD}] && H_1 = 1 \text{ with probability } 1, \quad H_2 = 5 \text{ with probability } 1, \\
&[\textbf{Case GI}] && H_k = \begin{cases} 1, & \text{with probability } 1/2, \\ 5, & \text{with probability } 1/2, \end{cases} \qquad k = 1, 2,
\end{aligned}$$

where H_k ($k = 1, 2$) denotes a generic random variable representing a service time of a class k customer. Note here that the overall service time distributions in both cases are identical and $\rho_{\rm on} = 6\lambda = 0.75$.

Figure 3.1 plots the expected total queue lengths $\mathrm{E}[N]$ in Cases GD and GI as functions of α^{-1}. As α^{-1} goes to 0, the above model gets close to a work-conserving single-server queue (i.e., no service interruptions occur) with the same arrival process and service time distributions, where the

processing speed of the server is reduced by half. We observe that the difference in the expected total queue lengths in the two cases is kept almost constant regardless of the value of α^{-1} and gets large with constant batch size g.

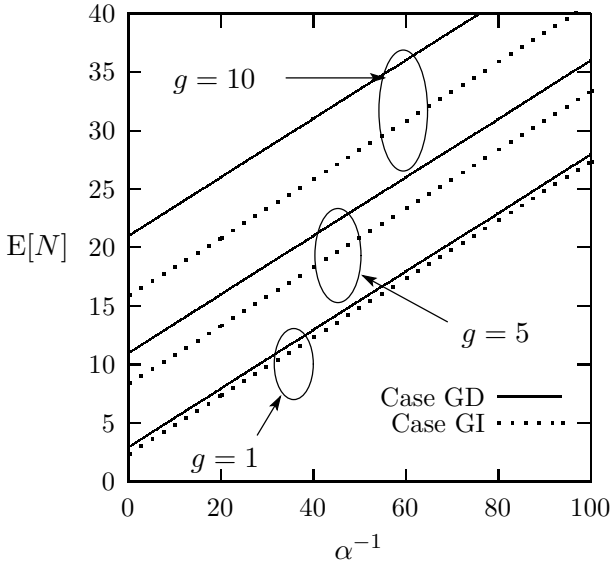

Figure 3.1: Expected total queue length E[N].

Table 3.1 shows the joint queue length distribution for $g = 1$. Let $\boldsymbol{p}_{\rm GD}(n_1, n_2)$ (resp. $\boldsymbol{p}_{\rm GI}(n_1, n_2)$) denote $\boldsymbol{p}(n_1, n_2)$ in Case GD (resp. GI). We observe that $\boldsymbol{p}_{\rm GI}(n_1, n_2)\boldsymbol{e} = \boldsymbol{p}_{\rm GI}(n_2, n_1)\boldsymbol{e}$ is due to the symmetry of input parameters of classes 1 and 2 in Case GI. Further, Table 3.1 shows that $\boldsymbol{p}_{\rm GD}(n_1, n_2)\boldsymbol{e} > \boldsymbol{p}_{\rm GI}(n_1, n_2)\boldsymbol{e}$ for $n_1 < n_2$, and vice versa. Thus for m, n such that $m > n$,

$$\boldsymbol{p}_{\rm GD}(m, n)\boldsymbol{e} < \boldsymbol{p}_{\rm GI}(m, n)\boldsymbol{e} = \boldsymbol{p}_{\rm GI}(n, m)\boldsymbol{e} < \boldsymbol{p}_{\rm GD}(n, m)\boldsymbol{e},$$

in this particular example. We conjecture that this phenomenon is caused by the fact that service times of class 2 customers are larger than those of class 1 customer in Case GD, and while a class 2 customer is being served, succeeding class 2 customers are likely to arrive back to back and stay in the system.

3.5.2 Impact of variation of on- and off-periods

Next, we discuss the impact of the variation in on- and off-periods on the total queue length N. We assume that the on-off process follows an alternating renewal process, and the on-off and arrival processes are mutually independent. Let $C^2_{v,\rm on}$ and $C^2_{v,\rm off}$ denote the squared coefficients of variation of on-periods and off-periods, respectively. To examine the impact of the variation of on-periods, the off-period distribution is fixed to be exponential with mean 100. For $C^2_{v,\rm on} = k^{-1} \le 1$ ($k = 1, 2, \ldots$), on-periods follow a k-stage Erlang distribution with mean 100, and for $C^2_{v,\rm on} > 1$, they follow a balanced hyper-exponential distribution $\psi(x)$ with mean 100, where

$$\psi(x) = 1 - p\exp(-0.02px) - (1-p)\exp\left[-0.02(1-p)x\right], \qquad 0 < p < 0.5.$$

Table 3.1: Joint queue length distribution $\boldsymbol{p}(n_1, n_2)\boldsymbol{e}$.
(Upper rows for Case GD and lower rows for Case GI)

n_2 \ n_1	0	1	2	5	10	20
0	1.34×10^{-1}	1.74×10^{-2}	5.67×10^{-3}	1.38×10^{-3}	1.79×10^{-4}	3.10×10^{-6}
	1.34×10^{-1}	3.41×10^{-2}	1.41×10^{-2}	2.51×10^{-3}	3.44×10^{-4}	6.80×10^{-6}
1	4.67×10^{-2}	9.37×10^{-3}	5.88×10^{-3}	2.12×10^{-3}	4.02×10^{-4}	1.14×10^{-5}
	3.41×10^{-2}	9.25×10^{-3}	6.25×10^{-3}	2.71×10^{-3}	6.02×10^{-4}	2.07×10^{-5}
2	2.18×10^{-2}	6.87×10^{-3}	5.31×10^{-3}	2.60×10^{-3}	6.48×10^{-4}	2.68×10^{-5}
	1.41×10^{-2}	6.25×10^{-3}	5.27×10^{-3}	3.03×10^{-3}	8.78×10^{-4}	4.38×10^{-5}
5	3.68×10^{-3}	3.26×10^{-3}	3.41×10^{-3}	2.87×10^{-3}	1.30×10^{-3}	1.24×10^{-4}
	2.51×10^{-3}	2.71×10^{-3}	3.03×10^{-3}	2.89×10^{-3}	1.50×10^{-3}	1.70×10^{-4}
10	5.02×10^{-4}	7.92×10^{-4}	1.09×10^{-3}	1.68×10^{-3}	1.54×10^{-3}	4.04×10^{-4}
	3.44×10^{-4}	6.02×10^{-4}	8.78×10^{-4}	1.50×10^{-3}	1.55×10^{-3}	4.73×10^{-4}
20	1.04×10^{-5}	2.97×10^{-5}	6.03×10^{-5}	2.15×10^{-4}	5.42×10^{-4}	6.12×10^{-4}
	6.80×10^{-6}	2.07×10^{-5}	4.38×10^{-5}	1.70×10^{-4}	4.73×10^{-4}	6.09×10^{-4}

Note that $C^2_{v,\mathrm{on}} = 1/\{2p(1-p)\} - 1$ in this case. On the other hand, in examining the impact of the variation of off-periods on the total queue length, the above on- and off-period distributions are exchanged.

As for the arrival process, we set $a = 0.1$, $\lambda = 0.125$ and $g = 1$ in (3.29) and (3.30). Besides, service times of each class are assumed to follow the same service time distribution as in Case GD of the preceding subsection.

Figures 3.2 and 3.3 plot the 99.9 percentile (99.9 PT) and expected value E[N] of the total queue length, respectively, as functions of the squared coefficient of variation $C^2_{v,\xi}$ ($\xi = \mathrm{on}, \mathrm{off}$), where the vertical axes are in log-scale. Note that in the case of $C^2_{v,\xi} = 1$ ($\xi = \mathrm{on}, \mathrm{off}$), the two models become identical with exponential on- and off-periods. We observe that both 99.9 PT and E[N] are monotone increasing functions of $C^2_{v,\xi}$ ($\xi = \mathrm{on}, \mathrm{off}$) and $C^2_{v,\mathrm{off}}$ has a more impact on the total queue length N than $C^2_{v,\mathrm{on}}$.

3.5.3 Impact of correlation in on- and off-periods

In this subsection, we examine the impact of the correlation in on- and off-periods on the total queue length N. For this purpose, we assume that the on-off and arrival processes are mutually independent and the marginal on-off process is given by

$$\boldsymbol{S}_{\mathrm{on}} = \boldsymbol{S}_{\mathrm{off}} = \begin{bmatrix} -1/40 & 0 \\ 0 & -1/160 \end{bmatrix}, \qquad \boldsymbol{T}_{\mathrm{on,off}} = \boldsymbol{T}_{\mathrm{off,on}} = \begin{bmatrix} p/40 & (1-p)/40 \\ (1-p)/160 & p/160 \end{bmatrix},$$

where $0 < p < 1$. Thus the marginal distributions of on- and off-periods are the same hyper-exponential distribution $\psi(x) = 1 - 0.5\exp(-x/40) - 0.5\exp(-x/160)$. Note that parameter p

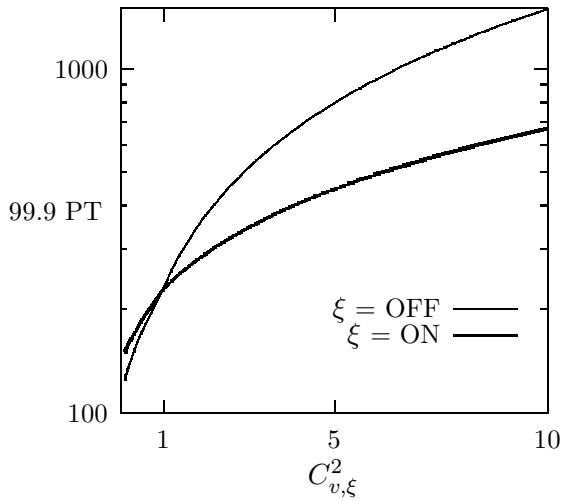

Figure 3.2: 99.9 percentile (99.9 PT) of the total queue length.

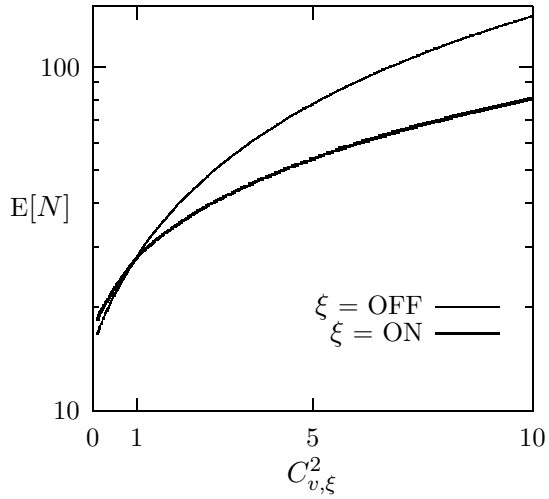

Figure 3.3: Expected total queue length E[N].

controls the correlation in consecutive on- and off-periods. Suppose the on-off process starts with an on-period. Let $I_{\rm on}(n)$ and $I_{\rm off}(n)$ $(n = 1, 2, \ldots)$ denote the lengths of the nth on- and off-periods, respectively. Then, both $\mathrm{Cov}[I_{\rm on}(n), I_{\rm off}(n)]$ and $\mathrm{Cov}[I_{\rm off}(n), I_{\rm on}(n+1)]$ are negative for $0 < p < 0.5$, equal to zero for $p = 0.5$, and positive for $0.5 < p < 1$. We set $a = 0.1$, $\lambda = 0.125$ and $g = 1$ in (3.29) and (3.30). Service times of each class follow the same distribution as in Case GD of subsection 3.5.1.

Figures 3.4 and 3.5 plot the 99.9 percentile (99.9 PT) and expected value E[N] of the total queue length, respectively, as functions of p. From these figures, we observe the followings. As p goes to zero, both 99.9 PT and E[N] rapidly increase. This phenomenon is due to the fact that once the on-off process is in a long off-period, long off-periods and short on-periods are likely to repeat alternately, and during those intervals, many customers are accumulated in the system. As p becomes large, however, this effect is weakened, and finally, both 99.9 PT and E[N] take their minimums and turn to increase. This implies that there exists some factor to make the queue length increase with p. In what follows, we examine this phenomenon more closely.

Let $\Psi_{\rm short-on}$, $\Psi_{\rm long-on}$, $\Psi_{\rm short-off}$ and $\Psi_{\rm long-off}$ denote the events that the on-off process is in a short on-period, long on-period, short off-period and long off-period, respectively. Figures 3.6 and 3.7 plot the conditional expected total queue lengths given those events as functions of p. From Figure 3.6, we observe that as expected, $\mathrm{E}[N \mid \Psi_{\rm long-off}]$ is always larger than $\mathrm{E}[N \mid \Psi_{\rm short-off}]$, so that the total queue length in an on-period following a long off-period is likely to be larger than that in an on-period following a short off-period, regardless of the value of p. Note here that as p goes to one, the contribution of the total queue length in an on-period following a long off-period to $\mathrm{E}[N \mid \Psi_{\rm long-on}]$ becomes large, and we conjecture that this factor makes $\mathrm{E}[N \mid \Psi_{\rm long-on}]$ increase in the region where p is close to one, as shown in Figure 3.7. Moreover, once $\mathrm{E}[N \mid \Psi_{\rm long-on}]$ turns to increase, this affects the total queue length in the following off-period, and as p goes to one, off-periods following long on-periods are likely to be long off-periods. Thus $\mathrm{E}[N \mid \Psi_{\rm long-off}]$ turns to increase after $\mathrm{E}[N \mid \Psi_{\rm long-on}]$ does, as shown in Figures 3.6 and 3.7.

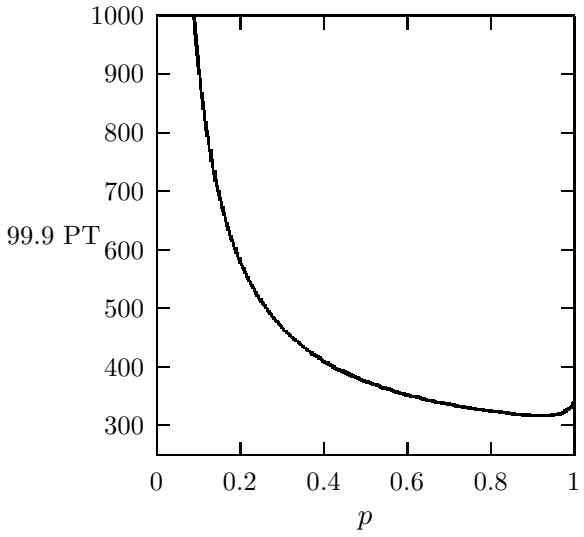

Figure 3.4: 99.9 percentile (99.9 PT) of the total queue length.

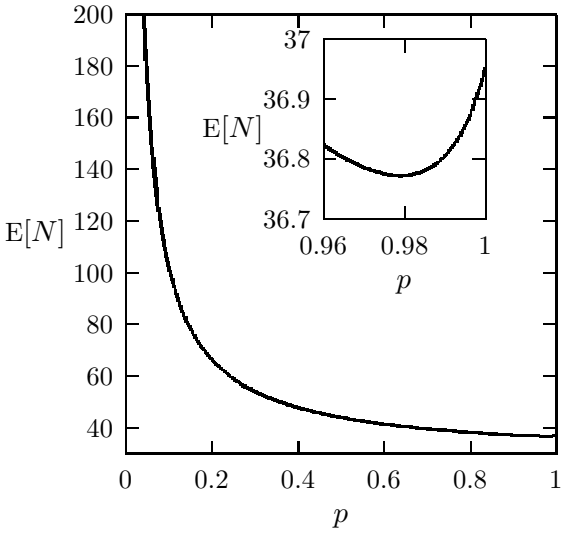

Figure 3.5: Expected total queue length E[N].

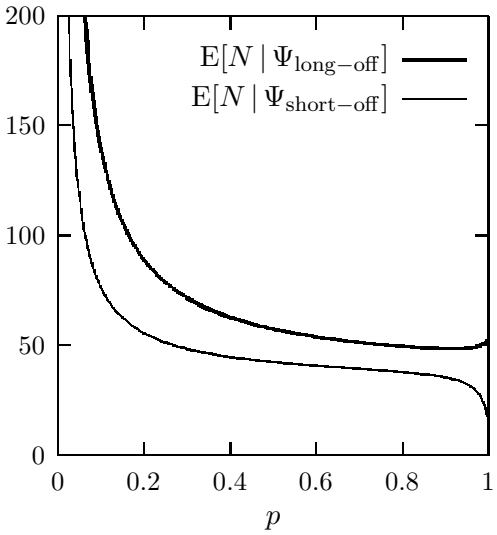

Figure 3.6: Conditional expected total queue length.

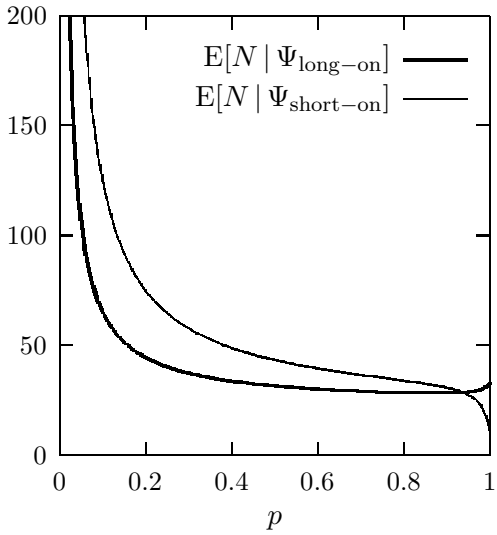

Figure 3.7: Conditional expected total queue length.

Note here that in this particular example,

$$\Pr(\Psi_{\text{long}-\text{on}}) = \Pr(\Psi_{\text{long}-\text{off}}) = 0.4, \quad \Pr(\Psi_{\text{short}-\text{on}}) = \Pr(\Psi_{\text{short}-\text{off}}) = 0.1.$$

Therefore the contributions of $\mathrm{E}[N \mid \Psi_{\text{long}-\text{on}}]$ and $\mathrm{E}[N \mid \Psi_{\text{long}-\text{off}}]$ to $\mathrm{E}[N]$ are four times as large as those of $\mathrm{E}[N \mid \Psi_{\text{short}-\text{on}}]$ and $\mathrm{E}[N \mid \Psi_{\text{short}-\text{off}}]$. As a result, $\mathrm{E}[N]$ increases for p near one. A similar observation can be applied to the 99.9 percentile, too.

3.5.4 Impact of correlation between on-off and arrival processes

Finally, we examine the impact of the correlation between on-off and arrival processes on the total queue length N. We consider the following three models, where service times of each class follow the same distribution as in Case GD of subsection 3.5.1.

In Model 1, the on-off and arrival processes have a correlation, and they are represented by

$$\boldsymbol{C} = \left[\begin{array}{c|c} -0.125-\alpha & \alpha \\ \hline \alpha & -0.125-\alpha \end{array}\right], \quad \boldsymbol{D}_1(1) = \left[\begin{array}{c|c} 0.125 & 0 \\ \hline 0 & 0 \end{array}\right], \quad \boldsymbol{D}_2(1) = \left[\begin{array}{c|c} 0 & 0 \\ \hline 0 & 0.125 \end{array}\right],$$

and $\boldsymbol{D}_k(n) = \boldsymbol{O}$ $(k = 1, 2)$ for $n = 2, 3, \ldots$, where $M_{\text{on}} = M_{\text{off}} = 1$. In Model 2, the on-off and arrival processes are mutually independent. As for the arrival process, we set $a = \alpha$, $\lambda = 0.125$ and $g = 1$ in (3.29) and (3.30). The on-off process is the same as that in subsection 3.5.1. In Model 3, the on-off and arrival processes have a correlation, and they are represented by

$$\boldsymbol{C} = \left[\begin{array}{c|c} -0.125-\alpha & \alpha \\ \hline \alpha & -0.125-\alpha \end{array}\right], \quad \boldsymbol{D}_1(1) = \left[\begin{array}{c|c} 0 & 0 \\ \hline 0 & 0.125 \end{array}\right], \quad \boldsymbol{D}_2(1) = \left[\begin{array}{c|c} 0.125 & 0 \\ \hline 0 & 0 \end{array}\right],$$

and $\boldsymbol{D}_k(n) = \boldsymbol{O}$ $(k = 1, 2)$ for $n = 2, 3, \ldots$, where $M_{\text{on}} = M_{\text{off}} = 1$. Note here that the marginal on-off processes in the three models are identical, and so are the marginal arrival processes.

Figures 3.8 and 3.9 show the 99.9 percentile (99.9 PT) and expected value $\mathrm{E}[N]$ of the total queue length, respectively, as functions of α^{-1}. Note here that as α^{-1} goes to zero, all the three models get close to a work-conserving single-server queue, where the arrival process follows a Poisson process with rate 0.125, and service times are i.i.d. and take 2 or 10 with equal probability. This is a reason why both 99.9 PTs and $\mathrm{E}[N]$'s in all the three models converge the same values, respectively, as $\alpha^{-1} \to 0$. We also observe that 99.9PT and $\mathrm{E}[N]$ in Model 1 (resp. Model 3) are always larger (resp. smaller) than those in Model 2. This is due to the fact the amount of work brought into the system during off-periods in Model 1 (resp. Model 3) is likely to be larger (resp. smaller) than that in Model 2.

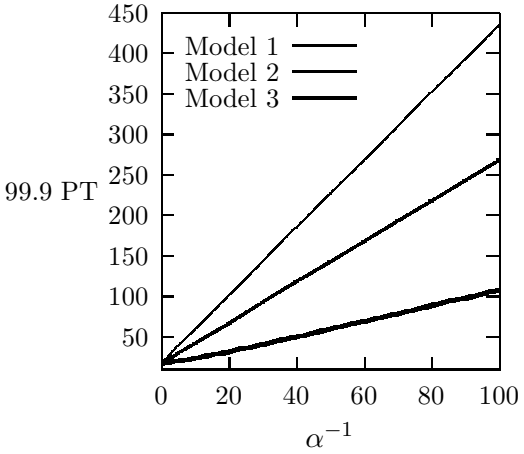

Figure 3.8: 99.9 percentile (99.9 PT) of the total queue length.

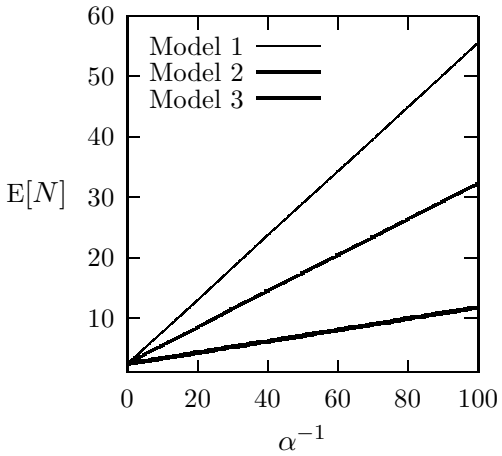

Figure 3.9: Expected total queue length E[N].

Chapter 4

Infinite-Server Queue

4.1 Introduction

For a few decades in the past, several studies have been done on infinite-server queues with batch arrivals, e.g., [Holm82, Holm83, Liu90, Liu91, Shan66]. In particular, Liu and Templeton [Liu91] studied an infinite-server queue, where the arrival process is assumed to be a Markov renewal process, the batch size distribution may depend on the state of the Markov renewal process, and service times of individual customers in the same batch are independent and identically distributed (i.i.d.), given the state of the Markov renewal process on arrival. To the best of our knowledge, this queue is the most general one among those studied in the past. They derived the time-dependent generating function and binomial moments of the number of customers in the system. However, most of those results are given in terms of integrals with respect to a certain Markov renewal function, and therefore numerical computations with those results are not easy to conduct, as stated in [Liu91].

This chapter considers an infinite-server queue fed by the Markovian input process defined in subsection 1.1.2. The input process has multiple arrival streams, and customer arrivals from respective streams occur with a predefined probability when transitions of the underlying Markov chain happen. Further, the service time distribution of customers may be different for different arrival streams. For such a queue, we first derive a system of ordinary differential equations for the time-dependent matrix joint generating function of the number of customers in the system. This is considered as a generalization of the known results for the N/G/∞ queue [Rama78] and for the $PH/G/\infty$ queue [Rama80]. As shown in [Rama80], applying a general-purpose numerical algorithm, we can compute the time-dependent joint distribution and the joint factorial moments of the number of customers in the system. Next we derive the time-dependent joint binomial moments of customers in the system, and assuming phase-type service times, we obtain explicit and numerically feasible expressions of the time-dependent and limiting joint binomial moments. Furthermore, through numerical examples, we reveal the impact of system parameters on the time-dependent and limiting performance.

The remainder of this chapter is organized as follows. In section 2, we mention the input process in this chapter. In section 3, we derive a system of ordinary differential equations for the time-

dependent matrix joint generating function of the number of customers in the system. In section 4, we derive the time-dependent joint binomial moments, and then assuming phase-type service times, we obtain numerically feasible expressions of the time-dependent and limiting joint binomial moments. Finally, in section 5, we show some numerical examples and discuss the impact of system parameters on the performance.

4.2 Model

This chapter considers an infinite-server queue with the Markovian input process introduced in subsection 1.1.2. Recall that the counting process of arrivals is characterized by $(\boldsymbol{C}, \boldsymbol{D}(\boldsymbol{n}))$. Customer arrivals can be viewed as follows. Let $S(t)$ denote the state of the underlying Markov chain (which governs the input process) at time t. Given $S(t) = i$ ($i \in \mathcal{M} = \{1, \ldots, M\}$), the conditional joint probability that n_k ($k \in \mathcal{K} = \{1, \ldots, K\}$) class k customers simultaneously arrive at the queue during time interval $(t, t+\delta t]$ and $S(t+\delta t) = j$ is given by $D_{i,j}(\boldsymbol{n})\delta t + o(\delta t)$. Besides, given $S(t) = i$, event $S(t+\delta t) = j$ ($j \neq i$) with no arrivals happens with probability $C_{i,j}\delta t + o(\delta t)$. As for service times, class k ($k \in \mathcal{K}$) customers have a service time distribution $H_k(t)$ ($t \geq 0$) with finite mean h_k.

In what follows, we assume that the arrival rate λ_k ($k \in \mathcal{K}$) of class k is positive and no customers are present in the system at time 0.

4.3 Time-Dependent Joint Distribution of the Number of Customers

In this section, we consider the time-dependent joint generating function of the number of customers of each class in the system. Let T_m denote the mth ($m = 1, 2, \ldots$) arrival epoch after time 0, where $0 < T_1 < T_2 < \cdots$. Let $X_{m,k}$ ($k \in \mathcal{K}$) denote a random variable representing the number of class k customers arriving at time T_m. We define $N_k(t)$ as a random variable representing the number of class k customers in the system at time t.

When no arrivals happen in time interval $(0, t]$, we have $N_k(t) = 0$ for all k ($k \in \mathcal{K}$). Thus

$$\mathrm{E}\left[z_1^{N_1(t)} \cdots z_K^{N_K(t)}\, 1\{S(t) = j,\, T_1 > t\} \,\middle|\, S(0) = i\right] = \left[e^{\boldsymbol{C}t}\right]_{i,j}, \tag{4.1}$$

where $1\{\xi\}$ denotes an indicator function of event ξ and $[\boldsymbol{A}]_{i,j}$ denotes the (i,j)th element of matrix $\boldsymbol{A}$.

Next we consider the case $T_\nu \leq t < T_{\nu+1}$ for some $\nu \geq 1$. Let ξ_ν ($\nu = 1, 2, \ldots$) denote the event $T_\nu \leq t < T_{\nu+1}$. Given the event ξ_ν happens, we define $\boldsymbol{x}_m$ ($m = 1, \ldots, \nu$) as $(X_{m,1}, \ldots, X_{m,K})$. Because customers are served independently, we have for $|z_k| \leq 1$ ($k \in \mathcal{K}$)

$$\begin{aligned}
&\mathrm{E}\left[z_1^{N_1(t)} \cdots z_K^{N_K(t)} \,\middle|\, \xi_\nu, T_m = t_m \text{ and } \boldsymbol{x}_m = (n_{m,1}, \ldots, n_{m,K}) \text{ for } m = 1, \ldots, \nu\right] \\
&\qquad = \prod_{m=1}^{\nu} \prod_{k_m \in \mathcal{K}} \left[H_{k_m}(t - t_m) + z_{k_m} \overline{H}_{k_m}(t - t_m)\right]^{n_{m,k_m}}
\end{aligned}$$

$$= \prod_{m=1}^{\nu} \prod_{k_m \in \mathcal{K}} \left[1 + (z_{k_m} - 1)\overline{H}_{k_m}(t - t_m)\right]^{n_{m,k_m}}, \tag{4.2}$$

where $\overline{H}_k(t) = 1 - H_k(t)$. Because events ξ_ν $(\nu = 1, 2, \ldots)$ are exclusive, it follows from (4.2) that

$$\begin{aligned}
&\mathrm{E}\left[z_1^{N_1(t)} \cdots z_K^{N_K(t)}\, 1\{S(t) = j \text{ and } T_1 < t\} \,\middle|\, S(0) = i\right] \\
&\quad = \sum_{\nu=1}^{\infty} \sum_{\boldsymbol{n}_1 \in \mathcal{Z}^+} \sum_{\boldsymbol{n}_2 \in \mathcal{Z}^+} \cdots \sum_{\boldsymbol{n}_\nu \in \mathcal{Z}^+} \int_0^t dt_1 \int_{t_1}^t dt_2 \cdots \int_{t_{\nu-1}}^t dt_\nu \\
&\qquad \cdot \prod_{m=1}^{\nu} \prod_{k_m \in \mathcal{K}} \left[1 + (z_{k_m} - 1)\overline{H}_{k_m}(t - t_m)\right]^{n_{m,k_m}} \\
&\qquad \cdot \left[e^{\boldsymbol{C}t_1} \boldsymbol{D}(\boldsymbol{n}_1) e^{\boldsymbol{C}(t_2 - t_1)} \boldsymbol{D}(\boldsymbol{n}_2) \cdot \cdots \cdot e^{\boldsymbol{C}(t_\nu - t_{\nu-1})} \boldsymbol{D}(\boldsymbol{n}_\nu) e^{\boldsymbol{C}(t - t_\nu)}\right]_{i,j},
\end{aligned} \tag{4.3}$$

where $\mathcal{Z}^+ = \{(n_1, \ldots, n_K); n_k\text{'s } (k \in \mathcal{K}) \text{ are nonnegative integers and at least one of them is strictly positive}\}$ and $\boldsymbol{n}_m = (n_{m,1}, \ldots, n_{m,K})$ for $m = 1, 2, \ldots$.

We now define $\boldsymbol{G}^*(t, \boldsymbol{z})$ as an $M \times M$ matrix whose (i, j)th $(i, j \in \mathcal{M})$ element represents

$$\mathrm{E}\left[z_1^{N_1(t)} \cdots z_K^{N_K(t)}\, 1\{S(t) = j\} \,\middle|\, S(0) = i\right],$$

where $\boldsymbol{z} = (z_1, \ldots, z_K)$ with $|z_k| \le 1$ for all k $(k \in \mathcal{K})$. It then follows from (4.1) and (4.3) that

$$\begin{aligned}
\boldsymbol{G}^*(t, \boldsymbol{z}) &= e^{\boldsymbol{C}t} + \sum_{\nu=1}^{\infty} \sum_{\boldsymbol{n}_1 \in \mathcal{Z}^+} \sum_{\boldsymbol{n}_2 \in \mathcal{Z}^+} \cdots \sum_{\boldsymbol{n}_\nu \in \mathcal{Z}^+} \int_0^t dt_1 \int_{t_1}^t dt_2 \cdots \int_{t_{\nu-1}}^t dt_\nu \\
&\quad \cdot \prod_{m=1}^{\nu} \prod_{k_m \in \mathcal{K}} \left[1 + (z_{k_m} - 1)\overline{H}_{k_m}(t - t_m)\right]^{n_{m,k_m}} \\
&\quad \cdot e^{\boldsymbol{C}t_1} \boldsymbol{D}(\boldsymbol{n}_1) e^{\boldsymbol{C}(t_2 - t_1)} \boldsymbol{D}(\boldsymbol{n}_2) \cdot \cdots \cdot e^{\boldsymbol{C}(t_\nu - t_{\nu-1})} \boldsymbol{D}(\boldsymbol{n}_\nu) e^{\boldsymbol{C}(t - t_\nu)}.
\end{aligned}$$

Further letting $u_j = t - t_{\nu+1-j}$ $(j = 1, \ldots, \nu)$ and rearranging terms, we have

$$\begin{aligned}
\boldsymbol{G}^*(t, \boldsymbol{z}) &= e^{\boldsymbol{C}t} + \sum_{\nu=1}^{\infty} \sum_{\boldsymbol{n}_\nu \in \mathcal{Z}^+} \sum_{\boldsymbol{n}_{\nu-1} \in \mathcal{Z}^+} \cdots \sum_{\boldsymbol{n}_1 \in \mathcal{Z}^+} \int_0^t du_\nu \int_0^{u_\nu} du_{\nu-1} \cdots \int_0^{u_2} du_1 \\
&\quad \cdot \prod_{m=1}^{\nu} \prod_{k_m \in \mathcal{K}} \left[1 + (z_{k_m} - 1)\overline{H}_{k_m}(u_m)\right]^{n_{m,k_m}} \\
&\quad \cdot e^{\boldsymbol{C}(t - u_\nu)} \boldsymbol{D}(\boldsymbol{n}_\nu) e^{\boldsymbol{C}(u_\nu - u_{\nu-1})} \boldsymbol{D}(\boldsymbol{n}_{\nu-1}) \cdot \cdots \cdot e^{\boldsymbol{C}(u_2 - u_1)} \boldsymbol{D}(\boldsymbol{n}_1) e^{\boldsymbol{C}u_1} \\
&= e^{\boldsymbol{C}t} + \sum_{\nu=1}^{\infty} \int_0^t du_\nu \int_0^{u_\nu} du_{\nu-1} \cdots \int_0^{u_2} du_1 e^{\boldsymbol{C}(t - u_\nu)} \boldsymbol{D}^*(u_\nu, \boldsymbol{z}) \\
&\quad \cdot e^{\boldsymbol{C}(u_\nu - u_{\nu-1})} \boldsymbol{D}^*(u_{\nu-1}, \boldsymbol{z}) \cdot \cdots \cdot e^{\boldsymbol{C}(u_2 - u_1)} \boldsymbol{D}^*(u_1, \boldsymbol{z}) e^{\boldsymbol{C}u_1},
\end{aligned} \tag{4.4}$$

where

$$\boldsymbol{D}^*(t, \boldsymbol{z}) = \sum_{\boldsymbol{n} \in \mathcal{Z}^+} \prod_{k \in \mathcal{K}} \left[1 + (z_k - 1)\overline{H}_k(t)\right]^{n_k} \boldsymbol{D}(\boldsymbol{n}). \tag{4.5}$$

Theorem 4.1 *For any closed interval of t where all $\overline{H}_\nu(t)$ are continuous, $\boldsymbol{G}^*(t, \boldsymbol{z})$ in (4.4) satisfies the following differential equation.*

$$\frac{\partial}{\partial t} \boldsymbol{G}^*(t, \boldsymbol{z}) = \left[\boldsymbol{C} + \boldsymbol{D}^*(t, \boldsymbol{z})\right] \boldsymbol{G}^*(t, \boldsymbol{z}). \tag{4.6}$$

Further (4.6) has a unique continuous solution in $[0, \infty)$ with $\boldsymbol{G}^*(0, \boldsymbol{z}) = \boldsymbol{I}$.

Proof. Pre-multiplying both sides of (4.4) by $\exp(-\boldsymbol{C}t)$, we have

$$\begin{aligned} e^{-\boldsymbol{C}t}\boldsymbol{G}^*(t,\boldsymbol{z}) &= \boldsymbol{I} + \sum_{\nu=1}^{\infty}\int_0^t du_\nu \int_0^{u_\nu} du_{\nu-1}\cdots\int_0^{u_2} du_1 e^{-\boldsymbol{C}u_\nu}\boldsymbol{D}^*(u_\nu,\boldsymbol{z}) \\ &\quad \cdot e^{\boldsymbol{C}(u_\nu-u_{\nu-1})}\boldsymbol{D}^*(u_{\nu-1},\boldsymbol{z})\cdot\cdots\cdot e^{\boldsymbol{C}(u_2-u_1)}\boldsymbol{D}^*(u_1,\boldsymbol{z})e^{\boldsymbol{C}u_1}. \end{aligned} \tag{4.7}$$

Differentiating both sides of (4.7) with respect to t yields

$$\begin{aligned} & e^{-\boldsymbol{C}t}\frac{\partial}{\partial t}\boldsymbol{G}^*(t,\boldsymbol{z}) - e^{-\boldsymbol{C}t}\boldsymbol{C}\boldsymbol{G}^*(t,\boldsymbol{z}) \\ &= e^{-\boldsymbol{C}t}\boldsymbol{D}^*(t,\boldsymbol{z})\Big[e^{\boldsymbol{C}t} + \sum_{\nu=2}^{\infty}\int_0^t du_{\nu-1}\int_0^{u_{\nu-1}} du_{\nu-2}\cdots\int_0^{u_2} du_1 e^{\boldsymbol{C}(t-u_{\nu-1})}\boldsymbol{D}^*(u_{\nu-1},\boldsymbol{z}) \\ &\qquad \cdot e^{\boldsymbol{C}(u_{\nu-1}-u_{\nu-2})}\boldsymbol{D}^*(u_{\nu-2},\boldsymbol{z})\cdot\cdots\cdot e^{\boldsymbol{C}(u_2-u_1)}\boldsymbol{D}^*(u_1,\boldsymbol{z})e^{\boldsymbol{C}u_1} \\ &= e^{-\boldsymbol{C}t}\boldsymbol{D}^*(t,\boldsymbol{z})\boldsymbol{G}^*(t,\boldsymbol{z}), \end{aligned} \tag{4.8}$$

where we use (4.4) in the last equality. Rearranging terms in (4.8) and pre-multiplying both sides by $\exp(\boldsymbol{C}t)$, we obtain (4.6). The uniqueness of the solution of (4.6) with $\boldsymbol{G}^*(0,\boldsymbol{z}) = \boldsymbol{I}$ follows from the well-known result of ordinary differential equations (see [Bell97], p.167, Theorem 1). ■

Remark 4.1 *For the $N/G/\infty$ queue and the $PH/G/\infty$ queue, systems of ordinary differential equations for the generating function of the number of customers are derived in [Rama78] and [Rama80], respectively. Theorem 4.1 is considered as a generalization of those results.*

We now consider the time-dependent joint distribution of the number of customers of each class in the system. Let $\boldsymbol{L}(t,\boldsymbol{n})$ $(\boldsymbol{n}\in\mathcal{Z})$ denote an $M\times M$ matrix whose (i,j)th element represents

$$\Pr[N_1(t)=n_1,\ldots,N_K(t)=n_K, S(t)=j \mid S(0)=i],$$

where $\mathcal{Z}=\{(n_1,\ldots,n_K); n_k=0,1,\ldots \text{ for all } k\in\mathcal{K}\}$. Let $\boldsymbol{D}(t,\boldsymbol{n})$ $(\boldsymbol{n}\in\mathcal{Z})$ denote an $M\times M$ matrix satisfying

$$\boldsymbol{D}^*(t,\boldsymbol{z}) = \sum_{\boldsymbol{n}\in\mathcal{Z}} z^{n_1}\cdots z^{n_K}\boldsymbol{D}(t,\boldsymbol{n}),$$

where $\boldsymbol{D}^*(t,\boldsymbol{z})$ is given in (4.5). For a given $\boldsymbol{n}\in\mathcal{Z}$, comparing the coefficient matrices of $z^{n_1}\cdots z^{n_K}$ on both sides of (4.6), we obtain

$$\frac{d}{dt}\boldsymbol{L}(t,\boldsymbol{n}) = (\boldsymbol{C}+\boldsymbol{D}(t,\boldsymbol{0}))\boldsymbol{L}(t,\boldsymbol{n}) + \sum_{\substack{\boldsymbol{0}\le\boldsymbol{m}\le\boldsymbol{n}\\ \boldsymbol{m}\neq\boldsymbol{0}}} \boldsymbol{D}(t,\boldsymbol{m})\boldsymbol{L}(t,\boldsymbol{n}-\boldsymbol{m}), \qquad \boldsymbol{n}\in\mathcal{Z}.$$

Therefore, for any given $\boldsymbol{n}\in\mathcal{Z}$, $\boldsymbol{L}(t,\boldsymbol{m})$'s $(\boldsymbol{m}\in\mathcal{Z},\ \boldsymbol{0}\le\boldsymbol{m}\le\boldsymbol{n})$ can be numerically obtained by solving the above system of ordinary differential equations with a general-purpose numerical algorithm (e.g., see [Rama80]).

Also, from (4.6), we can derive a system of ordinary differential equations for the joint factorial moments for the number of customers of each class in the system at time t. For this purpose, we

define $\boldsymbol{\Omega}(t,\boldsymbol{n})$ and $\boldsymbol{\Psi}(t,\boldsymbol{n})$ as

$$\begin{aligned}\boldsymbol{\Omega}(t,\boldsymbol{n}) &= \lim_{z_1\to 1}\cdots\lim_{z_K\to 1}\frac{\partial^{n_1}}{\partial z_1^{n_1}}\cdots\frac{\partial^{n_K}}{\partial z_K^{n_K}}\boldsymbol{G}^*(t,\boldsymbol{z}), \qquad \boldsymbol{n}\in\mathcal{Z},\\ \boldsymbol{\Psi}(t,\boldsymbol{n}) &= \lim_{z_1\to 1}\cdots\lim_{z_K\to 1}\frac{\partial^{n_1}}{\partial z_1^{n_1}}\cdots\frac{\partial^{n_K}}{\partial z_K^{n_K}}\boldsymbol{D}^*(t,\boldsymbol{z}), \qquad \boldsymbol{n}\in\mathcal{Z},\end{aligned}$$

respectively. By differentiating both sides of (4.6) n_k times with respect to z_k and setting $z_k = 1$ for all k ($k\in\mathcal{K}$), we obtain

$$\frac{\partial}{\partial t}\boldsymbol{\Omega}(t,\boldsymbol{n}) = (\boldsymbol{C}+\boldsymbol{D})\boldsymbol{\Omega}(t,\boldsymbol{n}) + \sum_{\substack{\boldsymbol{0}\le\boldsymbol{m}\le\boldsymbol{n}\\ \boldsymbol{m}\neq\boldsymbol{0}}}\prod_{k\in\mathcal{K}}\binom{n_k}{m_k}\boldsymbol{\Psi}(t,\boldsymbol{m})\boldsymbol{\Omega}(t,\boldsymbol{n}-\boldsymbol{m}), \qquad \boldsymbol{n}\in\mathcal{Z}.$$

Thus, the $\boldsymbol{\Omega}(t,\boldsymbol{n})$ is also obtained by solving the above system of ordinary differential equations.

4.4 Numerically Feasible Formulas for Phase-Type Service Times

In this section, we consider the joint binomial moments of the number of customers of each class in the system and develop numerically feasible formulas for the joint binomial moments, assuming phase-type service times. We define $\widehat{\boldsymbol{B}}(t,\boldsymbol{m})$ ($\boldsymbol{m}\in\mathcal{Z}^+$) as an $M\times M$ matrix whose (i,j)th element represents

$$\left[\widehat{\boldsymbol{B}}(t,\boldsymbol{m})\right]_{i,j} = \mathrm{E}\left[\prod_{k\in\mathcal{K}}\binom{N_k(t)}{m_k}1\{S(t)=j\}\,\middle|\,S(0)=i\right].$$

$\widehat{\boldsymbol{B}}(t,\boldsymbol{m})$ is called the $\boldsymbol{m}$th joint binomial moment matrix of the number of customers of each class in the system at time t.

4.4.1 Time-dependent joint binomial moments

We define $\boldsymbol{G}_B^*(t,\boldsymbol{\omega})$ as the matrix joint binomial moment generating function of the number of customers of each class in the system at time t, i.e.,

$$\boldsymbol{G}_B^*(t,\boldsymbol{\omega}) = e^{(\boldsymbol{C}+\boldsymbol{D})t} + \sum_{\boldsymbol{m}\in\mathcal{Z}^+}\omega_1^{m_1}\cdots\omega_K^{m_K}\widehat{\boldsymbol{B}}(t,\boldsymbol{m}), \tag{4.9}$$

where $\boldsymbol{\omega} = (\omega_1,\ldots,\omega_K)$ and $|\omega_k+1|\le 1$ for all k ($k\in\mathcal{K}$). Note here that $\boldsymbol{G}_B^*(t,\boldsymbol{\omega})$ is given in terms of $\boldsymbol{G}^*(t,\boldsymbol{z})$:

$$\boldsymbol{G}_B^*(t,\boldsymbol{\omega}) = \boldsymbol{G}^*(t,\omega_1+1,\ldots,\omega_K+1). \tag{4.10}$$

Theorem 4.2 *The time-dependent matrix joint binomial moment generating function $\boldsymbol{G}_B^*(t,\boldsymbol{\omega})$ is given by*

$$\begin{aligned}\boldsymbol{G}_B^*(t,\boldsymbol{\omega}) &= e^{(\boldsymbol{C}+\boldsymbol{D})t} + \sum_{\nu=1}^{\infty}\int_0^t du_\nu\int_0^{u_\nu}du_{\nu-1}\cdots\int_0^{u_2}du_1\, e^{(\boldsymbol{C}+\boldsymbol{D})(t-u_\nu)}\boldsymbol{D}_B^*(u_\nu,\boldsymbol{\omega})\\ &\qquad\cdot e^{(\boldsymbol{C}+\boldsymbol{D})(u_\nu-u_{\nu-1})}\boldsymbol{D}_B^*(u_{\nu-1},\boldsymbol{\omega})\\ &\qquad\qquad\cdots\cdots e^{(\boldsymbol{C}+\boldsymbol{D})(u_2-u_1)}\boldsymbol{D}_B^*(u_1,\boldsymbol{\omega})e^{(\boldsymbol{C}+\boldsymbol{D})u_1},\end{aligned} \tag{4.11}$$

where

$$\boldsymbol{D}_B^*(t,\boldsymbol{\omega}) = \boldsymbol{D}^*(t,\omega_1+1,\ldots,\omega_K+1) - \boldsymbol{D}. \tag{4.12}$$

Proof. It is easy to see from (4.6) that $\boldsymbol{G}_B^*(t,\boldsymbol{\omega})$ in (4.10) satisfies the following differential equation.

$$\begin{aligned}
\frac{\partial}{\partial t}\boldsymbol{G}_B^*(t,\boldsymbol{\omega}) &= [\boldsymbol{C}+\boldsymbol{D}+\boldsymbol{D}_B^*(t,\boldsymbol{\omega})]\,\boldsymbol{G}_B^*(t,\boldsymbol{\omega}), &(4.13)\\
\boldsymbol{G}_B^*(0,\boldsymbol{\omega}) &= \boldsymbol{I}. &(4.14)
\end{aligned}$$

It is clear that $\boldsymbol{G}_B^*(t,\boldsymbol{\omega})$ in (4.11) satisfies (4.14). In what follows, we shall show (4.11) is a solution of the differential equation (4.13).

Pre-multiplying both sides of (4.11) by $\exp[-(\boldsymbol{C}+\boldsymbol{D})t]$, we have

$$\begin{aligned}
e^{-(\boldsymbol{C}+\boldsymbol{D})t}\boldsymbol{G}_B^*(t,\boldsymbol{\omega}) &= \boldsymbol{I} + \sum_{\nu=1}^{\infty}\int_0^t du_\nu \int_0^{u_\nu} du_{\nu-1}\cdots\int_0^{u_2} du_1 e^{-(\boldsymbol{C}+\boldsymbol{D})u_\nu}\boldsymbol{D}_B^*(u_\nu,\boldsymbol{\omega})\\
&\qquad \cdot e^{(\boldsymbol{C}+\boldsymbol{D})(u_\nu-u_{\nu-1})}\boldsymbol{D}_B^*(u_{\nu-1},\boldsymbol{\omega})\\
&\qquad \cdots\cdots e^{(\boldsymbol{C}+\boldsymbol{D})(u_2-u_1)}\boldsymbol{D}_B^*(u_1,\boldsymbol{\omega})e^{(\boldsymbol{C}+\boldsymbol{D})u_1}. &(4.15)
\end{aligned}$$

Differentiating both sides of (4.15) with respect to t yields

$$\begin{aligned}
&e^{-(\boldsymbol{C}+\boldsymbol{D})t}\frac{\partial}{\partial t}\boldsymbol{G}_B^*(t,\boldsymbol{\omega}) - e^{-(\boldsymbol{C}+\boldsymbol{D})t}(\boldsymbol{C}+\boldsymbol{D})\boldsymbol{G}_B^*(t,\boldsymbol{\omega})\\
&\quad= e^{-(\boldsymbol{C}+\boldsymbol{D})t}\boldsymbol{D}_B^*(t,\boldsymbol{\omega})\Big[e^{(\boldsymbol{C}+\boldsymbol{D})t} + \sum_{\nu=2}^{\infty}\int_0^t du_{\nu-1}\int_0^{u_{\nu-1}} du_{\nu-2}\cdots\cdots\int_0^{u_2} du_1\\
&\qquad\qquad \cdot e^{(\boldsymbol{C}+\boldsymbol{D})(t-u_{\nu-1})}\boldsymbol{D}_B^*(u_{\nu-1},\boldsymbol{\omega})e^{(\boldsymbol{C}+\boldsymbol{D})(u_{\nu-1}-u_{\nu-2})}\boldsymbol{D}_B^*(u_{\nu-2},\boldsymbol{\omega})\\
&\qquad\qquad\qquad \cdots\cdots e^{(\boldsymbol{C}+\boldsymbol{D})(u_2-u_1)}\boldsymbol{D}_B^*(u_1,\boldsymbol{\omega})e^{(\boldsymbol{C}+\boldsymbol{D})u_1}\Big]\\
&\quad= e^{-(\boldsymbol{C}+\boldsymbol{D})t}\boldsymbol{D}_B^*(t,\boldsymbol{\omega})\boldsymbol{G}_B^*(t,\boldsymbol{\omega}), &(4.16)
\end{aligned}$$

where we use (4.11) in the last equality. Thus, pre-multiplying both sides of (4.16) by $\exp[(\boldsymbol{C}+\boldsymbol{D})t]$, we see that $\boldsymbol{G}_B^*(t,\boldsymbol{\omega})$ in (4.11) satisfies (4.13). ∎

We now rewrite $\boldsymbol{G}_B^*(t,\boldsymbol{\omega})$ in (4.11) to be a more appealing form. First, from (4.5) and (4.12), we have

$$\begin{aligned}
\boldsymbol{D}_B^*(t,\boldsymbol{\omega}) &= \sum_{\boldsymbol{n}\in\mathcal{Z}^+}\prod_{k\in\mathcal{K}}\left[1+\omega_k\overline{H}_k(t)\right]^{n_k}\boldsymbol{D}(\boldsymbol{n}) - \boldsymbol{D}\\
&= \sum_{\boldsymbol{n}\in\mathcal{Z}^+}\prod_{k\in\mathcal{K}}\left[\sum_{m_k=0}^{n_k}\binom{n_k}{m_k}\omega_k^{m_k}\overline{H}_k^{m_k}(t)\right]\boldsymbol{D}(\boldsymbol{n}) - \boldsymbol{D}\\
&= \sum_{\boldsymbol{m}\in\mathcal{Z}}\left[\prod_{k\in\mathcal{K}}\omega_k^{m_k}\overline{H}_k^{m_k}(t)\right]\sum_{\substack{\boldsymbol{n}\geq\boldsymbol{m}\\ \boldsymbol{n}\neq\boldsymbol{0}}}\prod_{k\in\mathcal{K}}\binom{n_k}{m_k}\boldsymbol{D}(\boldsymbol{n}) - \boldsymbol{D}\\
&= \sum_{\boldsymbol{n}\in\mathcal{Z}^+}\boldsymbol{D}(\boldsymbol{n}) + \sum_{\boldsymbol{m}\in\mathcal{Z}^+}\left[\prod_{k\in\mathcal{K}}\omega_k^{m_k}\overline{H}_k^{m_k}(t)\right]\sum_{\boldsymbol{n}\geq\boldsymbol{m}}\prod_{k\in\mathcal{K}}\binom{n_k}{m_k}\boldsymbol{D}(\boldsymbol{n}) - \boldsymbol{D}
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{\boldsymbol{m}\in\mathcal{Z}^+}\left[\prod_{k\in\mathcal{K}}\omega_k^{m_k}\overline{H}_k^{m_k}(t)\right]\sum_{\boldsymbol{n}\geq\boldsymbol{m}}\prod_{k\in\mathcal{K}}\binom{n_k}{m_k}\boldsymbol{D}(\boldsymbol{n})\\
&= \sum_{\boldsymbol{m}\in\mathcal{Z}^+}\omega^{(\boldsymbol{m})}\overline{H}^{(\boldsymbol{m})}(t)\widehat{\boldsymbol{D}}(\boldsymbol{m}),
\end{aligned}
\tag{4.17}
$$

where for $\boldsymbol{m}\in\mathcal{Z}^+$,

$$
\omega^{(\boldsymbol{m})}=\prod_{k\in\mathcal{K}}\omega_k^{m_k},\qquad \overline{H}^{(\boldsymbol{m})}(t)=\prod_{k\in\mathcal{K}}\overline{H}_k^{m_k}(t),\qquad \widehat{\boldsymbol{D}}(\boldsymbol{m})=\sum_{\boldsymbol{n}\geq\boldsymbol{m}}\prod_{k\in\mathcal{K}}\binom{n_k}{m_k}\boldsymbol{D}(\boldsymbol{n}).
$$

Thus, with (4.17), $\boldsymbol{G}_B^*(t,\boldsymbol{\omega})$ in (4.11) is rewritten to be

$$
\begin{aligned}
\boldsymbol{G}_B^*(t,\boldsymbol{\omega}) &= e^{(\boldsymbol{C}+\boldsymbol{D})t}+\sum_{\nu=1}^{\infty}\sum_{\boldsymbol{l}_1\in\mathcal{Z}^+}\sum_{\boldsymbol{l}_2\in\mathcal{Z}^+}\cdots\sum_{\boldsymbol{l}_\nu\in\mathcal{Z}^+}\omega^{(\boldsymbol{l}_1+\cdots+\boldsymbol{l}_\nu)}\\
&\qquad\cdot\int_0^t du_\nu\int_0^{u_\nu}du_{\nu-1}\cdots\int_0^{u_2}du_1\overline{H}^{(\boldsymbol{l}_\nu)}(u_\nu)\overline{H}^{(\boldsymbol{l}_{\nu-1})}(u_{\nu-1})\cdot\cdots\cdot\overline{H}^{(\boldsymbol{l}_1)}(u_1)\\
&\qquad\quad\cdot e^{(\boldsymbol{C}+\boldsymbol{D})(t-u_\nu)}\widehat{\boldsymbol{D}}(\boldsymbol{l}_\nu)e^{(\boldsymbol{C}+\boldsymbol{D})(u_\nu-u_{\nu-1})}\widehat{\boldsymbol{D}}(\boldsymbol{l}_{\nu-1})\\
&\qquad\qquad\cdots\cdot e^{(\boldsymbol{C}+\boldsymbol{D})(u_2-u_1)}\widehat{\boldsymbol{D}}(\boldsymbol{l}_1)e^{(\boldsymbol{C}+\boldsymbol{D})u_1}\\
&= e^{(\boldsymbol{C}+\boldsymbol{D})t}+\sum_{\boldsymbol{m}\in\mathcal{Z}^+}\omega^{(\boldsymbol{m})}\sum_{\nu=1}^{|\boldsymbol{m}|}\sum_{\vec{\boldsymbol{l}}_\nu\in\mathcal{L}_\nu(\boldsymbol{m})}\\
&\qquad\cdot\int_0^t du_\nu\int_0^{u_\nu}du_{\nu-1}\cdots\int_0^{u_2}du_1\,\overline{H}^{(\boldsymbol{l}_\nu)}(u_\nu)\overline{H}^{(\boldsymbol{l}_{\nu-1})}(u_{\nu-1})\cdot\cdots\cdot\overline{H}^{(\boldsymbol{l}_1)}(u_1)\\
&\qquad\quad\cdot e^{(\boldsymbol{C}+\boldsymbol{D})(t-u_\nu)}\widehat{\boldsymbol{D}}(\boldsymbol{l}_\nu)e^{(\boldsymbol{C}+\boldsymbol{D})(u_\nu-u_{\nu-1})}\widehat{\boldsymbol{D}}(\boldsymbol{l}_{\nu-1})\\
&\qquad\qquad\cdots\cdot e^{(\boldsymbol{C}+\boldsymbol{D})(u_2-u_1)}\widehat{\boldsymbol{D}}(\boldsymbol{l}_1)e^{(\boldsymbol{C}+\boldsymbol{D})u_1},
\end{aligned}
\tag{4.18}
$$

where

$$
\begin{aligned}
|\boldsymbol{m}| &= \sum_{k\in\mathcal{K}}m_k,\quad \boldsymbol{m}\in\mathcal{Z},\\
\boldsymbol{l}_j &= (l_{j,1},\ldots,l_{j,K})\in\mathcal{Z}^+,\\
\vec{\boldsymbol{l}}_\nu &= \left\{(\boldsymbol{l}_1,\ldots,\boldsymbol{l}_\nu)\,\middle|\,\boldsymbol{l}_j\in\mathcal{Z}^+,\ j=1,\ldots,\nu\right\},\\
\mathcal{L}_\nu(\boldsymbol{m}) &= \left\{\vec{\boldsymbol{l}}_\nu=(\boldsymbol{l}_1,\ldots,\boldsymbol{l}_\nu)\,\middle|\,\boldsymbol{l}_1+\cdots+\boldsymbol{l}_\nu=\boldsymbol{m},\widehat{\boldsymbol{D}}(\boldsymbol{l}_j)\neq\boldsymbol{O},\,\boldsymbol{l}_j\in\mathcal{Z}^+,j=1,\ldots,\nu\right\}.
\end{aligned}
$$

Therefore comparing (4.9) and (4.18), we obtain the following theorem.

Theorem 4.3 *The $\boldsymbol{m}$th ($\boldsymbol{m}\in\mathcal{Z}^+$) joint binomial moment matrix $\widehat{\boldsymbol{B}}(t,\boldsymbol{m})$ is given by*

$$
\begin{aligned}
\widehat{\boldsymbol{B}}(t,\boldsymbol{m}) &= \sum_{\nu=1}^{|\boldsymbol{m}|}\sum_{\vec{\boldsymbol{l}}_\nu\in\mathcal{L}_\nu(\boldsymbol{m})}\int_0^t du_\nu\int_0^{u_\nu}du_{\nu-1}\cdots\int_0^{u_2}du_1\\
&\qquad\cdot\overline{H}^{(\boldsymbol{l}_\nu)}(u_\nu)\overline{H}^{(\boldsymbol{l}_{\nu-1})}(u_{\nu-1})\cdot\cdots\cdot\overline{H}^{(\boldsymbol{l}_1)}(u_1)\\
&\qquad\quad\cdot e^{(\boldsymbol{C}+\boldsymbol{D})(t-u_\nu)}\widehat{\boldsymbol{D}}(\boldsymbol{l}_\nu)e^{(\boldsymbol{C}+\boldsymbol{D})(u_\nu-u_{\nu-1})}\widehat{\boldsymbol{D}}(\boldsymbol{l}_{\nu-1})\\
&\qquad\qquad\cdots\cdot e^{(\boldsymbol{C}+\boldsymbol{D})(u_2-u_1)}\widehat{\boldsymbol{D}}(\boldsymbol{l}_1)e^{(\boldsymbol{C}+\boldsymbol{D})u_1}.
\end{aligned}
\tag{4.19}
$$

Remark 4.2 *Note that* $\overline{H}^{(\boldsymbol{l})}(t)$ $(\boldsymbol{l} = (l_1, \ldots, l_K) \in \mathcal{Z}^+)$ *has probabilistic meanings. To see this, suppose* l_k $(k \in \mathcal{K})$ *class* k *customers simultaneously arrive. Let* $H_{k,i}$ $(k \in \mathcal{K}, i = 1, \ldots, l_k)$ *denote a random variable representing the service time of the* i*th class* k *customer.* $\overline{H}^{(\boldsymbol{l})}(t)$ *then represents the complementary distribution function of* $\min_{k\in\mathcal{K}, i=1,\ldots,l_k} H_{k,i}$*, because*

$$\begin{aligned}\overline{H}^{(\boldsymbol{l})}(t) &= \prod_{k\in\mathcal{K}}\prod_{i=1}^{l_k} \Pr(H_{k,i} > t) \\ &= \Pr\left(\min_{k\in\mathcal{K}, i=1,\ldots,l_k} H_{k,i} > t\right).\end{aligned}$$

Remark 4.3 $\widehat{\boldsymbol{B}}(t, \boldsymbol{m})$ *is independent of all* $\widehat{\boldsymbol{D}}(\boldsymbol{l})$ *(*$\boldsymbol{l} \geq \boldsymbol{m}$*,* $\boldsymbol{l} \neq \boldsymbol{m}$*). Thus the time-dependent binomial moment of the number of customers in the system depends on the batch size distribution only through its moments of the same and less order.*

4.4.2 Time-dependent formula for phase-type services

We now assume that service times of class k customers follow a phase-type distribution with representation $(\boldsymbol{\beta}_k, \boldsymbol{T}_k)$, i.e.,

$$\overline{H}_k(t) = \boldsymbol{\beta}_k e^{\boldsymbol{T}_k t}\boldsymbol{e}, \quad k \in \mathcal{K}.$$

Then using the following properties of Kronecker product $\otimes$ and Kronecker sum $\oplus$,

$$\begin{aligned}(\boldsymbol{U}_1\boldsymbol{U}_2 \cdot \cdots \cdot \boldsymbol{U}_n) \otimes (\boldsymbol{V}_1\boldsymbol{V}_2 \cdot \cdots \cdot \boldsymbol{V}_n) &= (\boldsymbol{U}_1 \otimes \boldsymbol{V}_1)(\boldsymbol{U}_2 \otimes \boldsymbol{V}_2) \cdot \cdots \cdot (\boldsymbol{U}_n \otimes \boldsymbol{V}_n), \quad \forall n \geq 1, \qquad (4.20)\\ \exp(\boldsymbol{U}) \otimes \exp(\boldsymbol{V}) &= \exp(\boldsymbol{U} \oplus \boldsymbol{V}), \qquad (4.21)\end{aligned}$$

we rewrite $\overline{H}^{(\boldsymbol{l})}(t)$ to be

$$\begin{aligned}\overline{H}^{(\boldsymbol{l})}(t) &= \prod_{k\in\mathcal{K}} \left(\boldsymbol{\beta}_k e^{\boldsymbol{T}_k t}\boldsymbol{e}\right)^{l_k} \\ &= \boldsymbol{\beta}^{<\boldsymbol{l}>} \exp\left(\boldsymbol{T}^{[\boldsymbol{l}]} t\right)\boldsymbol{e},\end{aligned}$$

where $\boldsymbol{\beta}^{<\boldsymbol{l}>}$ and $\boldsymbol{T}^{[\boldsymbol{l}]}$ $(\boldsymbol{l} \in \mathcal{Z}^+)$ are given by

$$\begin{aligned}\boldsymbol{\beta}^{<\boldsymbol{l}>} &= \underbrace{\boldsymbol{\beta}_1 \otimes \cdots \otimes \boldsymbol{\beta}_1}_{l_1} \otimes \underbrace{\boldsymbol{\beta}_2 \otimes \cdots \otimes \boldsymbol{\beta}_2}_{l_2} \otimes \cdots \otimes \underbrace{\boldsymbol{\beta}_K \otimes \cdots \otimes \boldsymbol{\beta}_K}_{l_K}, \\ \boldsymbol{T}^{[\boldsymbol{l}]} &= \underbrace{\boldsymbol{T}_1 \oplus \cdots \oplus \boldsymbol{T}_1}_{l_1} \oplus \underbrace{\boldsymbol{T}_2 \oplus \cdots \oplus \boldsymbol{T}_2}_{l_2} \oplus \cdots \oplus \underbrace{\boldsymbol{T}_K \oplus \cdots \oplus \boldsymbol{T}_K}_{l_K},\end{aligned}$$

respectively. Thus $\widehat{\boldsymbol{B}}(t, \boldsymbol{m})$ in (4.19) is rewritten to be

$$\widehat{\boldsymbol{B}}(t, \boldsymbol{m}) = \sum_{\nu=1}^{|\boldsymbol{m}|} \sum_{\vec{\boldsymbol{l}}_\nu \in \mathcal{L}_\nu(\boldsymbol{m})} \boldsymbol{F}_\nu(t, \vec{\boldsymbol{l}}_\nu), \qquad (4.22)$$

where

$$\begin{aligned}\boldsymbol{F}_\nu(t,\vec{\boldsymbol{l}}_\nu) &= \int_0^t du_\nu \int_0^{u_\nu} du_{\nu-1}\cdots\int_0^{u_2} du_1 \prod_{j=1}^{\nu}\left[\boldsymbol{\beta}^{<\boldsymbol{l}_j>}\exp\left(\boldsymbol{T}^{[\boldsymbol{l}_j]}u_j\right)\boldsymbol{e}\right] \\ &\quad\cdot e^{(\boldsymbol{C}+\boldsymbol{D})(t-u_\nu)}\widehat{\boldsymbol{D}}(\boldsymbol{l}_\nu)e^{(\boldsymbol{C}+\boldsymbol{D})(u_\nu-u_{\nu-1})}\widehat{\boldsymbol{D}}(\boldsymbol{l}_{\nu-1}) \\ &\quad\cdots\cdots e^{(\boldsymbol{C}+\boldsymbol{D})(u_2-u_1)}\widehat{\boldsymbol{D}}(\boldsymbol{l}_1)e^{(\boldsymbol{C}+\boldsymbol{D})u_1}. \end{aligned} \tag{4.23}$$

To obtain a numerically feasible formula for $\boldsymbol{F}_\nu(t,\vec{\boldsymbol{l}}_\nu)$, we shall rewrite (4.23) by considering the time-reversed arrival process. We define $\boldsymbol{C}^-$ and $\boldsymbol{D}^-(\boldsymbol{n})$ as

$$\boldsymbol{C}^- = \mathrm{diag}\,(\boldsymbol{\pi})^{-1}\,\boldsymbol{C}^{\mathrm{T}}\mathrm{diag}\,(\boldsymbol{\pi})\,, \qquad \boldsymbol{D}^-(\boldsymbol{n}) = \mathrm{diag}\,(\boldsymbol{\pi})^{-1}\,\boldsymbol{D}(\boldsymbol{n})^{\mathrm{T}}\mathrm{diag}\,(\boldsymbol{\pi})\,,$$

respectively, where $\mathrm{diag}\,(\boldsymbol{\pi})$ denotes an $M\times M$ diagonal matrix whose ith diagonal element is equal to the ith element of $\boldsymbol{\pi}$. The time-reversed arrival process can then be characterized by $\boldsymbol{C}^-$ and $\boldsymbol{D}^-(\boldsymbol{n})$. Therefore (4.23) is rewritten to be

$$\begin{aligned}\boldsymbol{F}_\nu(t,\vec{\boldsymbol{l}}_\nu) &= \left[\prod_{\eta=1}^{\nu} h(\boldsymbol{l}_\eta)d(\boldsymbol{l}_\eta)\right]\mathrm{diag}\,(\boldsymbol{\pi})^{-1} \\ &\quad\cdot\Bigg[\int_0^t du_\nu\int_0^{u_\nu} du_{\nu-1}\cdots\int_0^{u_2} du_1\prod_{j=1}^{\nu}\left[\boldsymbol{\kappa}^{(\boldsymbol{l}_j)}\exp\left(\boldsymbol{T}^{[\boldsymbol{l}_j]}u_j\right)\left(-\boldsymbol{T}^{[\boldsymbol{l}_j]}\right)\boldsymbol{e}\right] \\ &\qquad\cdot\exp\left(\boldsymbol{Q}^-u_1\right)\widehat{\boldsymbol{D}}^-(\boldsymbol{l}_1)\exp\left[\boldsymbol{Q}^-(u_2-u_1)\right]\widehat{\boldsymbol{D}}^-(\boldsymbol{l}_2) \\ &\qquad\cdots\cdots\exp\left[\boldsymbol{Q}^-(u_\nu-u_{\nu-1})\right]\widehat{\boldsymbol{D}}^-(\boldsymbol{l}_\nu)\exp\left[\boldsymbol{Q}^-(t-u_\nu)\right]\Bigg]^T\mathrm{diag}\,(\boldsymbol{\pi})\,, \end{aligned}$$

where

$$\boldsymbol{Q}^- = \mathrm{diag}\,(\boldsymbol{\pi})^{-1}\,(\boldsymbol{C}+\boldsymbol{D})^{\mathrm{T}}\mathrm{diag}\,(\boldsymbol{\pi})\,,$$

and for $\boldsymbol{l}\in\mathcal{Z}^+$,

$$h(\boldsymbol{l}) = \int_0^\infty dt\,\boldsymbol{\beta}^{<\boldsymbol{l}>}\exp\left(\boldsymbol{T}^{[\boldsymbol{l}]}t\right)\boldsymbol{e} = \boldsymbol{\beta}^{<\boldsymbol{l}>}\left(-\boldsymbol{T}^{[\boldsymbol{l}]}\right)^{-1}\boldsymbol{e}, \tag{4.24}$$

$$d(\boldsymbol{l}) = \max_{i\in\mathcal{M}}\left[\mathrm{diag}\,(\boldsymbol{\pi})^{-1}\,\widehat{\boldsymbol{D}}(\boldsymbol{l})^{\mathrm{T}}\mathrm{diag}\,(\boldsymbol{\pi})\,\boldsymbol{e}\right]_i, \tag{4.25}$$

$$\widehat{\boldsymbol{D}}^-(\boldsymbol{l}) = d(\boldsymbol{l})^{-1}\,\mathrm{diag}\,(\boldsymbol{\pi})^{-1}\,\widehat{\boldsymbol{D}}(\boldsymbol{l})^{\mathrm{T}}\mathrm{diag}\,(\boldsymbol{\pi})\,,$$

$$\boldsymbol{\kappa}^{(\boldsymbol{l})} = \frac{\boldsymbol{\beta}^{<\boldsymbol{l}>}\left(-\boldsymbol{T}^{[\boldsymbol{l}]}\right)^{-1}}{\boldsymbol{\beta}^{<\boldsymbol{l}>}\left(-\boldsymbol{T}^{[\boldsymbol{l}]}\right)^{-1}\boldsymbol{e}}.$$

Remark 4.4 *$\boldsymbol{Q}^-$ represents the infinitesimal generator of the time-reversed process of the underlying Markov chain that governs the arrival process, and satisfies $\boldsymbol{\pi}\boldsymbol{Q}^- = \boldsymbol{0}$. Note also that $\widehat{\boldsymbol{D}}^-(\boldsymbol{l})$ is a nonnegative matrix whose row sums are all equal to or less than one, i.e., $\widehat{\boldsymbol{D}}^-(\boldsymbol{l})$ is a substochastic matrix. Further*

$$\boldsymbol{\kappa}^{(\boldsymbol{l}_j)}\exp\left(\boldsymbol{T}^{[\boldsymbol{l}_j]}u_j\right)\left(-\boldsymbol{T}^{[\boldsymbol{l}_j]}\right)\boldsymbol{e}, \qquad j=1,2,\ldots$$

is considered as the density function of the phase-type distribution with representation $(\boldsymbol{\kappa}^{(\boldsymbol{l}_j)}, \boldsymbol{T}^{[\boldsymbol{l}_j]})$.

We now define $\boldsymbol{F}_{\nu}^{-}(t,\vec{\boldsymbol{l}}_{\nu})$ $(\nu = 1,2,\ldots)$ as

$$\begin{aligned}\boldsymbol{F}_{\nu}^{-}(t,\vec{\boldsymbol{l}}_{\nu}) &= \int_0^t du_{\nu}\int_0^{u_{\nu}} du_{\nu-1}\cdots\int_0^{u_2} du_1 \prod_{j=1}^{\nu}\left[\boldsymbol{\kappa}^{(\boldsymbol{l}_j)}\exp\left(\boldsymbol{T}^{[\boldsymbol{l}_j]}u_j\right)\left(-\boldsymbol{T}^{[\boldsymbol{l}_j]}\right)\boldsymbol{e}\right] \\ &\quad\cdot\exp\left(\boldsymbol{Q}^{-}u_1\right)\widehat{\boldsymbol{D}}^{-}(\boldsymbol{l}_1)\exp\left[\boldsymbol{Q}^{-}(u_2-u_1)\right]\widehat{\boldsymbol{D}}^{-}(\boldsymbol{l}_2) \\ &\quad\cdot\cdots\cdot\exp\left[\boldsymbol{Q}^{-}(u_{\nu}-u_{\nu-1})\right]\widehat{\boldsymbol{D}}^{-}(\boldsymbol{l}_{\nu})\exp\left[\boldsymbol{Q}^{-}(t-u_{\nu})\right].\end{aligned}$$

$\boldsymbol{F}_{\nu}(t,\vec{\boldsymbol{l}}_{\nu})$ can then be rewritten in terms of $\boldsymbol{F}_{\nu}^{-}(t,\vec{\boldsymbol{l}}_{\nu})$ as follows:

$$\boldsymbol{F}_{\nu}(t,\vec{\boldsymbol{l}}_{\nu}) = \left[\prod_{\eta=1}^{\nu} h(\boldsymbol{l}_{\eta})d(\boldsymbol{l}_{\eta})\right]\mathrm{diag}\left(\boldsymbol{\pi}\right)^{-1}\left[\boldsymbol{F}_{\nu}^{-}(t,\vec{\boldsymbol{l}}_{\nu})\right]^{\mathrm{T}}\mathrm{diag}\left(\boldsymbol{\pi}\right). \tag{4.26}$$

In what follows, we derive a numerically feasible formula for $\boldsymbol{F}_{\nu}^{-}(t,\boldsymbol{l}_{\nu})$.

We first consider $\boldsymbol{F}_{1}^{-}(t,\boldsymbol{l}_1)$;

$$\begin{aligned}\boldsymbol{F}_{1}^{-}(t,\boldsymbol{l}_1) &= \int_0^t du_1\left[\boldsymbol{\kappa}^{(\boldsymbol{l}_1)}\exp\left(\boldsymbol{T}^{[\boldsymbol{l}_1]}u_1\right)\left(-\boldsymbol{T}^{[\boldsymbol{l}_1]}\right)\boldsymbol{e}\right] \\ &\quad\cdot\exp(\boldsymbol{Q}^{-}u_1)\widehat{\boldsymbol{D}}^{-}(\boldsymbol{l}_1)\exp\left[\boldsymbol{Q}^{-}(t-u_1)\right]. \end{aligned}\tag{4.27}$$

Using (4.20) and (4.21), we rewrite (4.27) to be

$$\begin{aligned}\boldsymbol{F}_{1}^{-}(t,\boldsymbol{l}_1) &= \int_0^t du_1\left[\boldsymbol{\kappa}^{(\boldsymbol{l}_1)}\cdot\exp\left(\boldsymbol{T}^{[\boldsymbol{l}_1]}u_1\right)\cdot\left(-\boldsymbol{T}^{[\boldsymbol{l}_1]}\right)\boldsymbol{e}\right] \\ &\quad\cdot\left[\boldsymbol{I}_Q\cdot\exp(\boldsymbol{Q}^{-}u_1)\cdot\widehat{\boldsymbol{D}}^{-}(\boldsymbol{l}_1)\right]\exp\left[\boldsymbol{Q}^{-}(t-u_1)\right] \\ &= \left(\boldsymbol{\kappa}^{(\boldsymbol{l}_1)}\otimes\boldsymbol{I}_Q\right)\int_0^t du_1\,\exp\left(\boldsymbol{T}^{[\boldsymbol{l}_1]}\oplus\boldsymbol{Q}^{-}u_1\right) \\ &\quad\cdot\left[\left(-\boldsymbol{T}^{[\boldsymbol{l}_1]}\right)\boldsymbol{e}\otimes\widehat{\boldsymbol{D}}^{-}(\boldsymbol{l}_1)\right]\exp\left[\boldsymbol{Q}^{-}(t-u_1)\right],\end{aligned}$$

where $\boldsymbol{I}_Q$ denotes an identity matrix of the same size as $\boldsymbol{Q}^{-}$.

Similarly, $\boldsymbol{F}_{2}^{-}(t,\vec{\boldsymbol{l}}_2)$ is rewritten to be

$$\begin{aligned}\boldsymbol{F}_{2}^{-}(t,\vec{\boldsymbol{l}}_2) &= \int_0^t du_2\int_0^{u_2} du_1 \\ &\quad\cdot\left[\boldsymbol{\kappa}^{(\boldsymbol{l}_1)}\cdot\exp\left(\boldsymbol{T}^{[\boldsymbol{l}_1]}u_1\right)\cdot\left(-\boldsymbol{T}^{[\boldsymbol{l}_1]}\right)\boldsymbol{e}\cdot 1\cdot 1\cdot 1\right] \\ &\quad\cdot\left[\boldsymbol{\kappa}^{(\boldsymbol{l}_2)}\cdot\exp\left(\boldsymbol{T}^{[\boldsymbol{l}_2]}u_1\right)\cdot\boldsymbol{I}_1(\boldsymbol{l}_2)\cdot\exp\left[\boldsymbol{T}^{[\boldsymbol{l}_2]}(u_2-u_1)\right]\cdot\left(-\boldsymbol{T}^{[\boldsymbol{l}_2]}\right)\boldsymbol{e}\cdot 1\right] \\ &\quad\cdot\left[\boldsymbol{I}_Q\cdot\exp(\boldsymbol{Q}^{-}u_1)\cdot\widehat{\boldsymbol{D}}^{-}(\boldsymbol{l}_1)\cdot\ \exp[\boldsymbol{Q}^{-}(u_2-u_1)]\cdot\widehat{\boldsymbol{D}}^{-}(\boldsymbol{l}_2)\cdot\exp[\boldsymbol{Q}^{-}(t-u_2)]\right] \\ &= (\boldsymbol{\kappa}^{(\boldsymbol{l}_1)}\otimes\boldsymbol{\kappa}^{(\boldsymbol{l}_2)}\otimes\boldsymbol{I}_Q)\int_0^t du_2\int_0^{u_2} du_1\exp\left(\boldsymbol{T}^{[\boldsymbol{l}_1]}\oplus\boldsymbol{T}^{[\boldsymbol{l}_2]}\oplus\boldsymbol{Q}^{-}u_1\right) \\ &\quad\cdot\left[\left(-\boldsymbol{T}^{[\boldsymbol{l}_1]}\right)\boldsymbol{e}\otimes\boldsymbol{I}_1(\boldsymbol{l}_2)\otimes\widehat{\boldsymbol{D}}^{-}(\boldsymbol{l}_1)\right]\exp\left[\boldsymbol{T}^{[\boldsymbol{l}_2]}\oplus\boldsymbol{Q}^{-}(u_2-u_1)\right] \\ &\quad\cdot\left[\left(-\boldsymbol{T}^{[\boldsymbol{l}_2]}\right)\boldsymbol{e}\otimes\widehat{\boldsymbol{D}}^{-}(\boldsymbol{l}_2)\right]\exp[\boldsymbol{Q}^{-}(t-u_2)],\end{aligned}$$

where $\boldsymbol{I}_1(\boldsymbol{l}_2)$ denotes an identity matrix of the same size as $\boldsymbol{T}^{[\boldsymbol{l}_2]}$. Following the same manipulation as in the cases $\nu = 1$ and 2, we obtain for $\nu = 1,2,\ldots,$

$$\boldsymbol{F}_{\nu}^{-}(t,\vec{\boldsymbol{l}}_{\nu}) = \boldsymbol{J}_{\nu}(\vec{\boldsymbol{l}}_{\nu})\int_0^t du_{\nu}\int_0^{u_{\nu}} du_{\nu-1}\cdots\int_0^{u_2} du_1$$

$$\cdot \exp\left[\boldsymbol{U}_{\nu,1}(\vec{\boldsymbol{l}}_\nu)u_1\right]\boldsymbol{V}_{\nu,1}(\vec{\boldsymbol{l}}_\nu)\exp\left[\boldsymbol{U}_{\nu,2}(\vec{\boldsymbol{l}}_\nu)(u_2-u_1)\right]\boldsymbol{V}_{\nu,2}(\vec{\boldsymbol{l}}_\nu)$$
$$\cdot\cdots\cdot\exp\left[\boldsymbol{U}_{\nu,\nu}(\vec{\boldsymbol{l}}_\nu)(u_\nu-u_{\nu-1})\right]\boldsymbol{V}_{\nu,\nu}(\vec{\boldsymbol{l}}_\nu)\exp[\boldsymbol{Q}^-(t-u_\nu)], \tag{4.28}$$

where

$$\boldsymbol{J}_\nu(\vec{\boldsymbol{l}}_\nu) = \boldsymbol{\kappa}^{(l_1)}\otimes\cdots\otimes\boldsymbol{\kappa}^{(l_\nu)}\otimes\boldsymbol{I}_Q, \tag{4.29}$$

$$\boldsymbol{U}_{\nu,j}(\vec{\boldsymbol{l}}_\nu) = \boldsymbol{T}^{[l_j]}\oplus\cdots\oplus\boldsymbol{T}^{[l_\nu]}\oplus\boldsymbol{Q}^-, \qquad j=1,\ldots,\nu, \tag{4.30}$$

$$\boldsymbol{V}_{\nu,j}(\vec{\boldsymbol{l}}_\nu) = \begin{cases}\left(-\boldsymbol{T}^{[l_j]}\right)\boldsymbol{e}\otimes\boldsymbol{I}_{\nu-j}(\boldsymbol{l}_{j+1},\ldots,\boldsymbol{l}_\nu)\otimes\widehat{\boldsymbol{D}}^-(\boldsymbol{l}_j), & j=1,\ldots,\nu-1,\\ \left(-\boldsymbol{T}^{[l_\nu]}\right)\boldsymbol{e}\otimes\widehat{\boldsymbol{D}}^-(\boldsymbol{l}_\nu), & j=\nu,\end{cases} \tag{4.31}$$

and where $\boldsymbol{I}_{\nu-j}(\boldsymbol{l}_{j+1},\boldsymbol{l}_{j+2},\ldots,\boldsymbol{l}_\nu)$ $(\nu-j\geq 1)$ denotes an identity matrix whose size is equal to that of $\boldsymbol{T}^{[l_{j+1}]}\oplus\boldsymbol{T}^{[l_{j+2}]}\oplus\cdots\oplus\boldsymbol{T}^{[l_\nu]}$.

Lemma 4.1 *Let $\boldsymbol{U}_j$ $(j=1,\ldots,\nu)$ denote an $m_i\times m_i$ matrix and $\boldsymbol{V}_j$ $(j=0,\ldots,\nu)$ denote an $m_j\times m_{j+1}$ matrix. If $\boldsymbol{U}_j$'s and $\boldsymbol{V}_j$'s are all bounded, the following equation holds for all ν $(\nu=2,3\ldots)$.*

$$\boldsymbol{V}_0\int_0^t du_{\nu-1}\int_0^{u_{\nu-1}}du_{\nu-2}\cdots\int_0^{u_2}du_1 e^{\boldsymbol{U}_1u_1}\boldsymbol{V}_1e^{\boldsymbol{U}_2(u_2-u_1)}\boldsymbol{V}_2$$
$$\cdots\cdot e^{\boldsymbol{U}_{\nu-1}(u_{\nu-1}-u_{\nu-2})}\boldsymbol{V}_{\nu-1}e^{\boldsymbol{U}_\nu(t-u_{\nu-1})}\boldsymbol{V}_\nu$$
$$=\begin{bmatrix}\boldsymbol{V}_0 & \boldsymbol{O} & \cdots & \boldsymbol{O}\end{bmatrix}\exp\left[\begin{pmatrix}\boldsymbol{U}_1 & \boldsymbol{V}_1 & \boldsymbol{O} & \cdots & \boldsymbol{O}\\ \boldsymbol{O} & \boldsymbol{U}_2 & \boldsymbol{V}_2 & & \vdots\\ \vdots & & \ddots & \ddots & \boldsymbol{O}\\ \boldsymbol{O} & \cdots & \boldsymbol{O} & \boldsymbol{U}_{\nu-1} & \boldsymbol{V}_{\nu-1}\\ \boldsymbol{O} & \boldsymbol{O} & \cdots & \boldsymbol{O} & \boldsymbol{U}_\nu\end{pmatrix}t\right]\begin{bmatrix}\boldsymbol{O}\\ \vdots\\ \boldsymbol{O}\\ \boldsymbol{V}_\nu\end{bmatrix}. \tag{4.32}$$

The proof of Lemma 4.1 is given in Appendix G.

We define $\boldsymbol{T}_\nu(\vec{\boldsymbol{l}}_\nu)$ as

$$\boldsymbol{T}_\nu(\vec{\boldsymbol{l}}_\nu)=\begin{pmatrix}\boldsymbol{U}_{\nu,1}(\vec{\boldsymbol{l}}_\nu) & \boldsymbol{V}_{\nu,1}(\vec{\boldsymbol{l}}_\nu) & \boldsymbol{O} & \cdots & \boldsymbol{O}\\ \boldsymbol{O} & \boldsymbol{U}_{\nu,2}(\vec{\boldsymbol{l}}_\nu) & \boldsymbol{V}_{\nu,2}(\vec{\boldsymbol{l}}_\nu) & & \vdots\\ \vdots & & \ddots & \ddots & \boldsymbol{O}\\ \boldsymbol{O} & \cdots & \boldsymbol{O} & \boldsymbol{U}_{\nu,\nu}(\vec{\boldsymbol{l}}_\nu) & \boldsymbol{V}_{\nu,\nu}(\vec{\boldsymbol{l}}_\nu)\\ \boldsymbol{O} & \boldsymbol{O} & \cdots & \boldsymbol{O} & \boldsymbol{Q}^-\end{pmatrix}, \tag{4.33}$$

where $\boldsymbol{U}_{\nu,j}(\vec{\boldsymbol{l}}_\nu)$ and $\boldsymbol{V}_{\nu,j}(\vec{\boldsymbol{l}}_\nu)$ $(j=1,\ldots,\nu)$ are given in (4.30) and (4.31), respectively. Applying Lemma 4.1 to (4.28) and taking account of (4.22) and (4.26), we have the following theorem.

Theorem 4.4 $\widehat{\boldsymbol{B}}(t,\boldsymbol{m})$ *is given by*

$$\widehat{\boldsymbol{B}}(t,\boldsymbol{m})=\sum_{\nu=1}^{|\boldsymbol{m}|}\sum_{\vec{\boldsymbol{l}}_\nu\in\mathcal{L}_\nu(\boldsymbol{m})}\left[\prod_{\eta=1}^{\nu}h(\boldsymbol{l}_\eta)d(\boldsymbol{l}_\eta)\right]\mathrm{diag}\,(\boldsymbol{\pi})^{-1}\left[\boldsymbol{F}_\nu^-(t,\vec{\boldsymbol{l}}_\nu)\right]^{\mathrm{T}}\mathrm{diag}\,(\boldsymbol{\pi}), \tag{4.34}$$

where $h(\boldsymbol{l}_\eta)$ *and* $d(\boldsymbol{l}_\eta)$ *(*$\boldsymbol{l}_\eta \in \mathcal{Z}^+$*) are given in (4.24) and (4.25), respectively. Further* $\boldsymbol{F}_\nu^-(t, \vec{\boldsymbol{l}}_\nu)$ $(\nu = 1, 2, \ldots)$ *is given by*

$$\boldsymbol{F}_\nu^-(t, \vec{\boldsymbol{l}}_\nu) = \left[\begin{array}{cccc} \boldsymbol{J}_\nu(\vec{\boldsymbol{l}}_\nu) & \boldsymbol{O} & \cdots & \boldsymbol{O} \end{array} \right] \exp\left[\boldsymbol{T}_\nu(\vec{\boldsymbol{l}}_\nu)t\right] \left[\begin{array}{c} \boldsymbol{O} \\ \vdots \\ \boldsymbol{O} \\ \boldsymbol{I}_Q \end{array} \right], \tag{4.35}$$

where $\boldsymbol{J}_\nu(\vec{\boldsymbol{l}}_\nu)$ *and* $\boldsymbol{T}_\nu(\vec{\boldsymbol{l}}_\nu)$ *are given in (4.29) and (4.33), respectively.*

Remark 4.5 $\boldsymbol{T}_\nu(\vec{\boldsymbol{l}}_\nu)$ *in (4.33) is considered as the defective infinitesimal generator of an absorbing Markov chain. Namely,* $\boldsymbol{T}_\nu(\vec{\boldsymbol{l}}_\nu)$ *has negative diagonal elements and nonnegative off-diagonal elements, all its row sums are not positive, and at least one row sum is strictly negative. Therefore applying the uniformization technique (see Appendix A), we can readily compute* $\exp[\boldsymbol{T}_\nu(\vec{\boldsymbol{l}}_\nu)t]$. *Further since* $\exp[\boldsymbol{T}_\nu(\vec{\boldsymbol{l}}_\nu)t]$, $\boldsymbol{J}_\nu(\vec{\boldsymbol{l}}_\nu)$, $\boldsymbol{I}_Q$, $\mathrm{diag}\,(\boldsymbol{\pi})$, $\mathrm{diag}\,(\boldsymbol{\pi})^{-1}$, $h(\boldsymbol{l}_\eta)$ *and* $d(\boldsymbol{l}_\eta)$ *are all nonnegative, the computation of* $\widehat{\boldsymbol{B}}(t, \boldsymbol{m})$ *is numerically stable.*

4.4.3 Limiting formula for phase-type services

In this subsection, assuming phase-type service times, we derive an explicit and numerically feasible formula for the limit $\widehat{\boldsymbol{B}}(\boldsymbol{m})$ of $\widehat{\boldsymbol{B}}(t, \boldsymbol{m})$:

$$\widehat{\boldsymbol{B}}(\boldsymbol{m}) = \lim_{t\to\infty} \widehat{\boldsymbol{B}}(t, \boldsymbol{m}), \qquad \boldsymbol{m} \in \mathcal{Z}^+.$$

We define $\boldsymbol{F}_\nu^-(\vec{\boldsymbol{l}}_\nu)$ as

$$\boldsymbol{F}_\nu^-(\vec{\boldsymbol{l}}_\nu) = \lim_{t\to\infty} \boldsymbol{F}_\nu^-(t, \vec{\boldsymbol{l}}_\nu).$$

Theorem 4.5 $\widehat{\boldsymbol{B}}(\boldsymbol{m})$ *(*$\boldsymbol{m} \in \mathcal{Z}^+$*) is given by*

$$\widehat{\boldsymbol{B}}(\boldsymbol{m}) = \sum_{\nu=1}^{|\boldsymbol{m}|} \sum_{\vec{\boldsymbol{l}}_\nu \in \mathcal{L}_\nu(\boldsymbol{m})} \left[\prod_{\eta=1}^{\nu} h(\boldsymbol{l}_\eta) d(\boldsymbol{l}_\eta) \right] \mathrm{diag}\,(\boldsymbol{\pi})^{-1} \boldsymbol{F}_\nu^-(\vec{\boldsymbol{l}}_\nu)^{\mathrm{T}} \mathrm{diag}\,(\boldsymbol{\pi}), \tag{4.36}$$

where

$$\boldsymbol{F}_\nu^-(\vec{\boldsymbol{l}}_\nu) = \boldsymbol{J}_\nu(\vec{\boldsymbol{l}}_\nu) \left[-\boldsymbol{U}_{\nu,1}(\vec{\boldsymbol{l}}_\nu)\right]^{-1} \boldsymbol{V}_{\nu,1}(\vec{\boldsymbol{l}}_\nu) \cdot \cdots \cdot \left[-\boldsymbol{U}_{\nu,\nu}(\vec{\boldsymbol{l}}_\nu)\right]^{-1} \boldsymbol{V}_{\nu,\nu}(\vec{\boldsymbol{l}}_\nu) \boldsymbol{e}\boldsymbol{\pi}. \tag{4.37}$$

Note that $\boldsymbol{J}_\nu(\vec{\boldsymbol{l}}_\nu)$, $\boldsymbol{U}_{\nu,j}(\vec{\boldsymbol{l}}_\nu)$ *and* $\boldsymbol{V}_{\nu,j}(\vec{\boldsymbol{l}}_\nu)$ $(j = 1, \ldots, \nu)$ *are given in (4.29), (4.30) and (4.31), respectively.*

Proof. Taking the limit of both sides of (4.34) as $t \to \infty$ yields (4.36). In what follows, we show (4.37). We define a submatrix $\boldsymbol{T}'_\nu(\vec{\boldsymbol{l}}_\nu)$ of $\boldsymbol{T}_\nu(\vec{\boldsymbol{l}}_\nu)$ as

$$\boldsymbol{T}'_\nu(\vec{\boldsymbol{l}}_\nu) = \left[\begin{array}{ccccc} \boldsymbol{U}_{\nu,1}(\vec{\boldsymbol{l}}_\nu) & \boldsymbol{V}_{\nu,1}(\vec{\boldsymbol{l}}_\nu) & \boldsymbol{O} & \cdots & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{U}_{\nu,2}(\vec{\boldsymbol{l}}_\nu) & \boldsymbol{V}_{\nu,2}(\vec{\boldsymbol{l}}_\nu) & & \vdots \\ \vdots & & \ddots & \ddots & \boldsymbol{O} \\ \boldsymbol{O} & \cdots & \boldsymbol{O} & \boldsymbol{U}_{\nu,\nu-1}(\vec{\boldsymbol{l}}_\nu) & \boldsymbol{V}_{\nu,\nu-1}(\vec{\boldsymbol{l}}_\nu) \\ \boldsymbol{O} & \boldsymbol{O} & \cdots & \boldsymbol{O} & \boldsymbol{U}_{\nu,\nu}(\vec{\boldsymbol{l}}_\nu) \end{array} \right]. \tag{4.38}$$

Then $\boldsymbol{F}_{\nu}^{-}(t,\vec{\boldsymbol{l}}_{\nu})$ in (4.35) can be rewritten to be

$$\boldsymbol{F}_{\nu}^{-}(t,\vec{\boldsymbol{l}}_{\nu}) = \left[\begin{array}{cccc|c} \boldsymbol{J}_{\nu}(\vec{\boldsymbol{l}}_{\nu}) & \boldsymbol{O} & \cdots & \boldsymbol{O} & \boldsymbol{O} \end{array}\right] \exp\left(\left[\begin{array}{c|c} \boldsymbol{T}'_{\nu}(\vec{\boldsymbol{l}}_{\nu}) & \begin{bmatrix} \boldsymbol{O} \\ \vdots \\ \boldsymbol{O} \\ \boldsymbol{V}_{\nu,\nu}(\vec{\boldsymbol{l}}_{\nu}) \end{bmatrix} \\ \hline \boldsymbol{O} & \boldsymbol{Q}^{-} \end{array}\right] t \right) \left[\begin{array}{c} \boldsymbol{O} \\ \vdots \\ \boldsymbol{O} \\ \hline \boldsymbol{I}_{Q} \end{array}\right].$$

Applying Lemma 4.1 to the above equation yields

$$\boldsymbol{F}_{\nu}^{-}(t,\vec{\boldsymbol{l}}_{\nu}) = \left[\begin{array}{cccc} \boldsymbol{J}_{\nu}(\vec{\boldsymbol{l}}_{\nu}) & \boldsymbol{O} & \cdots & \boldsymbol{O} \end{array}\right] \boldsymbol{\Theta}_{\nu}(t,\vec{\boldsymbol{l}}_{\nu}), \tag{4.39}$$

where

$$\boldsymbol{\Theta}_{\nu}(t,\vec{\boldsymbol{l}}_{\nu}) = \int_{0}^{t} du \exp\left[\boldsymbol{T}'_{\nu}(\vec{\boldsymbol{l}}_{\nu})u\right] \begin{bmatrix} \boldsymbol{O} \\ \vdots \\ \boldsymbol{O} \\ \boldsymbol{V}_{\nu,\nu}(\vec{\boldsymbol{l}}_{\nu}) \end{bmatrix} \exp[\boldsymbol{Q}^{-}(t-u)]. \tag{4.40}$$

Note that $\boldsymbol{T}'_{\nu}(\vec{\boldsymbol{l}}_{\nu})$ in (4.38) is regarded as the defective infinitesimal generator of an absorbing Markov chain in transient states, and therefore all row sums of $\boldsymbol{T}'_{\nu}(\vec{\boldsymbol{l}}_{\nu})$ are strictly negative. Thus we have

$$\int_{0}^{\infty} du \exp\left[\boldsymbol{T}'_{\nu}(\vec{\boldsymbol{l}}_{\nu})u\right] = \left[-\boldsymbol{T}'_{\nu}(\vec{\boldsymbol{l}}_{\nu})\right]^{-1}. \tag{4.41}$$

Further, since $\boldsymbol{Q}^{-}$ is the infinitesimal generator of an irreducible Markov chain with the stationary probability vector $\boldsymbol{\pi}$ (see Remark 4.4), we have

$$\lim_{t\to\infty} \exp(\boldsymbol{Q}^{-}t) = \boldsymbol{e}\boldsymbol{\pi}. \tag{4.42}$$

It then follows from (4.40), (4.41) and (4.42) that

$$\begin{aligned} \boldsymbol{\Theta}_{\nu}(\vec{\boldsymbol{l}}_{\nu}) &= \lim_{t\to\infty} \boldsymbol{\Theta}_{\nu}(t,\vec{\boldsymbol{l}}_{\nu}) \\ &= \int_{0}^{\infty} du \exp\left[\boldsymbol{T}'_{\nu}(\vec{\boldsymbol{l}}_{\nu})u\right] \begin{bmatrix} \boldsymbol{O} \\ \vdots \\ \boldsymbol{O} \\ \boldsymbol{V}_{\nu,\nu}(\vec{\boldsymbol{l}}_{\nu}) \end{bmatrix} \exp(-\boldsymbol{Q}^{-}u) \cdot \lim_{t\to\infty} \exp(\boldsymbol{Q}^{-}t) \\ &= \int_{0}^{\infty} du \exp\left[\boldsymbol{T}'_{\nu}(\vec{\boldsymbol{l}}_{\nu})u\right] \cdot \begin{bmatrix} \boldsymbol{O} \\ \vdots \\ \boldsymbol{O} \\ \boldsymbol{V}_{\nu,\nu}(\vec{\boldsymbol{l}}_{\nu}) \end{bmatrix} \boldsymbol{e}\boldsymbol{\pi} \\ &= \left[-\boldsymbol{T}'_{\nu}(\vec{\boldsymbol{l}}_{\nu})\right]^{-1} \begin{bmatrix} \boldsymbol{O} \\ \vdots \\ \boldsymbol{O} \\ \boldsymbol{V}_{\nu,\nu}(\vec{\boldsymbol{l}}_{\nu}) \end{bmatrix} \boldsymbol{e}\boldsymbol{\pi}, \end{aligned} \tag{4.43}$$

where we use $\exp(\boldsymbol{Q}^- t)\boldsymbol{e} = \boldsymbol{e}$ $(t \geq 0)$. Finally, from (4.39) and (4.43), we obtain

$$\begin{aligned}
\boldsymbol{F}_\nu^-(\vec{\boldsymbol{l}}_\nu) &= \left[\begin{array}{cccc} \boldsymbol{J}_\nu(\vec{\boldsymbol{l}}_\nu) & \boldsymbol{O} & \cdots & \boldsymbol{O} \end{array} \right] \boldsymbol{\Theta}_\nu(\vec{\boldsymbol{l}}_\nu) \\
&= \left[\begin{array}{cccc} \boldsymbol{J}_\nu(\vec{\boldsymbol{l}}_\nu) & \boldsymbol{O} & \cdots & \boldsymbol{O} \end{array} \right] \left[-\boldsymbol{T}'_\nu(\vec{\boldsymbol{l}}_\nu) \right]^{-1} \left[\begin{array}{c} \boldsymbol{O} \\ \vdots \\ \boldsymbol{O} \\ \boldsymbol{V}_{\nu,\nu}(\vec{\boldsymbol{l}}_\nu) \end{array} \right] \boldsymbol{e}\boldsymbol{\pi},
\end{aligned}$$

from which Theorem 4.5 follows. ■

4.5 Numerical Examples

In this section, we show some numerical examples based on Theorems 4.4 and 4.5, and discuss the impact of system parameters on the mean and variance of the number of customers in the system under the assumption that the arrival process is stationary. Note that when the arrival process is stationary, the mean number $\mathrm{E}[N_k(t)]$ $(k \in \mathcal{K})$ of class k customers at time t is given by

$$\mathrm{E}[N_k(t)] = \boldsymbol{\pi}\widehat{\boldsymbol{B}}(t, \boldsymbol{e}_k)\boldsymbol{e} = \lambda_k \int_0^t du\, \overline{H}_k(u),$$

where λ_k is the arrival rate of class k customers and $\boldsymbol{e}_k$ is given in (1.4). Further the limit $\mathrm{E}[N_k]$ of $\mathrm{E}[N_k(t)]$ is given by

$$\mathrm{E}[N_k] = \boldsymbol{\pi}\widehat{\boldsymbol{B}}(\boldsymbol{e}_k)\boldsymbol{e} = \lambda_k h_k,$$

where h_k denotes the mean service time of class k customers. Therefore the mean $\mathrm{E}[N(t)]$ of the total number of customers at time t and its limit $E[N]$ are given by

$$\mathrm{E}[N(t)] = \sum_{k\in\mathcal{K}} \lambda_k \int_0^t du\, \overline{H}_k(u), \qquad \mathrm{E}[N] = \sum_{k\in\mathcal{K}} \lambda_k h_k,$$

respectively.

On the other hand, the variance $\mathrm{Var}[N_k(t)]$ $(k \in \mathcal{K})$ of the number of class k customers at time t and its limit $\mathrm{Var}[N_k]$ are given by

$$\begin{aligned}
\mathrm{Var}[N_k(t)] &= 2\boldsymbol{\pi}\widehat{\boldsymbol{B}}(t, 2\boldsymbol{e}_k)\boldsymbol{e} + \mathrm{E}[N_k(t)] - \mathrm{E}[N_k(t)]^2, \\
\mathrm{Var}[N_k] &= 2\boldsymbol{\pi}\widehat{\boldsymbol{B}}(2\boldsymbol{e}_k)\boldsymbol{e} + \mathrm{E}[N_k] - \mathrm{E}[N_k]^2,
\end{aligned}$$

respectively. Further the covariance $\mathrm{Cov}[N_k(t), N_{k'}(t)]$ $(k, k' \in \mathcal{K})$ of the numbers of class k and k' customers at time t and its limit $\mathrm{Cov}[N_k, N_{k'}]$ are given by

$$\begin{aligned}
\mathrm{Cov}[N_k(t), N_{k'}(t)] &= \boldsymbol{\pi}\widehat{\boldsymbol{B}}(t, \boldsymbol{e}_k + \boldsymbol{e}_{k'})\boldsymbol{e} - \mathrm{E}[N_k(t)]\mathrm{E}[N_{k'}(t)], \\
\mathrm{Cov}[N_k, N_{k'}] &= \boldsymbol{\pi}\widehat{\boldsymbol{B}}(\boldsymbol{e}_k + \boldsymbol{e}_{k'})\boldsymbol{e} - \mathrm{E}[N_k]\mathrm{E}[N_{k'}],
\end{aligned}$$

respectively. The variance $\mathrm{Var}[N(t)]$ of the total number of customers at time t and its limit $\mathrm{Var}[N]$ are given in terms of the marginal variances and covariances:

$$\begin{aligned}\mathrm{Var}[N(t)] &= \sum_{k\in\mathcal{K}} \mathrm{Var}[N_k(t)] + 2\sum_{k=1}^{K-1}\sum_{k'=k+1}^{K} \mathrm{Cov}[N_k(t), N_{k'}(t)],\\ \mathrm{Var}[N] &= \sum_{k\in\mathcal{K}} \mathrm{Var}[N_k] + 2\sum_{k=1}^{K-1}\sum_{k'=k+1}^{K} \mathrm{Cov}[N_k, N_{k'}],\end{aligned}$$

respectively.

4.5.1 Impact of service time distribution on $\mathrm{Var}[N]$

We first show the impact of the service time distribution on the limiting variance of the number of customers. To do so, we consider the following batch MMPP. There is only one class, i.e., $\mathcal{K} = \{1\}$, and

$$\boldsymbol{C} = \begin{bmatrix} -9 & 1 \\ 1 & -3 \end{bmatrix}, \qquad \boldsymbol{D} = \begin{bmatrix} 8 & 0 \\ 0 & 2 \end{bmatrix},$$

$$\boldsymbol{D}(n) = g(n)\boldsymbol{D}, \quad g(n) = \begin{cases} 1/2, & \text{if } n = 1, \\ 1/2, & \text{if } n = 3, \\ 0, & \text{otherwise.} \end{cases}$$

Note that the mean arrival rate λ_1 is equal to 10. As for the service time distribution, we fix the mean service time to be one (i.e., $\mathrm{E}[N] = 10$), and consider the following three distributions:
[k-stage Erlang distribution]

$$\overline{H}_1(t) = \Big[1 \underbrace{0 \ \ldots \ 0}_{k-1}\Big] \exp\left[\begin{pmatrix} -k & k & 0 & \cdots & 0 \\ 0 & -k & k & \cdots & 0 \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & -k & k \\ 0 & 0 & \cdots & 0 & -k \end{pmatrix} t\right] \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 1 \end{bmatrix} \quad k = 2, 3, \ldots, \tag{4.44}$$

[Exponential distribution]

$$\overline{H}_1(t) = e^{-t}, \tag{4.45}$$

[Two-state balanced hyper-exponential distribution]

$$\overline{H}_1(t) = \Big[p \quad 1-p\Big] \exp\left[\begin{pmatrix} -2p & 0 \\ 0 & -2(1-p) \end{pmatrix} t\right] \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad 0 < p < 0.5. \tag{4.46}$$

Let C_v denote the coefficient of variation of the service time distribution. Note that $C_v = 1/\sqrt{k}$ $(k = 2, 3, \ldots)$ for k-stage Erlang distribution, $C_v = 1$ for exponential distribution and $C_v = \sqrt{1/\{2p(1-p)\} - 1}$ $(0 < p < 0.5)$ for two-state balanced hyper-exponential distribution. Figure 4.1 plots the limiting variance $\mathrm{Var}[N]$ of the number of customers in the system as a function of C_v. We observe that $\mathrm{Var}[N]$ decreases as C_v increases.

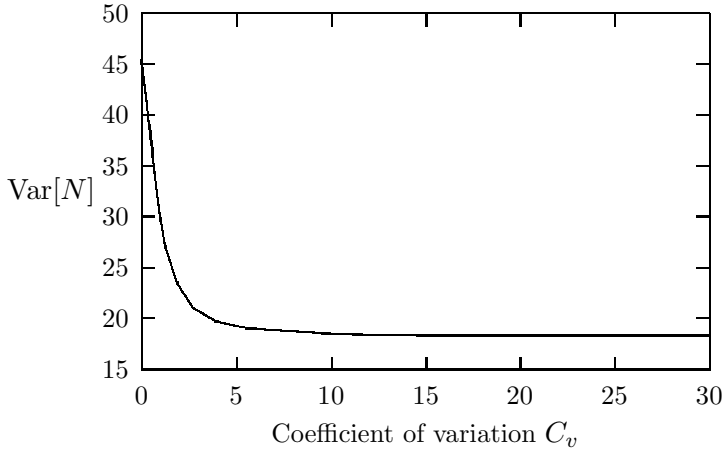

Figure 4.1: Limiting variance of the number of customers.

Remark 4.6 *For the $M(t)/G/\infty$ queue with a sinusoidal arrival rate, Eick et al. [Eick93] show that as C_v increases, the deviation of the arrival process has a less impact on the mean number of customers in the system. Thus our observation coincides with that in [Eick93].*

4.5.2 Impact of arrival process

We consider the impact of the correlation in the arrival process on the variance of the number of customers. For this purpose, we assume that batches arrive according to the following two-state Markov modulated Poisson process:

$$\boldsymbol{C} = \begin{bmatrix} -\lambda_1 - c & c \\ c & -\lambda_2 - c \end{bmatrix}, \quad \boldsymbol{D} = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}, \quad c > 0,$$

$$\boldsymbol{D}(n) = g(n)\boldsymbol{D}, \qquad n = 1, 2, \ldots,$$

where $\{g(n); n = 1, 2, \ldots\}$ denotes the batch size distribution, i.e., $\sum_{n=1}^{\infty} g(n) = 1$. Let G denote a random variable representing the batch size. We define $A(t, \tau)$ $(0 \le t < \tau)$ as the number of customers arriving in time interval $(t, \tau]$. We then have

$$\mathrm{Cov}[A(t, t+s), A(t+s+\tau, t+2s+\tau)] = \frac{\mathrm{E}[G]^2 \left(\dfrac{\lambda_1 - \lambda_2}{2}\right)^2 (1 - e^{-2cs})^2}{(2c)^2} e^{-2c\tau}.$$

Note that the time-correlation of the arrival process increases with the mean sojourn time $1/c$ in each state of the underlying Markov chain. As for the service time distribution, we consider three cases: the two-stage Erlang distribution in (4.44) with $k = 2$, the exponential distribution in (4.45), and the two-state hyper-exponential distribution in (4.46) with $p = 0.25$.

Figure 4.2 plots the limiting variance Var[N] of the number of customers in the system as a function of $1/2c$, where $\lambda_1 = 8$, $\lambda_2 = 2$ and

$$g(n) = \begin{cases} 1/2, & \text{if } n = 1, \\ 1/2, & \text{if } n = 3, \\ 0, & \text{otherwise.} \end{cases}$$

We observe that Var[N] increases with the correlation in arrivals.

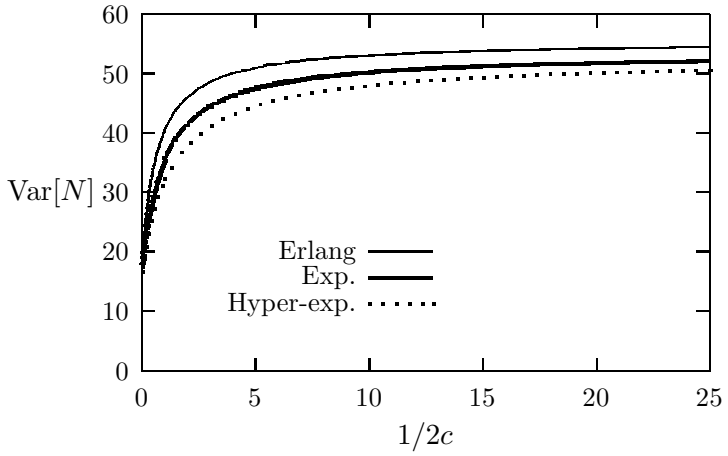

Figure 4.2: Limiting variance of the number of customers.

Next we consider the time-dependent variance Var[$N(t)$] with two stationary arrival streams. We fix the marginal characteristics of each arrival stream, which is represented by

$$\begin{aligned} \boldsymbol{C} &= \begin{bmatrix} -6 & 1 \\ 1 & -1 \end{bmatrix}, \qquad \boldsymbol{D} = \begin{bmatrix} 5 & 0 \\ 0 & 0 \end{bmatrix}, \\ g(n) &= \begin{cases} 1/2, & \text{if } n = 1, \\ 1/2, & \text{if } n = 3, \\ 0, & \text{otherwise,} \end{cases} \qquad \boldsymbol{D}(n) = g(n)\boldsymbol{D}, \\ \overline{H}(t) &= \begin{bmatrix} 0.5 & 0.5 \end{bmatrix} \exp\left[\begin{pmatrix} -0.75 & 0 \\ 0 & -1.5 \end{pmatrix} t \right] \begin{bmatrix} 1 \\ 1 \end{bmatrix}. \end{aligned}$$

where $\overline{H}(t)$ denotes the complementary distribution of service times. Under this restriction, we consider the following three cases.

Case N: Negatively correlated arrival streams.

Two arrival streams are negatively correlated. Namely, the two arrival streams are governed by a single two-state underlying Markov chain and when it is in state k ($k = 1, 2$), only class k customers can arrive. More precisely, the overall arrival process is characterized by

$$\boldsymbol{C} = \begin{bmatrix} -6 & 1 \\ 1 & -6 \end{bmatrix}, \qquad \boldsymbol{D}_1 = \begin{bmatrix} 5 & 0 \\ 0 & 0 \end{bmatrix}, \qquad \boldsymbol{D}_2 = \begin{bmatrix} 0 & 0 \\ 0 & 5 \end{bmatrix},$$

$$
\begin{aligned}
g(n) &= \begin{cases} 1/2, & \text{if } n = 1, \\ 1/2, & \text{if } n = 3, \\ 0, & \text{otherwise.} \end{cases} \\
\boldsymbol{D}(n_1, n_2) &= \begin{cases} g(n_1)\boldsymbol{D}_1, & \text{if } n_1 \geq 1 \text{ and } n_2 = 0, \\ g(n_2)\boldsymbol{D}_2, & \text{if } n_1 = 0 \text{ and } n_2 \geq 1, \\ \boldsymbol{O} & \text{otherwise,} \end{cases} \\
\overline{H}_k(t) &= \begin{bmatrix} 0.5 & 0.5 \end{bmatrix} \exp\left[\begin{pmatrix} -0.75 & 0 \\ 0 & -1.5 \end{pmatrix} t \right] \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \qquad k = 1, 2.
\end{aligned}
$$

Case I: Superposition of two independent arrival streams.

Two arrival streams are independent of each other. Thus the overall arrival process is characterized by

$$
\begin{aligned}
\boldsymbol{C} &= \begin{bmatrix} -6 & 1 \\ 1 & -1 \end{bmatrix} \oplus \begin{bmatrix} -6 & 1 \\ 1 & -1 \end{bmatrix} = \begin{bmatrix} -12 & 1 & 1 & 0 \\ 1 & -7 & 0 & 1 \\ 1 & 0 & -7 & 1 \\ 0 & 1 & 1 & -2 \end{bmatrix}, \\
\boldsymbol{D}_1 &= \begin{bmatrix} 5 & 0 \\ 0 & 0 \end{bmatrix} \otimes \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 5 & 0 & 0 & 0 \\ 0 & 5 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \\
\boldsymbol{D}_2 &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \otimes \begin{bmatrix} 5 & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 5 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \\
g(n) &= \begin{cases} 1/2, & \text{if } n = 1, \\ 1/2, & \text{if } n = 3, \\ 0, & \text{otherwise,} \end{cases} \qquad \boldsymbol{D}(n_1, n_2) = \begin{cases} g(n_1)\boldsymbol{D}_1, & \text{if } n_1 \geq 1 \text{ and } n_2 = 0, \\ g(n_2)\boldsymbol{D}_2, & \text{if } n_1 = 0 \text{ and } n_2 \geq 1, \\ \boldsymbol{O}, & \text{otherwise,} \end{cases} \\
\overline{H}_k(t) &= \begin{bmatrix} 0.5 & 0.5 \end{bmatrix} \exp\left[\begin{pmatrix} -0.75 & 0 \\ 0 & -1.5 \end{pmatrix} t \right] \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \qquad k = 1, 2.
\end{aligned}
$$

Case P: Positively correlated arrival streams.

Two arrival streams are positively correlated. Namely, arrivals from both streams occur simultaneously with probability one. More precisely, the overall arrival process is characterized by

$$
\begin{aligned}
\boldsymbol{C} &= \begin{bmatrix} -6 & 1 \\ 1 & -1 \end{bmatrix}, \qquad \boldsymbol{D} = \begin{bmatrix} 5 & 0 \\ 0 & 0 \end{bmatrix}, \\
g(n) &= \begin{cases} 1/2, & \text{if } n = 1, \\ 1/2, & \text{if } n = 3, \\ 0, & \text{otherwise,} \end{cases} \qquad \boldsymbol{D}(n_1, n_2) = g(n_1)g(n_2)\boldsymbol{D},
\end{aligned}
$$

$$\overline{H}_k(t) = \left[\begin{array}{cc} 0.5 & 0.5 \end{array}\right] \exp\left[\left(\begin{array}{cc} -0.75 & 0 \\ 0 & -1.5 \end{array}\right) t\right] \left[\begin{array}{c} 1 \\ 1 \end{array}\right], \quad k = 1, 2.$$

Because the marginal characteristics of each stream in these three cases are the same, the time-dependent and limiting marginal distributions of the numbers of customers in respective classes are identical among these three cases. However, the covariances of the numbers of customers in respective classes are different. Figure 4.3 shows the time-dependent covariance $\mathrm{Cov}[N_1(t), N_2(t)]$ of the numbers of customers in respective classes as a function of t.

We observe that positive (resp. negative) correlation between two arrival streams leads to positive (resp. negative) covariance at any time t. Thus the positive (resp. negative) correlation in arrivals leads to large (resp. small) variance, compared with the superposition of independent streams, as shown in Figure 4.4.

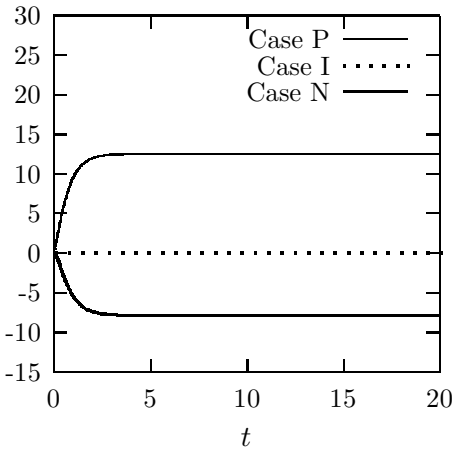

Figure 4.3: Time-dependent covariance $\mathrm{Cov}[N(t)]$ of the number of customers.

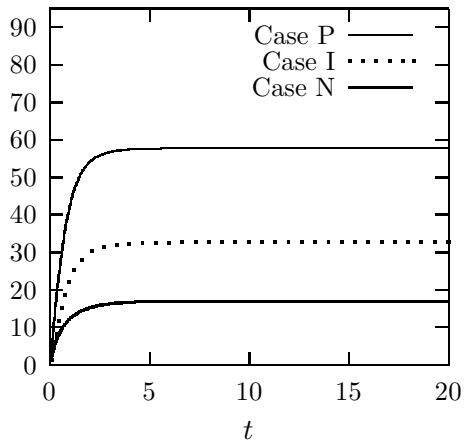

Figure 4.4: Time-dependent variance $\mathrm{Var}[N(t)]$ of the number of customers

4.5.3 Impact of correlation in service time sequence

Finally we consider the $M^X/SM/\infty$ queue to investigate the correlation in the service time sequence. We assume that batches arrive according to a Poisson process with rate 5, and batch sizes are i.i.d. according to the following two-points distribution: The size of the batch is equal to one with probability 1/2 and to 3 with probability 1/2. Let $H(t, n)$ denote the service time distribution of customers in the nth batch. We assume that the sequence $\{H(t, n);\ n = 1, 2, \ldots\}$ of service time distributions constitutes a semi-Markov process whose kernel $\boldsymbol{H}(t)$ is given by

$$\boldsymbol{H}(t) = \left(\begin{array}{cc} pH_1(t) & (1-p)H_2(t) \\ (1-p)H_1(t) & pH_2(t) \end{array}\right), \qquad 0 \le p < 1,$$

where $H_1(t)$ and $H_2(t)$ $(t \ge 0)$ are distribution functions. Note that the sequence has negative correlation for $0 \le p < 1/2$, it becomes i.i.d. for $p = 1/2$, and it has positive correlation for

$1/2 < p < 1$. We regard customers whose service times follow a distribution function $H_k(t)$ $(k = 1, 2)$ as class k customers. Formally this $M^X/SM/\infty$ queue is characterized by

$$\begin{aligned}
\boldsymbol{C} &= \begin{bmatrix} -5 & 0 \\ 0 & -5 \end{bmatrix}, \qquad \boldsymbol{D}_1 = \begin{bmatrix} 5p & 0 \\ 5(1-p) & 0 \end{bmatrix}, \qquad \boldsymbol{D}_2 = \begin{bmatrix} 0 & 5(1-p) \\ 0 & 5p \end{bmatrix}, \\
g(n) &= \begin{cases} 1/2, & \text{if } n = 1, \\ 1/2, & \text{if } n = 3, \\ 0, & \text{otherwise}, \end{cases} \\
\boldsymbol{D}(n_1, n_2) &= \begin{cases} g(n_1)\boldsymbol{D}_1, & \text{if } n_1 \geq 1 \text{ and } n_2 = 0, \\ g(n_2)\boldsymbol{D}_2, & \text{if } n_1 = 0 \text{ and } n_2 \geq 1, \\ \boldsymbol{O}, & \text{otherwise}, \end{cases}
\end{aligned}$$

where $0 \leq p < 1$. Note that $\lambda_1 = \lambda_2 = 5$ in this formulation. As for service time distributions $H_k(t)$ $(k = 1, 2)$, we fix $\lambda_1 h_1 + \lambda_2 h_2 = 10$ (i.e., $\mathrm{E}[N] = 10$), and consider the following three cases: Case 1:

$$H_1(t) = 1 - 0.25e^{-0.5t} - 0.75e^{-1.5t}, \qquad H_2(t) = 1 - e^{-t}.$$

Case 2:

$$H_1(t) = 1 - e^{-2t}, \qquad H_2(t) = 1 - e^{-\frac{2}{3}t}.$$

Case 3:

$$H_1(t) = 1 - e^{-5t}, \qquad H_2(t) = 1 - e^{-\frac{5}{9}t}.$$

In Case 1, the mean service time of each class is equal to one. Thus the autocovariance of the service time sequence is equal to zero for all p. On the other hand, in Case 2 and Case 3, the mean service times are different, so that the autocovariance of the service time sequence is positive for $p > 1/2$ and negative for $p < 1/2$. Note also that the difference between mean service times in Case 2 is equal to one, while that in Case 3 is equal to $8/5$. Thus the absolute value of the autocovariance in Case 3 is larger than in Case 2 for $p \neq 0$.

Figures 4.5, 4.6 and 4.7 plot the marginal variances $\mathrm{Var}[N_1]$ and $\mathrm{Var}[N_2]$, the covariance $\mathrm{Cov}[N_1, N_2]$ and the variance $\mathrm{Var}[N]$ of the total number of customers as functions of parameter p in Cases 1, 2 and 3, respectively. We observe that when the autocovariance of the service time sequence is equal to zero (Case 1), $\mathrm{Var}[N]$ is almost insensitive to parameter p. On the other hand, as shown in Figures 4.6 and 4.7, $\mathrm{Var}[N]$ becomes larger with the correlation in the service time sequence when the autocovariance is not equal to zero. Further, comparing Figure 4.7 with Figure 4.6, we observe that a stronger correlation leads to a larger value of $\mathrm{Var}[N]$.

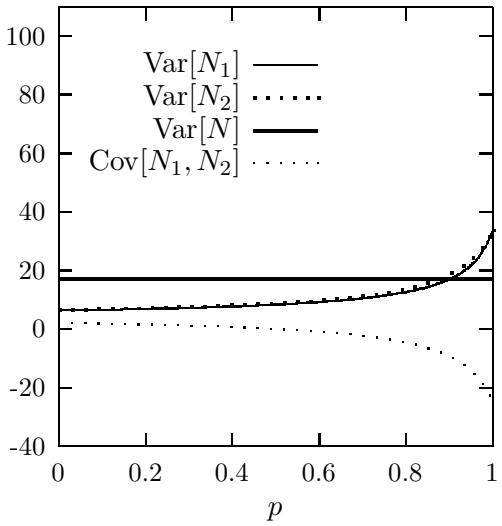

Figure 4.5: Variance and covariance in Case 1.

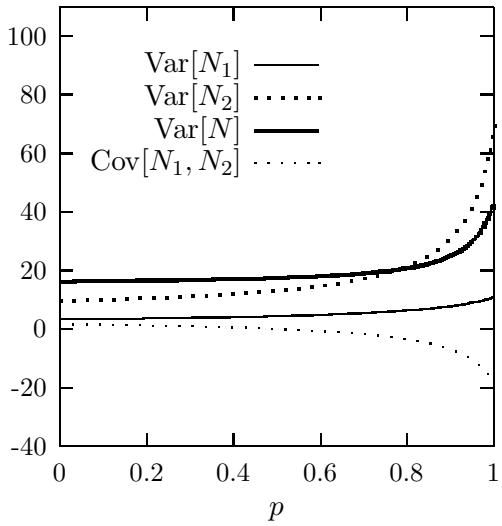

Figure 4.6: Variance and covariance in Case 2.

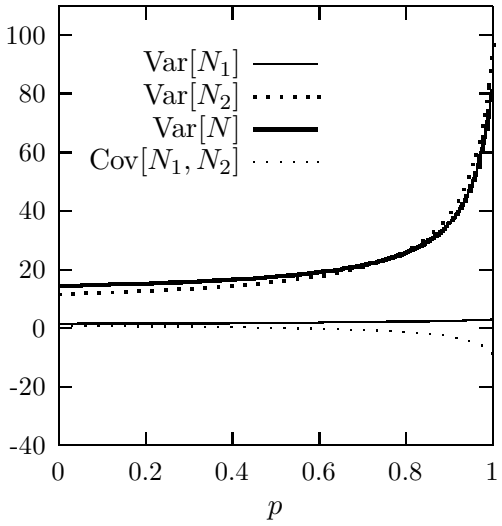

Figure 4.7: Variance and covariance in Case 3.

Chapter 5

Processor-Sharing Queue

5.1 Introduction

Processor-sharing queues are important models in computer engineering. Under the processor-sharing service discipline, the server is shared equally by all customers in the system. To put it more concretely, when n customers are present in the system, each customer receives service at rate $1/n$.

Many researchers have studied processor-sharing queues for a few decades. For the stationary M/M/1 processor-sharing queue (M/M/1-PS), Coffman, Muntz and Trotter [Coff70] derived the LST of the sojourn time distribution. By inverting this LST, Morrison [Morr85] obtained an integral representation for the complementary distribution of the sojourn time, which was also derived by Guillemin and Boyer with the spectral theory [Guil01]. Further Sengupta and Jagerman [Seng85] studied moments of the sojourn time conditioned on the number of customers found by arriving customers.

For the stationary GI/M/1-PS queue, the first two moments and the LST of the sojourn time were derived in Ramaswami [Rama84] and Jagerman and Sengupta [Jage91], respectively. Also, for the stationary M/G/1-PS queue, Ott [Ott84] and Yashkov [Yash83] derived the LST and first two moments of the sojourn time conditioned on the service time of an arriving customer. Besides, Grishechkin [Gris94] and Sengupta [Seng92] studied GI/G/1-PS queues.

To the best of our knowledge, all of previous studies on processor-sharing queues assume that inter-arrival times are independent and identically distributed (i.i.d.). Further there are no works focusing on the computation of the sojourn time distribution. Motivated by these facts, we consider the sojourn time distribution in a stationary processor-sharing queue with BMAP arrivals (see subsection 1.1.1) and exponential service times, i.e., a BMAP/M/1-PS queue.

The main contribution of this chapter is to provide a recursive formula for the stationary sojourn time distribution in a BMAP/M/1-PS queue. The derivation is very simple and the result is easy to be interpreted probabilistically. Moreover, in the single arrival case (i.e., MAP/M/1-PS queue), we discuss numerical procedure and accuracy guarantee.

The rest of this chapter is divided into four sections. In section 2, the mathematical model and some known results are described. In section 3, we derive a recursive formula for the complementary

distribution of the sojourn time in a BMAP/M/1-PS queue. In section 4, for a MAP/M/1-PS queue, we discuss the numerical procedure to control the absolute error in numerical results and show some numerical examples. Finally, in section 5, we provide concluding remarks on how to utilize the results in this chapter to obtain the waiting time distribution in a random order service MAP/M/1 queue.

5.2 Model and Known Results

This chapter considers a stationary BMAP/M/1 processor-sharing (BMAP/M/1-PS) queue. Service times are i.i.d. according to an exponential distribution with finite mean μ_i^{-1}. Customer arrivals follow a BMAP $(\boldsymbol{C}, \boldsymbol{D}(n))$ (see subsection 1.1.1). Broadly speaking, customer arrivals occur as follows. There is an irreducible Markov chain with finite state space $\mathcal{M} = \{1, \ldots, M\}$, which governs the BMAP. When a state transition driven by $M \times M$ matrix $\boldsymbol{D}(n)$ happens, n customers arrive simultaneously. On the other hand, when a state transition driven by $M \times M$ matrix $\boldsymbol{C}$ happens, no customers arrive.

Note that the infinitesimal generator of the underlying Markov chain is given by $\boldsymbol{C} + \boldsymbol{D}$, where $\boldsymbol{D} = \sum_{n=1}^{\infty} \boldsymbol{D}(n)$. Note also that the mean arrival rate λ is given by

$$\lambda = \boldsymbol{\pi} \sum_{n=1}^{\infty} n\boldsymbol{D}(n)\boldsymbol{e},$$

where $\boldsymbol{\pi}$ denotes the invariant probability vector for $(\boldsymbol{C} + \boldsymbol{D})$. To avoid trivialities, we assume $\boldsymbol{D} \neq \boldsymbol{O}$. Thus $\lambda > 0$. Let ρ denote the utilization factor, i.e., $\rho = \lambda\mu^{-1}$. In the remainder of this chapter, we assume that $\rho < 1$ and the system is in steady state.

We make a comment on the stationary queue length distribution in the BMAP/M/1 queue. Let N and S denote the number of customers in the system and the state of the underlying Markov chain, respectively, in steady state. We then define $\boldsymbol{\pi}_n$ $(n = 0, 1, \ldots)$ as a $1 \times M$ vector whose jth $(j \in \mathcal{M})$ element represents $\Pr[N = n, S = j]$. Owing to i.i.d. exponential services, the stationary queue length distribution in the BMAP/M/1-PS queue is identical to that in the BMAP/M/1-FCFS. Thus, $\boldsymbol{\pi}_n$'s $(n = 0, 1, \ldots)$ can be computed by the M/G/1 paradigm (see Appendix B.1).

5.3 Sojourn Time Distribution

In this section, we consider the sojourn time distribution in steady state. We denote, by customer C_n $(n \geq 0)$, a randomly chosen customer who is a member of a batch that increases the queue length to $n + 1$ on arrival. Let W_{C_n} denote a random variable representing the sojourn time of customer C_n. Let S_{C_n} denote a random variable representing the state of the underlying Markov chain immediately after the arrival of customer C_n. We then define $\overline{\boldsymbol{w}}_n(x)$ $(x \geq 0,\ n = 0, 1, \ldots)$ as an $M \times 1$ vector whose jth $(j \in \mathcal{M})$ element represents $\Pr[W_{\mathrm{C}_n} > x \mid S_{\mathrm{C}_n} = j]$.

We now consider the sojourn time of a customer C_n arriving at $t = 0$. Conditioning the event at $t = \Delta x$, we readily obtain

$$\overline{\boldsymbol{w}}_n(x + \Delta x) = \frac{n\mu}{n+1}\Delta x\overline{\boldsymbol{w}}_{n-1}(x) + \{\boldsymbol{I} + (\boldsymbol{C} - \mu\boldsymbol{I})\Delta x\}\overline{\boldsymbol{w}}_n(x)$$

$$+ \sum_{m=1}^{\infty} \boldsymbol{D}(m)\Delta x \overline{\boldsymbol{w}}_{n+m}(x) + o(\Delta x), \qquad n = 0, 1, \ldots, \tag{5.1}$$

where $\overline{\boldsymbol{w}}_{-1}(x) = \boldsymbol{0}$ for all $x \geq 0$. It then follows from (5.1) that

$$\frac{d}{dx}\overline{\boldsymbol{w}}_n(x) = \frac{n\mu}{n+1}\overline{\boldsymbol{w}}_{n-1}(x) + (\boldsymbol{C} - \mu\boldsymbol{I})\overline{\boldsymbol{w}}_n(x) + \sum_{m=1}^{\infty} \boldsymbol{D}(m)\overline{\boldsymbol{w}}_{n+m}(x), \quad n = 0, 1, \ldots, \tag{5.2}$$

where $\overline{\boldsymbol{w}}_{-1}(x) = \boldsymbol{0}$ for all $x \geq 0$. Note that the differential-difference equation (5.2) is an extension of equation (9.10) on page 111 in [Asmu03].

We now define $\overline{\boldsymbol{w}}(x)$ and $\boldsymbol{T}$ as

$$\overline{\boldsymbol{w}}(x) = \begin{bmatrix} \overline{\boldsymbol{w}}_0(x) \\ \overline{\boldsymbol{w}}_1(x) \\ \overline{\boldsymbol{w}}_2(x) \\ \vdots \end{bmatrix}, \tag{5.3}$$

$$\boldsymbol{T} = \begin{bmatrix} \boldsymbol{C} - \mu\boldsymbol{I} & \boldsymbol{D}(1) & \boldsymbol{D}(2) & \boldsymbol{D}(3) & \boldsymbol{D}(4) & \cdots \\ \frac{\mu\boldsymbol{I}}{2} & \boldsymbol{C} - \mu\boldsymbol{I} & \boldsymbol{D}(1) & \boldsymbol{D}(2) & \boldsymbol{D}(3) & \cdots \\ \boldsymbol{O} & \frac{2\mu\boldsymbol{I}}{3} & \boldsymbol{C} - \mu\boldsymbol{I} & \boldsymbol{D}(1) & \boldsymbol{D}(2) & \cdots \\ \boldsymbol{O} & \boldsymbol{O} & \frac{3\mu\boldsymbol{I}}{4} & \boldsymbol{C} - \mu\boldsymbol{I} & \boldsymbol{D}(1) & \cdots \\ \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} & \frac{4\mu\boldsymbol{I}}{5} & \boldsymbol{C} - \mu\boldsymbol{I} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix},$$

respectively. (5.2) is then rewritten to be

$$\frac{d}{dx}\overline{\boldsymbol{w}}(x) = \boldsymbol{T}\overline{\boldsymbol{w}}(x). \tag{5.4}$$

Note here that $\boldsymbol{T}$ can be considered as the defective infinitesimal generator of a continuous-time Markov chain with infinite states (n, j) $(n = 0, 1, \ldots, j \in \mathcal{M})$. Therefore the conditional sojourn time of customer C_n given $S_{\mathrm{C}_n} = j$ is equivalent to the first passage time from state (n, j) to an implicit absorbing state.

Note that $\overline{\boldsymbol{w}}_n(0) = \boldsymbol{e}$ for all $n = 0, 1, \ldots$. Then, the formal solution of (5.4) is given by

$$\overline{\boldsymbol{w}}(x) = \exp(\boldsymbol{T}x)\boldsymbol{e}.$$

Thus, applying the uniformization technique (see Appendix A), we obtain

$$\overline{\boldsymbol{w}}(x) = \sum_{k=0}^{\infty} e^{-(\theta+\mu)x} \frac{(\theta+\mu)^k x^k}{k!} \left[\boldsymbol{I} + \frac{1}{\theta+\mu}\boldsymbol{T}\right]^k \boldsymbol{e}, \tag{5.5}$$

where θ denotes the maximum absolute value of diagonal elements of $\boldsymbol{C}$. Let $\boldsymbol{h}_{n,k}$ $(n = 0, 1, \ldots, k = 0, 1, \ldots)$ denote an $M \times 1$ vector that satisfies

$$\begin{bmatrix} \boldsymbol{h}_{0,k} \\ \boldsymbol{h}_{1,k} \\ \vdots \end{bmatrix} = \left[\boldsymbol{I} + \frac{1}{\theta+\mu}\boldsymbol{T}\right]^k \boldsymbol{e}. \tag{5.6}$$

Note here that the jth ($j \in \mathcal{M}$) element of $\boldsymbol{h}_{n,k}$ represents the probability that the number of transitions in transient states of the uniformized Markov chain with transition matrix $\boldsymbol{I}+(\theta+\mu)^{-1}\boldsymbol{T}$ is not less than k during the first passage time from state (n, j) to the implicit absorbing state. Therefore $\boldsymbol{h}_{n,k}$ ($n = 0, 1, \ldots$, $k = 0, 1, \ldots$) satisfies

$$\boldsymbol{0} < \boldsymbol{h}_{n,k+1} < \boldsymbol{h}_{n,k} \leq \boldsymbol{e}, \tag{5.7}$$

where $\boldsymbol{h}_{n,0} = \boldsymbol{e}$ for $n = 0, 1, \ldots$.

We define $\overline{W}(x) = \Pr[W > x]$ for $x \geq 0$, where W denotes the sojourn time of a randomly chosen customer.

Theorem 5.1 *$\overline{\boldsymbol{w}}_n(x)$ ($x \geq 0$, $n = 0, 1, \ldots$) and $\overline{W}(x)$ ($x \geq 0$) are given by*

$$\overline{\boldsymbol{w}}_n(x) = \sum_{k=0}^{\infty} e^{-(\theta+\mu)x} \frac{(\theta+\mu)^k x^k}{k!} \boldsymbol{h}_{n,k}, \tag{5.8}$$

$$\overline{W}(x) = \sum_{n=0}^{\infty} \sum_{m=0}^{n} \frac{(m+1)\boldsymbol{\pi}_{n-m}\boldsymbol{D}(m+1)}{\lambda} \sum_{k=0}^{\infty} e^{-(\theta+\mu)x} \frac{(\theta+\mu)^k x^k}{k!} \boldsymbol{h}_{n,k}, \tag{5.9}$$

respectively, where the $\boldsymbol{h}_{n,k}$ is determined by the following recursion.

$$\boldsymbol{h}_{n,0} = \boldsymbol{e}, \qquad n = 0, 1, \ldots,$$

and for $k = 0, 1, \ldots$,

$$\boldsymbol{h}_{n,k+1} = \frac{1}{\theta+\mu}\left[\frac{n\mu}{n+1}\boldsymbol{h}_{n-1,k} + (\theta\boldsymbol{I}+\boldsymbol{C})\boldsymbol{h}_{n,k} + \sum_{m=1}^{\infty} \boldsymbol{D}(m)\boldsymbol{h}_{n+m,k}\right], \quad n = 0, 1, \ldots, \tag{5.10}$$

where $\boldsymbol{h}_{-1,k} = \boldsymbol{0}$.

Proof. It is clear from (5.3), (5.5) and (5.6) that $\overline{\boldsymbol{w}}_n(x)$ satisfies (5.8). Further it follows from (5.6) that

$$\begin{bmatrix} \boldsymbol{h}_{0,k+1} \\ \boldsymbol{h}_{1,k+1} \\ \vdots \end{bmatrix} = \left[\boldsymbol{I} + \frac{1}{\theta+\mu}\boldsymbol{T}\right] \begin{bmatrix} \boldsymbol{h}_{0,k} \\ \boldsymbol{h}_{1,k} \\ \vdots \end{bmatrix}, \quad k = 0, 1, \ldots,$$

which shows that the $\boldsymbol{h}_{n,k}$ satisfies (5.10).

We can show (5.9) as follows. Note that $\lambda^{-1}\sum_{m=0}^{n}(m+1)\boldsymbol{\pi}_{n-m}\boldsymbol{D}(m+1)$ ($n = 0, 1, \ldots$) is a $1 \times M$ vector whose the jth ($j \in \mathcal{M}$) element represents the joint probability that a customer C_n arrives at the stationary system and the underlying Markov chain is in state j immediately after his arrival [Neut89]. Thus, applying the law of total probabilities to (5.8), we obtain (5.9). ■

In what follows, we consider two special cases, MAP/M/1-PS and M/M/1-PS queues. In the former case, $\boldsymbol{D}(1) = \boldsymbol{D}$ and $\boldsymbol{D}(n) = \boldsymbol{O}$ for all $n = 2, 3, \ldots$. Further, $\boldsymbol{\pi}_n$'s are given by (see Corollary B.1)

$$\begin{aligned} \boldsymbol{\pi}_0 &= \boldsymbol{\pi}(\boldsymbol{I} - \boldsymbol{R}), \\ \boldsymbol{\pi}_n &= \boldsymbol{\pi}_0\boldsymbol{R}^n, \qquad n = 1, 2, \ldots, \end{aligned}$$

where $\boldsymbol{R}$ is the minimal nonnegative solution of

$$\boldsymbol{D} + \boldsymbol{R}(\boldsymbol{C} - \mu\boldsymbol{I}) + \mu\boldsymbol{R}^2 = \boldsymbol{O}.$$

Thus, the following corollary is readily obtained from Theorem 5.1.

Corollary 5.1 *For the MAP/M/1-PS queue, $\overline{\boldsymbol{w}}_n(x)$ $(x \geq 0,\ n = 0, 1, \ldots)$ is given by (5.8) and $\overline{W}(x)$ $(x \geq 0)$ is given by*

$$\overline{W}(x) = \frac{1}{\lambda} \sum_{n=0}^{\infty} \boldsymbol{\pi}_0 \boldsymbol{R}^n \boldsymbol{D} \sum_{k=0}^{\infty} e^{-(\theta+\mu)x} \frac{(\theta+\mu)^k x^k}{k!} \boldsymbol{h}_{n,k}, \tag{5.11}$$

where the $\boldsymbol{h}_{n,k}$ is determined by the following recursion.

$$\boldsymbol{h}_{n,0} = \boldsymbol{e}, \qquad n = 0, 1, \ldots,$$

and for $k = 0, 1, \ldots$,

$$\boldsymbol{h}_{n,k+1} = \frac{1}{\theta+\mu}\left[\frac{n\mu}{n+1}\boldsymbol{h}_{n-1,k} + (\theta\boldsymbol{I} + \boldsymbol{C})\boldsymbol{h}_{n,k} + \boldsymbol{D}\boldsymbol{h}_{n+1,k}\right], \quad n = 0, 1, \ldots, \tag{5.12}$$

where $\boldsymbol{h}_{-1,k} = \boldsymbol{0}$.

Next, we discuss an M/M/1-PS queue with arrival rate λ and service rate μ. Note that, in this queue, $\boldsymbol{\pi}_n$ and $\boldsymbol{h}_{n,k}$ defined in (5.6) are scalars. To make things clear, we here denote them by π_n and $h_{n,k}$, respectively. Further, we define $\overline{w}_n(x)$ $(x \geq 0,\ n = 0, 1, \ldots)$ as $\Pr[W_{\mathrm{C}_n} > x]$. The following result is the direct conclusion of Corollary 5.1.

Corollary 5.2 *For the M/M/1-PS queue, $\overline{w}_n(x)$ $(x \geq 0,\ n = 0, 1, \ldots)$ and $\overline{W}(x)$ $(x \geq 0)$ are given by*

$$\begin{aligned}
\overline{w}_n(x) &= \sum_{k=0}^{\infty} e^{-(\lambda+\mu)x} \frac{(\lambda+\mu)^k x^k}{k!} h_{n,k}, \\
\overline{W}(x) &= \sum_{n=0}^{\infty} (1-\rho)\rho^n \sum_{k=0}^{\infty} e^{-(\lambda+\mu)x} \frac{(\lambda+\mu)^k x^k}{k!} h_{n,k},
\end{aligned}$$

respectively, where the $h_{n,k}$ is determined by the following recursion.

$$h_{n,0} = 1, \qquad n = 0, 1, \ldots,$$

and for $k = 0, 1, \ldots$,

$$h_{n,k+1} = \frac{n}{n+1}\frac{\mu}{\lambda+\mu} h_{n-1,k} + \frac{\lambda}{\lambda+\mu} h_{n+1,k}, \qquad n = 0, 1, \ldots,$$

where $h_{-1,k} = 0$.

Remark 5.1 *For the M/M/1-PS queue, Asmussen obtained a recursive formula similar to Corollary 5.2 (see Theorem 9.5 on page 112 in [Asmu03]). However, it is numerically unstable because subtraction of positive numbers of similar magnitude is involved. On the other hand, our recursion (5.12) is constructed only by sums and products of nonnegative vectors and matrices, and therefore it is numerically stable.*

5.4 Numerical Examples

This section considers MAP/M/1-PS queues. We first discuss the numerical procedure for $\overline{W}(x)$ and the absolute error in numerical results. Next, we discuss the impact of the arrival process on the sojourn time distribution through some numerical examples.

5.4.1 Numerical procedure and accuracy guarantee

For a fixed x ($x > 0$), we compute an approximation to $\overline{W}(x)$ as follows. We first set some ε ($0 < \varepsilon \ll 1$) and obtain a minimum nonnegative integer $N(\varepsilon)$ satisfying

$$\frac{1}{\lambda}\sum_{n=0}^{N(\varepsilon)} \boldsymbol{\pi}_0 \boldsymbol{R}^n \boldsymbol{D}\boldsymbol{e} > 1 - \varepsilon. \tag{5.13}$$

Note that $N(\varepsilon)$ is uniquely determined because

$$\sum_{n=0}^{\infty} \boldsymbol{\pi}_0 \boldsymbol{R}^n \boldsymbol{D}\boldsymbol{e} = \lambda. \tag{5.14}$$

Next, we set some ε' ($0 < \varepsilon' \ll 1$) and obtain two nonnegative integers $L(\varepsilon', x)$ and $R(\varepsilon', x)$ satisfying $L(\varepsilon', x) \le R(\varepsilon', x)$ and

$$\sum_{k=L(\varepsilon',x)}^{R(\varepsilon',x)} \frac{(\theta+\mu)^k x^k}{k!} e^{-(\theta+\mu)x} \ge 1 - \varepsilon'. \tag{5.15}$$

One of algorithms to find $L(\varepsilon', x)$ and $R(\varepsilon', x)$ is proposed by Fox and Glynn [Fox88]. We use their algorithm for numerical examples in the following subsections.

Finally, we compute an approximation $\overline{W}_{\rm comp}(x)$ to $\overline{W}(x)$:

$$\overline{W}_{\rm comp}(x) = \frac{1}{\lambda}\sum_{n=0}^{N(\varepsilon)} \boldsymbol{\pi}_0 \boldsymbol{R}^n \boldsymbol{D} \sum_{k=L(\varepsilon',x)}^{R(\varepsilon',x)} \frac{(\theta+\mu)^k x^k}{k!} e^{-(\theta+\mu)x} \boldsymbol{h}_{n,k}. \tag{5.16}$$

Note here that we have to compute $\boldsymbol{h}_{n,k}$'s in the trapezoidal domain in Figure 5.1, although $\overline{W}_{\rm comp}(x)$ requires $\boldsymbol{h}_{n,k}$'s only in the shaded rectangular domain in Figure 5.1 (see (5.12)).

The bound for the difference between $\overline{W}(x)$ and $\overline{W}_{\rm comp}(x)$ can be obtained as follows. For simplicity in description, $L(\varepsilon', x)$ and $R(\varepsilon', x)$ are denoted by L and R, respectively. From (5.11) and (5.16), we have

$$\begin{aligned}\overline{W}(x) - \overline{W}_{\rm comp}(x) &= \frac{1}{\lambda}\sum_{n=N(\varepsilon)+1}^{\infty} \boldsymbol{\pi}_0 \boldsymbol{R}^n \boldsymbol{D} \sum_{k=0}^{\infty} \frac{(\theta+\mu)^k x^k}{k!} e^{-(\theta+\mu)x} \boldsymbol{h}_{n,k} \\ &\quad + \frac{1}{\lambda}\sum_{n=0}^{N(\varepsilon)} \boldsymbol{\pi}_0 \boldsymbol{R}^n \boldsymbol{D} \sum_{k<L,k>R} \frac{(\theta+\mu)^k x^k}{k!} e^{-(\theta+\mu)x} \boldsymbol{h}_{n,k}.\end{aligned}$$

Applying (5.7), (5.13), (5.14) and (5.15) to the above equation, we obtain

$$\overline{W}(x) - \overline{W}_{\rm comp}(x) < \frac{1}{\lambda}\sum_{n=N(\varepsilon)+1}^{\infty} \boldsymbol{\pi}_0 \boldsymbol{R}^n \boldsymbol{D}\boldsymbol{e} \sum_{k=0}^{\infty} \frac{(\theta+\mu)^k x^k}{k!} e^{-(\theta+\mu)x}$$

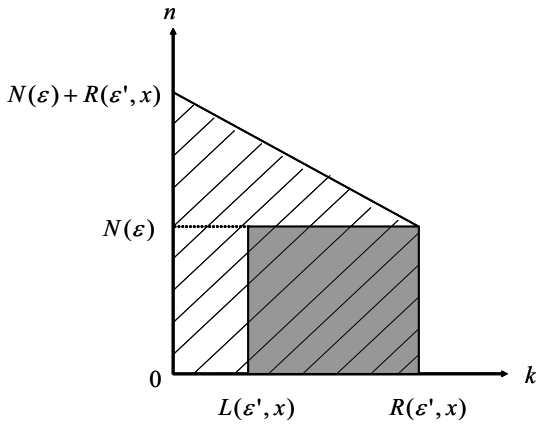

Figure 5.1: The computational domain of the $\boldsymbol{h}_{n,k}$.

$$
\begin{aligned}
& + \frac{1}{\lambda} \sum_{n=0}^{N(\varepsilon)} \boldsymbol{\pi}_0 \boldsymbol{R}^n \boldsymbol{D} \boldsymbol{e} \sum_{k<L,k>R} \frac{(\theta+\mu)^k x^k}{k!} e^{-(\theta+\mu)x} \\
\leq \quad & \varepsilon + \varepsilon'.
\end{aligned}
$$

For numerical examples in sections 5.4.2 and 5.4.3, we set $\varepsilon = 10^{-11}$ and $\varepsilon' = 10^{-10}$.

5.4.2 Impact of variation in inter-arrival times

In this subsection, we discuss the impact of variation of inter-arrival times on the sojourn time. For this purpose, we assume that inter-arrival times are i.i.d. and consider the following three types of inter-arrival time distributions with the same mean λ^{-1}; four-stage Erlang distribution, exponential distribution and two-stage balanced hyper-exponential distribution $\psi(x) = 1 - pe^{-2p\lambda x} - (1-p)e^{-2(1-p)\lambda x}$ with $p = 0.5 + \sqrt{15}/10$. Let $C_{\rm v}$ denote the coefficient of variation of the inter-arrival time distribution. Note that $C_{\rm v} = 0.5$ for the four-stage Erlang distribution, $C_{\rm v} = 1$ for the exponential distribution and $C_{\rm v} = 2$ for the above two-stage balanced hyper-exponential distribution.

We set $\lambda = 0.8$ and $\mu = 1$, so that $\rho = 0.8$ for all three cases. Figure 5.2 plots the complementary distributions for these cases. We observe that the sojourn time distribution becomes larger stochastically with the increase of $C_{\rm v}$.

5.4.3 Impact of correlation in inter-arrival times

Next we discuss the impact of correlation in inter-arrival times on the sojourn time. For this purpose, we set $\mu = 1$ and assume that the arrival process follows a stationary two-state MAP:

$$
\boldsymbol{C} = \begin{bmatrix} -2 & 0 \\ 0 & -0.5 \end{bmatrix}, \quad \boldsymbol{D} = \begin{bmatrix} 2p & 2(1-p) \\ 0.5(1-p) & 0.5p \end{bmatrix},
$$

where $0 \leq p < 1$. Note here that $\lambda = 0.8$ and hence $\rho = 0.8$.

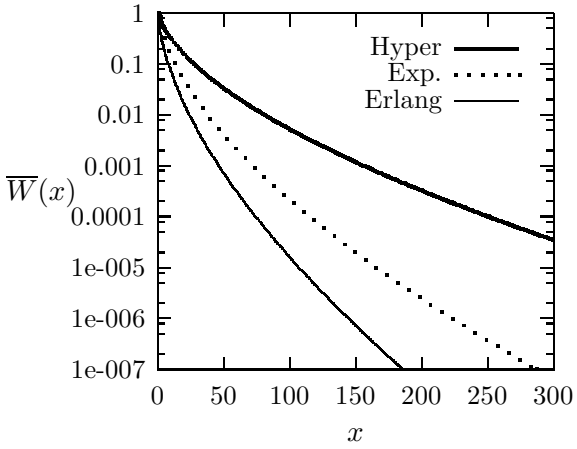

Figure 5.2: The complementary distribution $\overline{W}(x)$ of the sojourn time.

Let Y_n $(n = 1, 2, \ldots)$ denote a random variable representing the nth inter-arrival time. Note that the marginal distribution of Y_n is independent of p and follows a hyper-exponential distribution $\psi(x) = 1 - 0.5e^{-2x} - 0.5e^{-0.5x}$. Note also that the covariance of Y_n and Y_{n+1} is given by $9(2p-1)/16$. Therefore inter-arrival times are positively (resp. negatively) correlated for $0.5 < p < 1$ (resp. $0 \leq p < 0.5$), and are i.i.d. for $p = 0.5$.

Figure 5.3 plots the complementary distributions for the three cases, $p = 0.1$, 0.5 and 0.9. We observe that the positive (resp. negative) correlation of inter-arrival times makes the sojourn time stochastically larger (resp. smaller) than that for i.i.d. inter-arrival times.

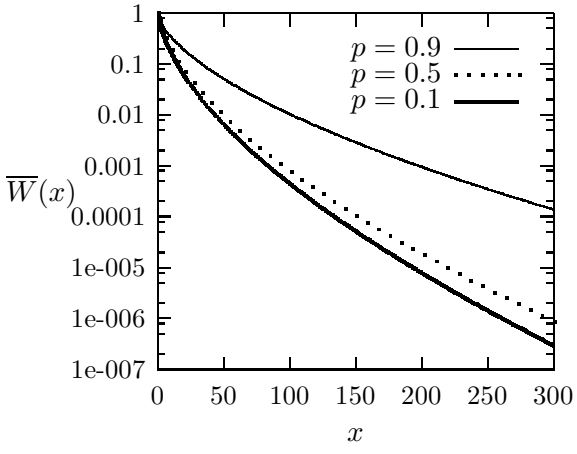

Figure 5.3: The complementary distribution $\overline{W}(x)$ of the sojourn time.

5.5 Concluding Remarks

In this chapter, we derived a numerically feasible formula to compute the sojourn time distribution in a MAP/M/1-PS queue. It is known that in the GI/M/1 queue, the sojourn time distribution for PS is identical to the conditional waiting time distribution for random order of service (ROS) given the waiting time is positive [Bors03, Cohe84]. Note that in the stationary MAP/M/1 queue, the sojourn time distribution for PS is closely related to the conditional waiting time distribution for ROS. To see this, we define $W_{\mathrm{C}_n}^{\mathrm{ROS}}$ as a generic random variable representing the waiting time of customers finding n customers in the ROS system on arrival. Let $\overline{\boldsymbol{w}}_n^{\mathrm{ROS}}(x)$ denote an $M \times 1$ vector whose jth ($j \in \mathcal{M}$) element represents $\Pr[W_{\mathrm{C}_n}^{\mathrm{ROS}} > x \mid S_{\mathrm{C}_n} = j]$, where S_{C_n} denotes the state of the underlying Markov chain immediately after arrivals finding n customers in the system. It then follows from a very similar reasoning in section 2 of [Bors03] that for $n = 0, 1, \ldots$,

$$\overline{\boldsymbol{w}}_{n+1}^{\mathrm{ROS}}(x) = \overline{\boldsymbol{w}}_n(x), \qquad x \geq 0,$$

where $\overline{\boldsymbol{w}}_n(x)$ is given in Corollary 5.1. Thus the complementary distribution $\overline{W}^{\mathrm{ROS}}(x)$ of waiting time in a MAP/M/1-ROS queue is given by

$$\overline{W}^{\mathrm{ROS}}(x) = \frac{1}{\lambda} \sum_{n=0}^{\infty} \boldsymbol{\pi}_0 \boldsymbol{R}^{n+1} \boldsymbol{D} \overline{\boldsymbol{w}}_n(x), \qquad x \geq 0,$$

which means that our results can be also used to obtain the waiting time distribution in a MAP/M/1-ROS queue.

Chapter 6

Conclusion

6.1 Summary of Results

This thesis studied queues with batch Markovian arrival streams, focusing on the computation of performance measures of interest. The summary of the results is described below.

(a) We analyzed FIFO single-server queues with/without service interruptions by the approach based on [Taki01a, Taki01b] and established a numerical algorithm to compute the joint queue length distribution. Further, we examined the impact of system parameters on the queue length distribution through some numerical examples.

(b) Extremely general infinite-server queues were analyzed. By simple probabilistic consideration, we obtained the system of linear differential equations for the time-dependent matrix joint generating function of the queue length. Further, assuming phase-type service times, we obtained explicit and numerically feasible expressions of the time-dependent and limiting joint binomial moments.

(c) The sojourn time distribution in a BMAP/M/1-PS queue was studied. Based on the uniformization technique (see Appendix A), the recursive formula for the sojourn time distribution was obtained. In addition, for a MAP/M/1-PS queue, we established a numerically feasible algorithm to compute the sojourn time distribution. The advantage of the algorithm is that the accuracy of computational results can be estimated in advance.

6.2 Future Work

This thesis studies some types of queues with batch Markovian arrival streams. However, the study of those queues has just started. The author describes some future works below.

(a) As for the single-server queue considered chapter 2, we can obtain the mean sojourn time of customers in each class by using Theorem 2.1 and the results on moments of the virtual waiting time distribution [Taki94a]. Further with Little's law, we can readily derive the mean

queue length of each class. We have, however, no way to compute higher-order moments such as a variance. This is a problem to be solved.

(b) The computational procedures proposed in chapters 2 and 3 are two-step algorithms to obtain the stationary joint queue length distribution. The first step is to obtain the joint queue length distribution immediately after departures, and the second step is to derive the joint queue length distribution at a random point in time, using the relationship of these two joint queue length distributions (see (2.8) and (2.9) in Corollary 2.1). The first step is numerically feasible, but the second step is not necessarily so because we have to subtract positive numbers of similar magnitude (see (2.9)). Therefore, to establish a numerically feasible algorithm directly to compute the stationary joint queue length distribution is a challenging problem.

(c) A multi-server queue with batch Markovian arrival streams is an interesting model. Such a queue is so flexible that it can cover most of the queues we can now think of. Thus, if we obtained analytic results suitable for computing performance measures of interest, we could establish unified numerical algorithms for almost all of the queues we often encounter.

Appendix A

Uniformization

We consider a continuous-time Markov chain $\{X(t); t \geq 0\}$ with state space $\mathcal{L}$ and infinitesimal generator $\widetilde{\boldsymbol{P}}$. Assuming that diagonal elements of $\widetilde{\boldsymbol{P}}$ are bounded from below, we choose a positive number θ such that

$$\sup_{j \in \mathcal{L}} |[\widetilde{\boldsymbol{P}}]_{j,j}| \leq \theta < \infty,$$

where $[\widetilde{\boldsymbol{P}}]_{i,j}$ denotes the (i,j)th element of $\widetilde{\boldsymbol{P}}$. Let $\widetilde{\boldsymbol{\Pi}}(t)$ denote a matrix whose (i,j)th element represents $\Pr[X(t) = j \mid X(0) = i]$. Then, $\widetilde{\boldsymbol{\Pi}}(t)$ is given by

$$\widetilde{\boldsymbol{\Pi}}(t) = \exp(\widetilde{\boldsymbol{P}}t) = \sum_{k=0}^{\infty} e^{-\theta t} \frac{(\theta t)^k}{k!} [\boldsymbol{I} + \theta^{-1} \widetilde{\boldsymbol{P}}]^k = \sum_{k=0}^{\infty} e^{-\theta t} \frac{(\theta t)^k}{k!} \boldsymbol{P}^k, \tag{A.1}$$

where

$$\boldsymbol{P} = \boldsymbol{I} + \theta^{-1} \widetilde{\boldsymbol{P}}.$$

Note that $\boldsymbol{P}$ is a nonnegative matrix because non-diagonal elements of $\widetilde{\boldsymbol{P}}$ are all nonnegative. Note also that $\boldsymbol{P}\boldsymbol{e} = \boldsymbol{e}$ owing to $\widetilde{\boldsymbol{P}}\boldsymbol{e} = \boldsymbol{0}$. These two facts mean that $\boldsymbol{P}$ is a stochastic matrix. Thus, if the state space $\mathcal{L}$ is finite, we can compute the transient distribution of the continuous-time Markov chain $\{X(t); t \geq 0\}$, using (A.1).

We now introduce a discrete-time Markov chain $\{X_k; k = 0, 1, \ldots\}$ with the same state space $\mathcal{L}$ as $\{X(t); t \geq 0\}$, whose transition probability matrix is given by $\boldsymbol{P}$. Noting that

$$\exp(\widetilde{\boldsymbol{P}}t) = \sum_{k=0}^{\infty} e^{-\theta t} \frac{(\theta t)^k}{k!} \boldsymbol{P}^k,$$

we see that if a continuous-time Markov chain $\{X(t); t \geq 0\}$ is irreducible, a discrete-time Markov chain $\{X_k; k = 0, 1, \ldots\}$ is also irreducible, and vice versa. Assuming that $\{X(t); t \geq 0\}$ is positive recurrent, let $\widetilde{\boldsymbol{\pi}}$ denote the stationary distribution of $\{X(t); t \geq 0\}$, i.e., $\widetilde{\boldsymbol{\pi}}\widetilde{\boldsymbol{P}} = \boldsymbol{0}$. We then have

$$\widetilde{\boldsymbol{\pi}}\boldsymbol{P} = \widetilde{\boldsymbol{\pi}}(\boldsymbol{I} + \theta^{-1}\widetilde{\boldsymbol{P}}) = \widetilde{\boldsymbol{\pi}}.$$

Thus, $\{X_k; k = 0, 1, \ldots\}$ is also positive recurrent and has the same stationary distribution $\widetilde{\boldsymbol{\pi}}$.

We call, *uniformization*, the above technique that converts a continuous-time Markov chain to a discrete-time Markov chain with the same stationary distribution [Tijm94]. In general, discrete-time Markov chains are more manageable than continuous-time Markov chains, because transition probability matrices, which characterize discrete-time Markov chains, are nonnegative. Therefore, the uniformization technique is very often used in algorithmic analysis of queues.

Appendix B

Queue Length Distribution in a BMAP/GI/1 Queue

We discuss a stationary FIFO single-server queue with a single batch Markovian arrival stream, i.e., a BMAP/GI/1 queue. Customer arrivals follow a BMAP characterized by $\boldsymbol{C}$ and $\boldsymbol{D}(n)$ (see subsection 1.1.1), and service times are i.i.d. according to a distribution function $H(x)$. Let $\boldsymbol{q}_{\mathrm{BMAP/GI/1}}(n)$ $(n = 0, 1, \ldots)$ denote a vector whose jth element represents the joint probability that the queue length is equal to n and the underlying Markov chain is in state j immediately after departures. Then the $\boldsymbol{q}_{\mathrm{BMAP/GI/1}}(n)$ is identical to the steady-state solution for the M/G/1-type Markov chain whose transition probability matrix is given by

$$\begin{bmatrix} \boldsymbol{B}_0 & \boldsymbol{B}_1 & \boldsymbol{B}_2 & \boldsymbol{B}_3 & \cdots \\ \boldsymbol{A}_0 & \boldsymbol{A}_1 & \boldsymbol{A}_2 & \boldsymbol{A}_3 & \cdots \\ \boldsymbol{O} & \boldsymbol{A}_0 & \boldsymbol{A}_1 & \boldsymbol{A}_2 & \cdots \\ \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{A}_0 & \boldsymbol{A}_1 & \cdots \\ \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{A}_0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \tag{B.1}$$

where $\boldsymbol{A}_n$ and $\boldsymbol{B}_n$ satisfy

$$\begin{aligned} \sum_{n=0}^{\infty} z^n \boldsymbol{A}_n &= \int_0^{\infty} \exp\left[\left(\boldsymbol{C} + \sum_{m=0}^{\infty} z^m \boldsymbol{D}(m)\right) x\right] dH(x), \\ \boldsymbol{B}_n &= (-\boldsymbol{C})^{-1} \boldsymbol{A}_n. \end{aligned} \tag{B.2}$$

Thus, $\boldsymbol{q}_{\mathrm{BMAP/GI/1}}(n)$'s can be computed by the M/G/1 paradigm (see subsection 1.2.2).

Remark B.1 *The recursion of $\boldsymbol{A}_n$'s in the BMAP/GI/1 queue is given in Lemma 2.2 with $K = 1$, and a remark on its implementation is found in section 2.7.*

We now define $\boldsymbol{p}_{\mathrm{BMAP/GI/1}}(n)$ $(n = 0, 1, \ldots)$ as a vector whose jth element represents the stationary joint probability that the queue length is equal to n and the underlying Markov chain is in state j at a random point in time. It is known that $\boldsymbol{p}_{\mathrm{BMAP/GI/1}}(n)$ and $\boldsymbol{q}_{\mathrm{BMAP/GI/1}}(n)$ have

the following relationship:

$$\boldsymbol{p}_{\mathrm{BMAP/GI/1}}(0) = \lambda \boldsymbol{q}_{\mathrm{BMAP/GI/1}}(0)(-\boldsymbol{C})^{-1}, \tag{B.3}$$

$$\boldsymbol{p}_{\mathrm{BMAP/GI/1}}(n) = \Bigg[\lambda\Big\{\boldsymbol{q}_{\mathrm{BMAP/GI/1}}(n) - \boldsymbol{q}_{\mathrm{BMAP/GI/1}}(n-1)\Big\} + \sum_{m=1}^{n} \boldsymbol{p}_{\mathrm{BMAP/GI/1}}(n-m)\boldsymbol{D}(m)\Bigg](-\boldsymbol{C})^{-1}, \tag{B.4}$$

where λ is given by (1.2). Thus, $\boldsymbol{p}_{\mathrm{BMAP/GI/1}}(n)$'s can be obtained in terms of $\boldsymbol{q}_{\mathrm{BMAP/GI/1}}(n)$'s. Note here that the above recursion is not always stable because we have to execute the subtraction $\boldsymbol{q}_{\mathrm{BMAP/GI/1}}(n) - \boldsymbol{q}_{\mathrm{BMAP/GI/1}}(n-1)$.

Recently, Takine established a stable and direct recursion for the $\boldsymbol{p}_{\mathrm{BMAP/GI/1}}(n)$ [Taki00]. Takine's recursion is based on the fact that the $\boldsymbol{p}_{\mathrm{BMAP/GI/1}}(n)$ is identical to the steady-state solution for a certain embedded Markov chain in the corresponding BMAP/GI/1 queue with multiple vacations and exhaustive services. In conclusion, it is shown that the $\boldsymbol{p}_{\mathrm{BMAP/GI/1}}(n)$ is identical to the steady-state solution for the M/G/1-type Markov chain with the following transition probability matrix [Taki00]:

$$\begin{bmatrix} \boldsymbol{A}_0 & \boldsymbol{A}_1 & \boldsymbol{A}_2 & \boldsymbol{A}_3 & \cdots \\ \boldsymbol{A}_0 & \boldsymbol{A}_1 & \boldsymbol{A}_2 & \boldsymbol{A}_3 & \cdots \\ \boldsymbol{O} & \boldsymbol{A}_0 & \boldsymbol{A}_1 & \boldsymbol{A}_2 & \cdots \\ \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{A}_0 & \boldsymbol{A}_1 & \cdots \\ \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{A}_0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \tag{B.5}$$

where $\boldsymbol{A}_n$'s satisfy (B.2). A related work is found in [Lee00], which considered an N/G/1 queue. Note that $\boldsymbol{K}$ in (1.11) is identical to fundamental matrix $\boldsymbol{G}$ for this special Markov chain of M/G/1 type (see subsection 1.2.2). Let $\boldsymbol{g}$ denote a probability vector satisfying $\boldsymbol{gG} = \boldsymbol{g}$. Then, Proposition 1.2 is reduced to the following result.

Proposition B.1 ([Taki00]) *$\boldsymbol{p}_{\mathrm{BMAP/GI/1}}(n)$'s ($n = 0, 1, \ldots$) are recursively determined in the following way.*

$$\boldsymbol{p}_{\mathrm{BMAP/GI/1}}(0) = (1-\rho)\boldsymbol{g},$$

where ρ denotes the utilization factor. For $n = 1, 2, \ldots$,

$$\boldsymbol{p}_{\mathrm{BMAP/GI/1}}(n) = \left[\boldsymbol{p}_{\mathrm{BMAP/GI/1}}(0)\overline{\boldsymbol{A}}(n) + \sum_{l=1}^{n-1} \boldsymbol{p}_{\mathrm{BMAP/GI/1}}(l)\overline{\boldsymbol{A}}(n-l+1)\right][\boldsymbol{I} - \overline{\boldsymbol{A}}(1)]^{-1}.$$

We next consider a special case of a BMAP/GI/1 queue, where the arrival process is a MAP characterized by $(\boldsymbol{C}, \boldsymbol{D})$ (see subsection 1.1.1) and service times follow a phase-type distribution $(\boldsymbol{\beta}, \boldsymbol{T})$. This queue is denoted by MAP/PH/1. Let $\boldsymbol{p}_{\mathrm{MAP/PH/1}}(0)$ denote a vector whose jth element represents the stationary probability that the system is idle and the underlying Markov chain is in state j. Also let $p_{\mathrm{MAP/PH/1},j,\nu}(n)$ ($n = 1, 2, \ldots$) denote the stationary joint probability that

the queue length is equal to n, the underlying Markov chain is in state j and the phase of service is ν. We construct a vector $\boldsymbol{p}_{\mathrm{MAP/PH/1}}(n)$ by arranging $p_{\mathrm{MAP/PH/1},j,\nu}(n)$'s in the lexicographical order, where the first index is the state j of the underlying Markov chain and the second index is the phase ν of service. Then the $\boldsymbol{p}_{\mathrm{MAP/PH/1}}(n)$ is identical to the steady-state solution for the continuous-time M/G/1-type Markov chain whose infinitesimal generator is given by

$$\begin{bmatrix} \boldsymbol{C} & \boldsymbol{D}\otimes\boldsymbol{\beta} & \boldsymbol{O} & \boldsymbol{O} & \cdots \\ \boldsymbol{I}_{\mathrm{MAP}}\otimes(-\boldsymbol{T})\boldsymbol{e} & \boldsymbol{C}\oplus\boldsymbol{T} & \boldsymbol{D}\otimes\boldsymbol{I}_{\mathrm{PH}} & \boldsymbol{O} & \cdots \\ \boldsymbol{O} & \boldsymbol{I}_{\mathrm{MAP}}\otimes(-\boldsymbol{T})\boldsymbol{e}\boldsymbol{\beta} & \boldsymbol{C}\oplus\boldsymbol{T} & \boldsymbol{D}\otimes\boldsymbol{I}_{\mathrm{PH}} & \cdots \\ \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{I}_{\mathrm{MAP}}\otimes(-\boldsymbol{T})\boldsymbol{e}\boldsymbol{\beta} & \boldsymbol{C}\oplus\boldsymbol{T} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \tag{B.6}$$

where $\boldsymbol{I}_{\mathrm{MAP}}$ and $\boldsymbol{I}_{\mathrm{PH}}$ denote identity matrices with the same dimensions as $\boldsymbol{C}$ and $\boldsymbol{T}$, respectively. Note that with the uniformization technique (see Appendix A), the above continuous-time Markov chain of G/M/1-type can be converted into a discrete-time Markov chain of G/M/1 type, whose transition probability matrix has the same structure as $\boldsymbol{P}_{\mathrm{R}}$ in (1.8), setting $\boldsymbol{E}_n = \boldsymbol{O}$ for $n \geq 3$. Thus, from Proposition 1.1, we can see that the $\boldsymbol{p}_{\mathrm{MAP/PH/1}}(n)$ is a matrix geometric solution. This result has an advantage over Takine's recursion (see Proposition B.1) if the number of phases of service is considerably small. In the case of exponential service, we recommend computing the stationary queue length distribution with the matrix geometric solution.

Finally, we show the matrix geometric solution of a MAP/M/1 queue. Let $\boldsymbol{p}_{\mathrm{MAP/M/1}}(n)$ $(n = 0, 1, \ldots)$ denote a vector whose jth element represents the stationary probability that the queue length is equal to n and the underlying Markov chain is in state j.

Corollary B.1 *$\boldsymbol{p}_{\mathrm{MAP/M/1}}(n)$'s are given by*

$$\begin{aligned} \boldsymbol{p}_{\mathrm{MAP/M/1}}(0) &= \boldsymbol{\pi}(\boldsymbol{I}-\boldsymbol{R}), \\ \boldsymbol{p}_{\mathrm{MAP/M/1}}(n) &= \boldsymbol{p}_{\mathrm{MAP/M/1}}(0)\boldsymbol{R}^n, \qquad n-1,2,\ldots, \end{aligned}$$

where $\boldsymbol{\pi}$ is a probability vector satisfying $\boldsymbol{\pi}(\boldsymbol{C}+\boldsymbol{D}) = \boldsymbol{0}$ and $\boldsymbol{R}$ is the minimal nonnegative solution of

$$\boldsymbol{D} + \boldsymbol{R}(\boldsymbol{C}-\mu\boldsymbol{I}) + \mu\boldsymbol{R}^2 = \boldsymbol{O}.$$

Appendix C

Total Queue Length Distribution in a FIFO Queue

We show a numerical algorithm to compute the stationary total queue length distribution, by modifying our algorithm for the joint queue length distribution. We define $\boldsymbol{p}^{(\mathrm{T})}(n)$ $(n = 0, 1, \ldots)$ and $\boldsymbol{q}_k^{(\mathrm{T})}(n)$ $(k \in \mathcal{K},\, n = 0, 1, \ldots)$ as

$$\boldsymbol{p}^{(\mathrm{T})}(n) = \sum_{\substack{\boldsymbol{n}\in\mathcal{Z} \\ |\boldsymbol{n}|=n}} \boldsymbol{p}(\boldsymbol{n}), \qquad \boldsymbol{q}_k^{(\mathrm{T})}(n) = \sum_{\substack{\boldsymbol{n}\in\mathcal{Z} \\ |\boldsymbol{n}|=n}} \boldsymbol{q}_k(\boldsymbol{n}),$$

respectively. Corollary 2.1 is then reduced to

Corollary C.1 *The $\boldsymbol{p}^{(\mathrm{T})}(n)$ is recursively determined in the following way:*

$$\boldsymbol{p}^{(\mathrm{T})}(0) = \sum_{k\in\mathcal{K}} \lambda_k \boldsymbol{q}_k^{(\mathrm{T})}(0)(-\boldsymbol{C})^{-1},$$

and for $n = 1, 2, \ldots$,

$$\boldsymbol{p}^{(\mathrm{T})}(n) = \sum_{k\in\mathcal{K}} \left[\lambda_k \left(\boldsymbol{q}_k^{(\mathrm{T})}(n) - \boldsymbol{q}_k^{(\mathrm{T})}(n-1) \right) + \sum_{m=1}^{n} \boldsymbol{p}^{(\mathrm{T})}(n-m)\boldsymbol{D}_k(m) \right] (-\boldsymbol{C})^{-1}.$$

Further, under Assumption 2.2, Theorem 2.4 is reduced to

$$\boldsymbol{q}_k^{(\mathrm{T})}(n) = \frac{1}{\lambda_k} \sum_{\substack{m_1+m_2+m_3 \\ +m_4=n}} \boldsymbol{v}_k^{(\mathrm{T})}(m_1)[\boldsymbol{\alpha}_k \otimes \boldsymbol{A}_k^{(\mathrm{T})}(m_2)]\boldsymbol{\Gamma}_k^{(\mathrm{T})}(m_3) \left[\{\boldsymbol{P}_k^{m_4}(\boldsymbol{I} - \boldsymbol{P}_k)\boldsymbol{e}\} \otimes \boldsymbol{I}(M) \right],$$

where

$$\begin{aligned} \boldsymbol{A}_k^{(\mathrm{T})}(n) &= \sum_{\substack{\boldsymbol{n}\in\mathcal{Z} \\ |\boldsymbol{n}|=n}} \boldsymbol{A}_k(\boldsymbol{n}), \qquad (\mathrm{C}.1) \\ \boldsymbol{v}_k^{(\mathrm{T})}(n) &= \sum_{\substack{\boldsymbol{n}\in\mathcal{Z} \\ |\boldsymbol{n}|=n}} \boldsymbol{v}_k(\boldsymbol{n}), \qquad \boldsymbol{\Gamma}_k^{(\mathrm{T})}(n) = \sum_{\substack{\boldsymbol{n}\in\mathcal{Z} \\ |\boldsymbol{n}|=n}} \boldsymbol{\Gamma}_k(\boldsymbol{n}). \end{aligned}$$

Thus the $\boldsymbol{p}^{(\mathrm{T})}(n)$ is obtained if we compute the $\boldsymbol{A}_k^{(\mathrm{T})}(n)$, the $\boldsymbol{v}_k^{(\mathrm{T})}(n)$ and the $\boldsymbol{\Gamma}_k^{(\mathrm{T})}(n)$.

Note here that $\boldsymbol{A}_k^{(\mathrm{T})}(n)$, $\boldsymbol{v}_k^{(\mathrm{T})}(n)$ and $\boldsymbol{\Gamma}_k^{(\mathrm{T})}(n)$ satisfy

$$\begin{aligned}
\sum_{n=0}^{\infty} z^n \boldsymbol{A}_k^{(\mathrm{T})}(n) &= \int_0^{\infty} dH_k(x) \exp\left[\left(\boldsymbol{C} + \sum_{k\in\mathcal{K}} \boldsymbol{D}_k^*(z)\right) x\right], \qquad \text{(C.2)}\\
\sum_{n=0}^{\infty} z^n \boldsymbol{v}_k^{(\mathrm{T})}(n) &= \int_0^{\infty} d\boldsymbol{v}(x)\boldsymbol{D}_k \exp\left[\left(\boldsymbol{C} + \sum_{k\in\mathcal{K}} \boldsymbol{D}_k^*(z)\right) x\right],\\
\sum_{n=0}^{\infty} z^n \boldsymbol{\Gamma}_k^{(\mathrm{T})}(n) &= \left[\boldsymbol{I} - \boldsymbol{P}_k \otimes \int_0^{\infty} dH_k(x) \exp\left[\left(\boldsymbol{C} + \sum_{k\in\mathcal{K}} \boldsymbol{D}_k^*(z)\right) x\right]\right]^{-1},
\end{aligned}$$

respectively. Thus $\boldsymbol{A}_k^{(\mathrm{T})}(n)$ and $\boldsymbol{v}_k^{(\mathrm{T})}(n)$ can be written to be

$$\begin{aligned}
\boldsymbol{A}_k^{(\mathrm{T})}(n) &= \sum_{m=0}^{\infty} \gamma_k^{(m)}(\theta)\boldsymbol{F}_m^{(\mathrm{T})}(n),\\
\boldsymbol{v}_k^{(\mathrm{T})}(n) &= \sum_{m=0}^{\infty} \boldsymbol{v}^{(m)}(\theta)\boldsymbol{D}_k\boldsymbol{F}_m^{(\mathrm{T})}(n),
\end{aligned}$$

respectively, where $\boldsymbol{F}_m^{(\mathrm{T})}(n)$ denotes an $M \times M$ matrix which satisfies

$$\sum_{n=0}^{\infty} z^n \boldsymbol{F}_m^{(\mathrm{T})}(n) = \left[\boldsymbol{I} + \theta^{-1}\left(\boldsymbol{C} + \sum_{k\in\mathcal{K}} \boldsymbol{D}_k^*(z)\right)\right]^m.$$

Further the $\boldsymbol{\Gamma}_k^{(\mathrm{T})}(n)$ ($k \in \mathcal{K}$, $n \geq 0$) is determined by the following recursion:

$$\boldsymbol{\Gamma}_k^{(\mathrm{T})}(0) = \left[\boldsymbol{I} - \boldsymbol{P}_k \otimes \boldsymbol{A}_k^{(\mathrm{T})}(0)\right]^{-1},$$

and for $n = 1, 2, \ldots$,

$$\boldsymbol{\Gamma}_k^{(\mathrm{T})}(n) = \sum_{l=1}^{n} \boldsymbol{\Gamma}_k^{(\mathrm{T})}(n-l)\left[\boldsymbol{P}_k \otimes \boldsymbol{A}_k^{(\mathrm{T})}(l)\right]\boldsymbol{\Gamma}_k^{(\mathrm{T})}(0).$$

Thus we can compute $\boldsymbol{A}_k^{(\mathrm{T})}(n)$, $\boldsymbol{v}_k^{(\mathrm{T})}(n)$ and $\boldsymbol{\Gamma}_k^{(\mathrm{T})}(n)$ by replacing Steps 3 and 4 with the followings.

Step 3. Compute $\breve{\boldsymbol{A}}_k^{(\mathrm{T})}(n)$ and $\breve{\boldsymbol{v}}_k^{(\mathrm{T})}(n)$ by the following procedure, where the initial values of $\breve{\boldsymbol{A}}_k^{(\mathrm{T})}(n)$ and $\breve{\boldsymbol{v}}_k^{(\mathrm{T})}(n)$ ($n \geq 0$) are assumed to be $\boldsymbol{O}$ and $\boldsymbol{0}$, respectively.

Step (3–a). Set $\breve{\boldsymbol{F}}_0^{(\mathrm{T})}(0) = \boldsymbol{I}$ and $n_F^{(0)} = 0$. Also set

$$\breve{\boldsymbol{A}}_k^{(\mathrm{T})}(0) = \gamma_k^{(0)}(\theta)\boldsymbol{I}, \quad \breve{\boldsymbol{v}}_k^{(\mathrm{T})}(0) = \boldsymbol{v}^{(0)}(\theta)\boldsymbol{D}_k, \quad \forall k \in \mathcal{K}.$$

Step (3–b). Set $n_F^{(1)} = \max_{k\in\mathcal{K}} n_g(k)$ and $m = 1$, and compute $\breve{\boldsymbol{F}}_1^{(\mathrm{T})}(n)$ by the following recursion:

$$\breve{\boldsymbol{F}}_1^{(\mathrm{T})}(0) = \boldsymbol{I} + \theta^{-1}\boldsymbol{C},$$

and for $n = 1, 2, \ldots, n_F^{(1)}$,

$$\breve{\boldsymbol{F}}_1^{(\mathrm{T})}(n) = \theta^{-1} \sum_{k\in\mathcal{K}} U(n_g(k) - n) g_k(n) \boldsymbol{D}_k.$$

Step (3–c). For each $k \in \mathcal{K}$, if $m \le m_\gamma(k)$, add $\gamma_k^{(m)}(\theta)\breve{\boldsymbol{F}}_m^{(\mathrm{T})}(n)$ to $\breve{\boldsymbol{A}}_k^{(\mathrm{T})}(n)$ for all $n \le n_F^{(m)}$. Also, for each $k \in \mathcal{K}$, if $m \le m_v(k)$, add $\boldsymbol{v}^{(m)}(\theta)\boldsymbol{D}_k\breve{\boldsymbol{F}}_m^{(\mathrm{T})}(n)$ to $\breve{\boldsymbol{v}}_k^{(\mathrm{T})}(n)$ for all $n \le n_F^{(m)}$.

Step (3–d). If $m \ge m_{\max}$, stop computing, and otherwise, add one to m and go to Step (3–e).

Step (3–e). For each $n = 0, 1, \ldots$, compute $\breve{\boldsymbol{F}}_m^{(\mathrm{T})}(n)$ by

$$\begin{aligned}\breve{\boldsymbol{F}}_m^{(\mathrm{T})}(n) &= U\left(n_F^{(m-1)} - n\right)\breve{\boldsymbol{F}}_{m-1}^{(\mathrm{T})}(n)(\boldsymbol{I} + \theta^{-1}\boldsymbol{C}) \\ &\quad + \sum_{l=1}^{\min(n,\, n_F^{(1)})} U\left(n_F^{(m-1)} - n + l\right)\breve{\boldsymbol{F}}_{m-1}^{(\mathrm{T})}(n-l)\breve{\boldsymbol{F}}_1^{(\mathrm{T})}(l),\end{aligned}$$

until $\breve{\boldsymbol{F}}_m^{(\mathrm{T})}(n)$'s satisfy $\sum_{n \le n^*}\breve{\boldsymbol{F}}_m^{(\mathrm{T})}(n)\boldsymbol{e} > (1-\varepsilon_F)^m\,\boldsymbol{e}$ for some n^*. Let $n_F^{(m)} = n^*$ and go to Step (3–c).

Step 4. Set

$$n_A(k) = \max\left(n_F^{(m)}; m = 0, 1, \ldots, m_\gamma(k)\right),$$

and for each $k \in \mathcal{K}$, compute $\breve{\boldsymbol{\Gamma}}_k^{(\mathrm{T})}(n)$ by the following recursion:

$$\breve{\boldsymbol{\Gamma}}_k^{(\mathrm{T})}(0) = \left[\boldsymbol{I} - \boldsymbol{P}_k \otimes \breve{\boldsymbol{A}}_k^{(\mathrm{T})}(0)\right]^{-1},$$

and for $n = 1, 2, \ldots$,

$$\breve{\boldsymbol{\Gamma}}_k^{(\mathrm{T})}(n) = \sum_{l=1}^{n} U\left(n_A(k) - l\right)\breve{\boldsymbol{\Gamma}}_k^{(\mathrm{T})}(n-l)\left[\boldsymbol{P}_k \otimes \breve{\boldsymbol{A}}_k^{(\mathrm{T})}(l)\right]\breve{\boldsymbol{\Gamma}}_k^{(\mathrm{T})}(0),$$

until $\breve{\boldsymbol{\Gamma}}_k^{(\mathrm{T})}(n)$'s satisfy

$$\begin{aligned}\sum_{n=0}^{n_\Gamma(k)}\breve{\boldsymbol{\Gamma}}_k^{(\mathrm{T})}(n)\boldsymbol{e} &> \left\{(\boldsymbol{I} - \boldsymbol{P}_k)^{-1}\,\boldsymbol{e}(M_k)\right\} \otimes \boldsymbol{e}(M) \\ &\quad - \varepsilon\left\{(\boldsymbol{I} - \boldsymbol{P}_k)^{-2}\,\boldsymbol{P}_k\boldsymbol{e}(M_k)\right\} \otimes \boldsymbol{e}(M),\end{aligned}$$

for some integer $n_\Gamma(k)$.

Remark C.1 *The above algorithm ensures that*

$$\sum_{n=0}^{n_A(k)}\breve{\boldsymbol{A}}_k^{(\mathrm{T})}(n)\boldsymbol{e} > (1-\varepsilon)\boldsymbol{e}, \quad \sum_{n=0}^{n_v(k)}\breve{\boldsymbol{v}}_k^{(\mathrm{T})}(n)\boldsymbol{e} > (1-\varepsilon)\lambda_k^{(\mathrm{B})},$$

respectively, where $n_v(k)$ is given by

$$n_v(k) = \max\left(n_F^{(m)}; m = 0, 1, \ldots, m_v(k)\right).$$

Further $\breve{\boldsymbol{\Gamma}}_k^{(\mathrm{T})}(n)$ satisfies

$$(\boldsymbol{\alpha}_k \otimes \boldsymbol{\pi})\sum_{n=0}^{n_\Gamma(k)}\breve{\boldsymbol{\Gamma}}_k^{(\mathrm{T})}(n)\boldsymbol{e} > \mathrm{E}[G_k] - \frac{1}{2}\mathrm{E}[G_k(G_k-1)]\varepsilon.$$

Appendix D

Proof of Lemma 3.2

From the definition (3.18) of $\overline{\boldsymbol{F}}_m(\boldsymbol{n})$, (3.19) is clearly satisfied. (3.20) is proved in the following way. Let $\boldsymbol{N}_{\rm off}(\boldsymbol{n})$ denote an $M_{\rm off} \times M_{\rm off}$ matrix satisfying

$$\sum_{\boldsymbol{n}\in\mathcal{Z}} z_1^{n_1}\cdots z_K^{n_K}\boldsymbol{N}_{\rm off}(\boldsymbol{n}) = \left(-\boldsymbol{C}_{\rm off} - \sum_{k\in\mathcal{K}} \boldsymbol{D}^*_{k,{\rm off}}(z_k)\right)^{-1}. \tag{D.1}$$

From (3.3), (3.18) and (D.1), we have for $m = 1, 2, \ldots$,

$$\begin{aligned}
&\sum_{\boldsymbol{n}\in\mathcal{Z}} z_1^{n_1}\cdots z_K^{n_K}\overline{\boldsymbol{F}}_m(\boldsymbol{n}) \\
&\quad= \sum_{\boldsymbol{n}\in\mathcal{Z}} z_1^{n_1}\cdots z_K^{n_K}\overline{\boldsymbol{F}}_{m-1}(\boldsymbol{n})\left[\boldsymbol{I} + \theta_{\rm on}^{-1}\left\{\boldsymbol{C}_{\rm on} + \sum_{k\in\mathcal{K}}\sum_{n_k=1}^{\infty} z_k^{n_k}\boldsymbol{D}_{k,{\rm on}}(n_k)\right.\right. \\
&\qquad\qquad\left.\left. + \boldsymbol{E}_{\rm on,off}\sum_{\boldsymbol{n}\in\mathcal{Z}} z_1^{n_1}\cdots z_K^{n_K}\boldsymbol{N}_{\rm off}(\boldsymbol{n})\boldsymbol{E}_{\rm off,on}\right\}\right] \\
&\quad= \sum_{\boldsymbol{n}\in\mathcal{Z}} z_1^{n_1}\cdots z_K^{n_K}\overline{\boldsymbol{F}}_{m-1}(\boldsymbol{n})(\boldsymbol{I} + \theta_{\rm on}^{-1}\boldsymbol{C}_{\rm on}) \\
&\qquad+ \sum_{\boldsymbol{n}\in\mathcal{Z}} z_1^{n_1}\cdots z_K^{n_K}\theta_{\rm on}^{-1}\sum_{k\in\mathcal{K}}\sum_{l_k=1}^{n_k}\overline{\boldsymbol{F}}_{m-1}(\boldsymbol{n}-l_k\boldsymbol{e}_k)\boldsymbol{D}_{k,{\rm on}}(l_k) \\
&\qquad+ \sum_{\boldsymbol{n}\in\mathcal{Z}} z_1^{n_1}\cdots z_K^{n_K}\theta_{\rm on}^{-1}\sum_{\boldsymbol{0}\le\boldsymbol{l}\le\boldsymbol{n}}\overline{\boldsymbol{F}}_{m-1}(\boldsymbol{n}-\boldsymbol{l})\boldsymbol{E}_{\rm on,off}\boldsymbol{N}_{\rm off}(\boldsymbol{l})\boldsymbol{E}_{\rm off,on}.
\end{aligned}$$

Comparing coefficient matrices of $z_1^{n_1}\cdots z_K^{n_K}$ on both sides of the above equation, we obtain (3.20). Next, we show (3.21) and (3.22). From (3.3) and (D.1), we have

$$\sum_{\boldsymbol{n}\in\mathcal{Z}} z_1^{n_1}\cdots z_K^{n_K}\boldsymbol{N}_{\rm off}(\boldsymbol{n})\left(-\boldsymbol{C}_{\rm off} - \sum_{k\in\mathcal{K}}\sum_{n_k=1}^{\infty} z_k^{n_k}\boldsymbol{D}_{k,{\rm off}}(n_k)\right) = \boldsymbol{I}.$$

Comparing coefficient matrices of $z_1^{n_1}\cdots z_K^{n_K}$ on both sides of the above equation yields

$$\begin{aligned}
\boldsymbol{N}_{\rm off}(\boldsymbol{0})\,(-\boldsymbol{C}_{\rm off}) &= \boldsymbol{I}, \\
\boldsymbol{N}_{\rm off}(\boldsymbol{n})\,(-\boldsymbol{C}_{\rm off}) &- \sum_{k\in\mathcal{K}}\sum_{l_k=1}^{n_k}\boldsymbol{N}_{\rm off}(\boldsymbol{n}-l_k\boldsymbol{e}_k)\boldsymbol{D}_{k,{\rm off}}(l_k) = \boldsymbol{O},
\end{aligned}$$

from which (3.21) and (3.22) follow. ■

Appendix E

Proof of Theorem 3.5

Post-multiplying both sides of (3.2) by $-\boldsymbol{C}_{\text{off}} - \overline{\boldsymbol{D}}^{*}_{\text{off}}(s)$ and substituting $\theta_{\text{on}} - \theta_{\text{on}} z$ for s, we obtain

$$\sum_{m=0}^{\infty} z^m \boldsymbol{v}^{(m)}_{\text{off}}(\theta_{\text{on}}) \left[-\boldsymbol{C}_{\text{off}} - \sum_{m=0}^{\infty} z^m \boldsymbol{D}^{(m)}_{\text{off}}(\theta_{\text{on}}) \right] = \frac{\sum_{m=0}^{\infty} z^m \boldsymbol{v}^{(m)}_{\text{on}}(\theta_{\text{on}}) \boldsymbol{E}_{\text{on,off}}}{\boldsymbol{\pi}_{\text{on}} \boldsymbol{E}_{\text{on,off}} \boldsymbol{e}} \frac{1}{\overline{I}_{\text{off}}}, \tag{E.1}$$

where we use (3.24) and (3.26). Comparing coefficient vectors of z^m $(m = 0, 1, \ldots)$ on both sides of (E.1) yields

$$\boldsymbol{v}^{(0)}_{\text{off}}(\theta_{\text{on}}) \left[-\boldsymbol{C}_{\text{off}} - \boldsymbol{D}^{(0)}_{\text{off}}(\theta_{\text{on}}) \right] = \frac{\boldsymbol{v}^{(0)}_{\text{on}}(\theta_{\text{on}}) \boldsymbol{E}_{\text{on,off}}}{\boldsymbol{\pi}_{\text{on}} \boldsymbol{E}_{\text{on,off}} \boldsymbol{e}} \frac{1}{\overline{I}_{\text{off}}},$$

and for $m = 1, 2, \ldots$,

$$\boldsymbol{v}^{(m)}_{\text{off}}(\theta_{\text{on}}) \left[-\boldsymbol{C}_{\text{off}} - \boldsymbol{D}^{(0)}_{\text{off}}(\theta_{\text{on}}) \right] - \sum_{l=0}^{m-1} \boldsymbol{v}^{(l)}_{\text{off}}(\theta_{\text{on}}) \boldsymbol{D}^{(m-l)}_{\text{off}}(\theta_{\text{on}}) = \frac{\boldsymbol{v}^{(m)}_{\text{on}}(\theta_{\text{on}}) \boldsymbol{E}_{\text{on,off}}}{\boldsymbol{\pi}_{\text{on}} \boldsymbol{E}_{\text{on,off}} \boldsymbol{e}} \frac{1}{\overline{I}_{\text{off}}}.$$

(3.27) and (3.28) follow the above two equations. ■

Appendix F

Total Queue Length Distribution in a Queue with Service Interruptions

This appendix summarizes the recursions for the total queue length distribution. Because they are readily derived from the results in Section 3.4, we omit the proofs.

We define $\boldsymbol{p}^{(\mathrm{T})}(n)$ $(n = 0, 1, \ldots)$ and $\boldsymbol{q}_k^{(\mathrm{T})}(n)$ $(k \in \mathcal{K}, n = 0, 1, \ldots)$ as the stationary total queue length distributions at a random point in time and at immediately after departures of class k, respectively.

$$\boldsymbol{p}^{(\mathrm{T})}(n) = \sum_{\substack{\boldsymbol{n}\in\mathcal{Z}\\|\boldsymbol{n}|=n}} \boldsymbol{p}(\boldsymbol{n}), \qquad \boldsymbol{q}_k^{(\mathrm{T})}(n) = \sum_{\substack{\boldsymbol{n}\in\mathcal{Z}\\|\boldsymbol{n}|=n}} \boldsymbol{q}_k(\boldsymbol{n}),$$

where $|\boldsymbol{n}| = |n_1| + \cdots + |n_K|$. Since $\boldsymbol{p}^{(\mathrm{T})}(n)$'s can be obtained in terms of $\boldsymbol{q}_k^{(\mathrm{T})}(n)$'s (see Corollary C.1), we hereafter consider the total queue length distribution $\boldsymbol{q}_k^{(\mathrm{T})}(n)$ immediately after departures.

We first introduce the following notations: For $k \in \mathcal{K}$, $\xi = \mathrm{on}, \mathrm{off}$, $n = 0, 1, \ldots$ and $m = 0, 1, \ldots$,

$$\begin{aligned}
\boldsymbol{q}_{k,\xi}^{(\mathrm{T})}(n) &= \sum_{\substack{\boldsymbol{n}\in\mathcal{Z}\\|\boldsymbol{n}|=n}} \boldsymbol{q}_{k,\xi}(\boldsymbol{n}), \quad \boldsymbol{\Gamma}_{k,\xi}^{(\mathrm{T})}(n) = \sum_{\substack{\boldsymbol{n}\in\mathcal{Z}\\|\boldsymbol{n}|=n}} \boldsymbol{\Gamma}_{k,\xi}(\boldsymbol{n}), \quad \boldsymbol{A}_{k,\xi}^{(\mathrm{T})}(n) = \sum_{\substack{\boldsymbol{n}\in\mathcal{Z}\\|\boldsymbol{n}|=n}} \boldsymbol{A}_{k,\xi}(\boldsymbol{n}),\\
\boldsymbol{v}_{k,\xi}^{(\mathrm{T})}(n) &= \sum_{\substack{\boldsymbol{n}\in\mathcal{Z}\\|\boldsymbol{n}|=n}} \boldsymbol{v}_{k,\xi}(\boldsymbol{n}), \quad \overline{\boldsymbol{F}}_m^{(\mathrm{T})}(n) = \sum_{\substack{\boldsymbol{n}\in\mathcal{Z}\\|\boldsymbol{n}|=n}} \overline{\boldsymbol{F}}_m(\boldsymbol{n}).
\end{aligned}$$

From Theorem 3.2, we have the following corollary.

Corollary F.1 *Under Assumption 3.1, the* $\boldsymbol{q}_k^{(\mathrm{T})}(n)$ *(*$k \in \mathcal{K}, n = 0, 1, \ldots$*) is given by*

$$\boldsymbol{q}_k^{(\mathrm{T})}(n) = \left(\frac{r_{\mathrm{on}}\lambda_{k,\mathrm{on}}}{\lambda_k} \boldsymbol{q}_{k,\mathrm{on}}^{(\mathrm{T})}(n) + \frac{r_{\mathrm{off}}\lambda_{k,\mathrm{off}}}{\lambda_k} \boldsymbol{q}_{k,\mathrm{off}}^{(\mathrm{T})}(n), 0, \ldots, 0 \right),$$

where the $\boldsymbol{q}_{k,\xi}^{(\mathrm{T})}(n)$ *(*$k \in \mathcal{K}$*,* $\xi = on, off$*,* $n = 0, 1, \ldots$*) is given by*

$$\begin{aligned}
\boldsymbol{q}_{k,\xi}^{(\mathrm{T})}(n) &= \frac{1}{\lambda_{k,\xi}} \sum_{\substack{m_1+m_2+m_3\\+m_4=n}} \boldsymbol{v}_{k,\xi}^{(\mathrm{T})}(m_1)[\boldsymbol{\alpha}_{k,\xi} \otimes \boldsymbol{A}_{k,\xi}^{(\mathrm{T})}(m_2)]\boldsymbol{\Gamma}_{k,\xi}^{(\mathrm{T})}(m_3)\\
&\qquad \cdot \left[\left\{ \boldsymbol{P}_{k,\xi}^{m_4}(\boldsymbol{I} - \boldsymbol{P}_{k,\xi})\boldsymbol{e} \right\} \otimes \boldsymbol{I}(M_{\mathrm{on}}) \right],
\end{aligned}$$

if $\lambda_{k,\xi} > 0$*, and otherwise* $\boldsymbol{q}_{k,\xi}^{(\mathrm{T})}(n) = \boldsymbol{0}$*.*

The following corollary for the $\mathbf{\Gamma}_{k,\xi}^{(\mathrm{T})}(n)$ can be readily obtained from Lemma 3.1.

Corollary F.2 *The $\mathbf{\Gamma}_{k,\xi}^{(\mathrm{T})}(n)$ ($k \in \mathcal{K}$, $\xi = on, off$, $n = 0, 1, \ldots$) is determined by the following recursion:*

$$\begin{aligned}
\mathbf{\Gamma}_{k,\xi}^{(\mathrm{T})}(0) &= \left[\boldsymbol{I} - \boldsymbol{P}_{k,\xi} \otimes \boldsymbol{A}_{k,\xi}^{(\mathrm{T})}(0)\right]^{-1}, \\
\mathbf{\Gamma}_{k,\xi}^{(\mathrm{T})}(n) &= \sum_{l=1}^{n} \mathbf{\Gamma}_{k,\xi}^{(\mathrm{T})}(n-l) \left[\boldsymbol{P}_{k,\xi} \otimes \boldsymbol{A}_{k,\xi}^{(\mathrm{T})}(l)\right] \mathbf{\Gamma}_{k,\xi}^{(\mathrm{T})}(0), \qquad n = 1, 2, \ldots.
\end{aligned}$$

As for the $\boldsymbol{A}_{k,\xi}^{(\mathrm{T})}(n)$, the following recursions can be obtained from Lemma 3.2 and Theorem 3.3.

Corollary F.3 *The $\boldsymbol{A}_{k,\xi}^{(\mathrm{T})}(n)$ is given by*

$$\boldsymbol{A}_{k,\xi}^{(\mathrm{T})}(n) = \sum_{m=0}^{\infty} \gamma_{k,\xi}^{(m)}(\theta_{\mathrm{on}}) \overline{\boldsymbol{F}}_{m}^{(\mathrm{T})}(n), \qquad k \in \mathcal{K},\ \xi = \mathrm{on}, \mathrm{off},\ n = 0, 1, \ldots,$$

where the $\overline{\boldsymbol{F}}_{m}^{(\mathrm{T})}(n)$ is recursively determined by

$$\overline{\boldsymbol{F}}_{0}^{(\mathrm{T})}(n) = \begin{cases} \boldsymbol{I}, & \text{if } n = 0, \\ \boldsymbol{O}, & \text{otherwise}, \end{cases}$$

and for $m = 1, 2, \ldots,$

$$\begin{aligned}
\overline{\boldsymbol{F}}_{m}^{(\mathrm{T})}(n) &= \overline{\boldsymbol{F}}_{m-1}^{(\mathrm{T})}(n)(\boldsymbol{I} + \theta_{\mathrm{on}}^{-1}\boldsymbol{C}_{\mathrm{on}}) + \theta_{\mathrm{on}}^{-1} \sum_{l=1}^{n} \overline{\boldsymbol{F}}_{m-1}^{(\mathrm{T})}(n-l) \sum_{k \in \mathcal{K}} \boldsymbol{D}_{k,\mathrm{on}}(l) \\
&\qquad + \theta_{\mathrm{on}}^{-1} \left[\sum_{l=0}^{n} \overline{\boldsymbol{F}}_{m-1}^{(\mathrm{T})}(n-l) \boldsymbol{E}_{\mathrm{on,off}} \boldsymbol{N}_{\mathrm{off}}^{(\mathrm{T})}(l)\right] \boldsymbol{E}_{\mathrm{off,on}}, \qquad n = 0, 1, \ldots,
\end{aligned}$$

with the $\boldsymbol{N}_{\mathrm{off}}^{(\mathrm{T})}(n)$, which is given by the following recursion:

$$\begin{aligned}
\boldsymbol{N}_{\mathrm{off}}^{(\mathrm{T})}(0) &= (-\boldsymbol{C}_{\mathrm{off}})^{-1}, \\
\boldsymbol{N}_{\mathrm{off}}^{(\mathrm{T})}(n) &= \left[\sum_{l=1}^{n} \boldsymbol{N}_{\mathrm{off}}^{(\mathrm{T})}(n-l) \sum_{k \in \mathcal{K}} \boldsymbol{D}_{k,\mathrm{off}}(l)\right] \boldsymbol{N}_{\mathrm{off}}^{(\mathrm{T})}(0), \qquad n = 1, 2, \ldots.
\end{aligned}$$

Finally, from Theorem 3.4, we obtain the following result.

Corollary F.4 *The $\boldsymbol{v}_{k,\mathrm{on}}^{(\mathrm{T})}(n)$ ($k \in \mathcal{K}$) and the $\boldsymbol{v}_{k,\mathrm{off}}^{(\mathrm{T})}(n)$ ($k \in \mathcal{K}$) are determined by*

$$\begin{aligned}
\boldsymbol{v}_{k,\mathrm{on}}^{(\mathrm{T})}(n) &= \sum_{m=0}^{\infty} \boldsymbol{v}_{\mathrm{on}}^{(m)}(\theta_{\mathrm{on}}) \boldsymbol{D}_{k,\mathrm{on}} \overline{\boldsymbol{F}}_{m}^{(\mathrm{T})}(n), \qquad n = 0, 1, \ldots, \\
\boldsymbol{v}_{k,\mathrm{off}}^{(\mathrm{T})}(n) &= \sum_{m=0}^{\infty} \boldsymbol{v}_{\mathrm{off}}^{(m)}(\theta_{\mathrm{on}}) \boldsymbol{D}_{k,\mathrm{off}} \sum_{l=0}^{n} \boldsymbol{N}_{\mathrm{off}}^{(\mathrm{T})}(n-l) \boldsymbol{E}_{\mathrm{off,on}} \overline{\boldsymbol{F}}_{m}^{(\mathrm{T})}(l), \qquad n = 0, 1, \ldots,
\end{aligned}$$

respectively.

Appendix G

Proof of Lemma 4.1

By mathematical induction, we prove Lemma 4.1. We first consider the case of $\nu = 2$. We define $\boldsymbol{G}_1(t, \boldsymbol{U}_1, \boldsymbol{U}_2, \boldsymbol{V}_1)$ as

$$\boldsymbol{G}_1(t, \boldsymbol{U}_1, \boldsymbol{U}_2, \boldsymbol{V}_1) = \int_0^t du\, e^{\boldsymbol{U}_1 u} \boldsymbol{V}_1 e^{\boldsymbol{U}_2 (t-u)}. \tag{G.1}$$

By differentiating (G.1) with respect to t, we have

$$\frac{d}{dt}\boldsymbol{G}_1(t, \boldsymbol{U}_1, \boldsymbol{U}_2, \boldsymbol{V}_1) - \boldsymbol{G}_1(t, \boldsymbol{U}_1, \boldsymbol{U}_2, \boldsymbol{V}_1)\boldsymbol{U}_2 = e^{\boldsymbol{U}_1 t}\boldsymbol{V}_1. \tag{G.2}$$

On the other hand, by straightforward calculation based on the definition of matrix exponential, we obtain

$$\exp\left[\begin{pmatrix} \boldsymbol{U}_1 & \boldsymbol{V}_1 \\ \boldsymbol{O} & \boldsymbol{U}_2 \end{pmatrix} t\right] = \begin{bmatrix} e^{\boldsymbol{U}_1 t} & \widetilde{\boldsymbol{G}}_1(t, \boldsymbol{U}_1, \boldsymbol{U}_2, \boldsymbol{V}_1) \\ \boldsymbol{O} & e^{\boldsymbol{U}_2 t} \end{bmatrix}, \tag{G.3}$$

where

$$\widetilde{\boldsymbol{G}}_1(t, \boldsymbol{U}_1, \boldsymbol{U}_2, \boldsymbol{V}_1) = \sum_{n=1}^{\infty} \frac{t^n}{n!} \sum_{j=0}^{n-1} \boldsymbol{U}_1^j \boldsymbol{V}_1 \boldsymbol{U}_2^{n-1-j}.$$

It is easy to verify that $\widetilde{\boldsymbol{G}}_1(t, \boldsymbol{U}_1, \boldsymbol{U}_2, \boldsymbol{V}_1)$ satisfies

$$\frac{d}{dt}\widetilde{\boldsymbol{G}}_1(t, \boldsymbol{U}_1, \boldsymbol{U}_2, \boldsymbol{V}_1) - \widetilde{\boldsymbol{G}}_1(t, \boldsymbol{U}_1, \boldsymbol{U}_2, \boldsymbol{V}_1)\boldsymbol{U}_2 = e^{\boldsymbol{U}_1 t}\boldsymbol{V}_1. \tag{G.4}$$

Taking the difference between (G.2) and (G.4), we have

$$\begin{aligned} &\frac{d}{dt}\left[\boldsymbol{G}_1(t, \boldsymbol{U}_1, \boldsymbol{U}_2, \boldsymbol{V}_1) - \widetilde{\boldsymbol{G}}_1(t, \boldsymbol{U}_1, \boldsymbol{U}_2, \boldsymbol{V}_1)\right] \\ &\qquad = \left[\boldsymbol{G}_1(t, \boldsymbol{U}_1, \boldsymbol{U}_2, \boldsymbol{V}_1) - \widetilde{\boldsymbol{G}}_1(t, \boldsymbol{U}_1, \boldsymbol{U}_2, \boldsymbol{V}_1)\right]\boldsymbol{U}_2, \end{aligned}$$

from which it follows that

$$\boldsymbol{G}_1(t, \boldsymbol{U}_1, \boldsymbol{U}_2, \boldsymbol{V}_1) - \widetilde{\boldsymbol{G}}_1(t, \boldsymbol{U}_1, \boldsymbol{U}_2, \boldsymbol{V}_1) = \boldsymbol{K} e^{\boldsymbol{U}_2 t},$$

for some constant matrix $\boldsymbol{K}$. Note here that

$$\widetilde{\boldsymbol{G}}_1(0, \boldsymbol{U}_1, \boldsymbol{U}_2, \boldsymbol{V}_1) = \boldsymbol{G}_1(0, \boldsymbol{U}_1, \boldsymbol{U}_2, \boldsymbol{V}_1) = \boldsymbol{O},$$

so that $\boldsymbol{K} = \boldsymbol{O}$. Thus we have

$$\widetilde{\boldsymbol{G}}_1(t, \boldsymbol{U}_1, \boldsymbol{U}_2, \boldsymbol{V}_1) = \boldsymbol{G}_1(t, \boldsymbol{U}_1, \boldsymbol{U}_2, \boldsymbol{V}_1), \quad t \geq 0,$$

and therefore (G.3) is rewritten to be

$$\exp\left[\begin{pmatrix} \boldsymbol{U}_1 & \boldsymbol{V}_1 \\ \boldsymbol{O} & \boldsymbol{U}_2 \end{pmatrix} t\right] = \begin{bmatrix} e^{\boldsymbol{U}_1 t} & \int_0^t du\, e^{\boldsymbol{U}_1 u} \boldsymbol{V}_1 e^{\boldsymbol{U}_2 (t-u)} \\ \boldsymbol{O} & e^{\boldsymbol{U}_2 t} \end{bmatrix}. \tag{G.5}$$

As a result, we have

$$\begin{aligned} \boldsymbol{V}_0 \int_0^t du\, e^{\boldsymbol{U}_1 u} \boldsymbol{V}_1 e^{\boldsymbol{U}_2 (t-u)} \boldsymbol{V}_2 &= \boldsymbol{V}_0 \begin{bmatrix} \boldsymbol{I} & \boldsymbol{O} \end{bmatrix} \exp\left[\begin{pmatrix} \boldsymbol{U}_1 & \boldsymbol{V}_1 \\ \boldsymbol{O} & \boldsymbol{U}_2 \end{pmatrix} t\right] \begin{bmatrix} \boldsymbol{O} \\ \boldsymbol{I} \end{bmatrix} \boldsymbol{V}_2 \\ &= \begin{bmatrix} \boldsymbol{V}_0 & \boldsymbol{O} \end{bmatrix} \exp\left[\begin{pmatrix} \boldsymbol{U}_1 & \boldsymbol{V}_1 \\ \boldsymbol{O} & \boldsymbol{U}_2 \end{pmatrix} t\right] \begin{bmatrix} \boldsymbol{O} \\ \boldsymbol{V}_2 \end{bmatrix}, \end{aligned} \tag{G.6}$$

which shows (4.32) holds for $\nu = 2$.

Suppose that (4.32) holds for some $\nu = m$ ($m \geq 2$), i.e.,

$$\begin{aligned} &\boldsymbol{V}_0 \int_0^t du_{m-1} \int_0^{u_{m-1}} du_{m-2} \cdots \int_0^{u_2} du_1 e^{\boldsymbol{U}_1 u_1} \boldsymbol{V}_1 e^{\boldsymbol{U}_2 (u_2 - u_1)} \boldsymbol{V}_2 \\ &\qquad \cdots\cdots e^{\boldsymbol{U}_{m-1}(u_{m-1} - u_{m-2})} \boldsymbol{V}_{m-1} e^{\boldsymbol{U}_m (t - u_{m-1})} \boldsymbol{V}_m \\ &= \begin{bmatrix} \boldsymbol{V}_0 & \boldsymbol{O} & \cdots & \boldsymbol{O} \end{bmatrix} \exp\left[\begin{pmatrix} \boldsymbol{U}_1 & \boldsymbol{V}_1 & \boldsymbol{O} & \cdots & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{U}_2 & \boldsymbol{V}_2 & & \vdots \\ \vdots & & \ddots & \ddots & \boldsymbol{O} \\ \boldsymbol{O} & \cdots & \boldsymbol{O} & \boldsymbol{U}_{m-1} & \boldsymbol{V}_{m-1} \\ \boldsymbol{O} & \boldsymbol{O} & \cdots & \boldsymbol{O} & \boldsymbol{U}_m \end{pmatrix} t\right] \begin{bmatrix} \boldsymbol{O} \\ \vdots \\ \boldsymbol{O} \\ \boldsymbol{V}_m \end{bmatrix}. \end{aligned}$$

We then have

$$\begin{aligned} &\int_0^t du_m \Bigg[\boldsymbol{V}_0 \int_0^{u_m} du_{m-1} \int_0^{u_{m-1}} du_{m-2} \cdots \int_0^{u_2} du_1 e^{\boldsymbol{U}_1 u_1} \boldsymbol{V}_1 e^{\boldsymbol{U}_2 (u_2 - u_1)} \boldsymbol{V}_2 \\ &\qquad \cdots\cdots e^{\boldsymbol{U}_m (u_m - u_{m-1})} \boldsymbol{V}_m \Bigg] e^{\boldsymbol{U}_{m+1}(t - u_m)} \boldsymbol{V}_{m+1} \\ &= \begin{bmatrix} \boldsymbol{V}_0 & \boldsymbol{O} & \cdots & \boldsymbol{O} \end{bmatrix} \\ &\quad \cdot \int_0^t du_m \exp\left[\begin{pmatrix} \boldsymbol{U}_1 & \boldsymbol{V}_1 & \boldsymbol{O} & \cdots & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{U}_2 & \boldsymbol{V}_2 & & \vdots \\ \vdots & & \ddots & \ddots & \boldsymbol{O} \\ \boldsymbol{O} & \cdots & \boldsymbol{O} & \boldsymbol{U}_{m-1} & \boldsymbol{V}_{m-1} \\ \boldsymbol{O} & \boldsymbol{O} & \cdots & \boldsymbol{O} & \boldsymbol{U}_m \end{pmatrix} u_m\right] \begin{bmatrix} \boldsymbol{O} \\ \vdots \\ \boldsymbol{O} \\ \boldsymbol{V}_m \end{bmatrix} e^{\boldsymbol{U}_{m+1}(t - u_m)} \boldsymbol{V}_{m+1}. \end{aligned} \tag{G.7}$$

Thus, applying (G.6) to (G.7), we obtain

$$\int_0^t du_m \Bigg[\boldsymbol{V}_0 \int_0^{u_m} du_{m-1} \int_0^{u_{m-1}} du_{m-2} \cdots \int_0^{u_2} du_1 e^{\boldsymbol{U}_1 u_1} \boldsymbol{V}_1 e^{\boldsymbol{U}_2 (u_2 - u_1)} \boldsymbol{V}_2$$

$$\cdots\cdots e^{\boldsymbol{U}_m(u_m-u_{m-1})}\boldsymbol{V}_m\Big]e^{\boldsymbol{U}_{m+1}(t-u_m)}\boldsymbol{V}_{m+1}$$

$$= \left[\begin{array}{c|c} \boldsymbol{V}_0 \;\; \boldsymbol{O} \;\; \cdots \;\; \boldsymbol{O} & \boldsymbol{O} \end{array}\right]$$

$$\cdot\exp\left[\left(\begin{array}{cccccc|c} \boldsymbol{U}_1 & \boldsymbol{V}_1 & \boldsymbol{O} & \cdots & \boldsymbol{O} & & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{U}_2 & \boldsymbol{V}_2 & & \vdots & & \vdots \\ \vdots & & \ddots & \ddots & \boldsymbol{O} & & \boldsymbol{O} \\ \boldsymbol{O} & \cdots & \boldsymbol{O} & \boldsymbol{U}_{m-1} & \boldsymbol{V}_{m-1} & & \boldsymbol{V}_m \\ \boldsymbol{O} & \boldsymbol{O} & \cdots & \boldsymbol{O} & \boldsymbol{U}_m & & \\ \hline \multicolumn{5}{c}{\boldsymbol{O}} & & \boldsymbol{U}_{m+1} \end{array}\right)t\right]\left[\begin{array}{c} \boldsymbol{O} \\ \vdots \\ \boldsymbol{O} \\ \hline \boldsymbol{V}_{m+1} \end{array}\right],$$

which shows that (4.32) holds for $\nu = m+1$, too. ■

Bibliography

[Asmu91] S. Asmussen, Ladder Heights and the Markov-Modulated M/G/1 Queue, *Stochastic Processes and their Applications*, 37 (1991) 313–326.

[Asmu93] S. Asmussen and G. Koole, Marked Point Processes as Limits of Markovian Arrival Streams, *Journal of Applied Probability*, 30 (1993) 365–372.

[Asmu03] S. Asmussen, *Applied Probability and Queues*, 2nd ed. (Springer Verlag, New York, 2003).

[Bell97] R. Bellman, *Introduction to Matrix Analysis (Classics in Applied Mathematics)*, 2nd ed. (SIAM, Philadelphia, 1997).

[Bors03] S. C. Borst, O. J. Boxma, J. A. Morrison and R. Núñez Queija, The Equivalence between Processor Sharing and Service in Random Order, *Operations Research Letters*, 31 (2003) 254–262.

[Coff70] E. G. Coffman, Jr., R. R. Muntz and H. Trotter, Waiting Time Distributions for Processor-Sharing Systems, *Journal of the Association for Computing Machinery*, 17 (1970) 123–130.

[Cohe84] J. W. Cohen, On Processor Sharing and Random Service, *Journal of Applied Probability*, 21 (1984) 937.

[Eick93] S. G. Eick, W. A. Massey and W. Whitt, $M_t/G/\infty$ Queues with Sinusoidal Arrival Rates, *Management Science*, 39 (1993) 241–252.

[Fede86] A. Federgruen and L. Green, Queueing Systems with Service Interruptions, *Operations Research*, 34 (1986) 752–768.

[Fede88] A. Federgruen and L. Green, Queueing Systems with Service Interruptions II, *Naval Research Logistics*, 35 (1988) 345–358.

[Fox88] B. L. Fox and P. W. Glynn, Computing Poisson Probabilities, *Communications of the ACM*, 31 (1988) 440–445.

[Gris94] S. Grishechkin, GI/G/1 Processor Sharing Queue in Heavy Traffic, *Advances in Applied Probability*, 26 (1994) 539–555.

[Guil01] F. Guillemin and J. Boyer, Analysis of the M/M/1 Queue with Processor Sharing via Spectral Theory, *Queueing Systems*, 39 (2001) 377–397.

[He96] Q.-M. He, Queues with Marked Customers, *Advances in Applied Probability*, 28 (1996) 567–587.

[He98] Q.-M. He and A. S. Alfa, The MMAP[K]/PH[K]/1 Queues with a Last-Come-First-Served Preemptive Service Discipline, *Queueing Systems*, 29 (1998) 269–291.

[He00a] Q.-M. He, Classification of Markov Processes of M/G/1 Type with a Tree Structure and Its Applications to Queueing Models, *Operations Research Letters*, 26 (2000) 67–80.

[He00b] Q.-M. He, Classification of Markov Processes of Matrix M/G/1 Type with a Tree Structure and Its Applications to the MMAP[K]/G[K]/1 Queues, *Stochastic Models*, 16 (2000) 407–433.

[He01] Q.-M. He, The Versatility of MMAP[K] and the MMAP[K]/G[K]/1 Queue, *Queueing Systems*, 38 (2001) 397–418.

[He03a] Q.-M. He, A Fixed Point Approach to the Classification of Markov Chains with a Tree Structure, *Stochastic Models*, 19 (2003) 75–111.

[He03b] Q.-M. He and H. Li, Stability Conditions of the MMAP[K]/G[K]/1/LCFS Preemptive Repeat Queue, *Queueing Systems*, 44 (2003) 137–160.

[Holm82] D. F. Holman, M. L. Chaudhry and B. R. K. Kashyap, On the Number in the System $GI^X/M/\infty$, *Sankhyā : The Indian Journal of Statistics*, 44, Series A, Part 1 (1982) 294–297.

[Holm83] D. F. Holman, M. L. Chaudhry and B. R. K. Kashyap, On the Service System $M^X/G/\infty$, *European Journal of Operational Research*, 13 (1983) 142–145.

[Jage91] D. L. Jagerman and B. Sengupta, The GI/M/1 Processor-Sharing Queue and Its Heavy Traffic Analysis, *Stochastic Models*, 7 (1991) 379–395.

[Lee00] G. Lee and J. Jeon, A New Approach to an N/G/1 Queue, *Queueing Systems*, 35 (200) 317–322.

[Liu90] L. Liu, B. R. K. Kashyap and J. G. C. Templeton, On the $GI^X/G/\infty$ System, *Journal of Applied Probability*, 27 (1990) 671–683.

[Liu91] L. Liu and J. G. C. Templeton, The $GR^{X_n}/G_n/\infty$ System: System Size, *Queueing Systems*, 8 (1991) 323–356.

[Loyn62] R. M. Loynes, The Stability of a Queue with Non-Independent Inter-Arrival and Service Times, *Proceedings of the Cambridge Philosophical Society*, 58 (1962) 497–520.

[Luca90] D. M. Lucantoni, K. S. Meier-Hellstern and M. F. Neuts, A Single-Server Queue with Server Vacations and a Class of Non-Renewal Arrival Processes, *Advances in Applied Probability*, 22 (1990) 676–705.

[Luca91] D. M. Lucantoni, New Results on the Single Server Queue with a Batch Markovian Arrival Process, *Stochastic Models*, 7 (1991) 1–46.

[Luca93] D. M. Lucantoni, The BMAP/G/1 Queue: A Tutorial, *Models and Techniques for Performance Evaluation of Computer and Communication Systems*, L. Donatiello and R. Nelson, eds., Springer Verlag (1993) 330–358.

[Mach86] F. Machihara, An Infinitely-Many-Server Queue Having Markov Renewal Arrivals and Hyperexponential Service Times, *Journal of the Operations Research Society of Japan*, 29 (1986) 338–351

[Mach93] F. Machihara, On the Queue with PH-Markov Renewal Preemptions, *Journal of the Operations Research Society of Japan*, 36 (1993) 13–28.

[Mach99] F. Machihara, A BMAP/SM/1 Queue with Service Times Depending on the Arrival Process, *Queueing Systems*, 33 (1999) 277–291.

[Masu02] H. Masuyama and T. Takine, Analysis of an Infinite-Server Queue with Batch Markovian Arrival Streams, *Queueing Systems*, 42 (2002) 269-296.

[Masu03a] H. Masuyama and T. Takine, Analysis and Computation of the Joint Queue Length Distribution in a FIFO Single-Server Queue with Multiple Batch Markovian Arrival Streams, *Stochastic Models*, 19 (2003) 349–381.

[Masu03b] H. Masuyama and T. Takine, Sojourn Time Distribution in a MAP/M/1 Processor-Sharing Queue, *Operations Research Letters*, 31 (2003) 406-412.

[Masu03c] H. Masuyama and T. Takine, Stationary Queue Length in a FIFO Single Server Queue with Service Interruptions and Multiple Batch Markovian Arrival Streams, *Journal of the Operations Research Society of Japan*, 46 (2003) 319–341.

[Morr85] J. A. Morrison, Response-Time Distribution for a Processor-Sharing System, *SIAM Journal on Applied Mathematics*, 45 (1985) 152–167.

[Neut79] M. F. Neuts, A Versatile Markovian Point Process, *Journal of Applied Probability*, 16 (1979) 764–779.

[Neut81] M. F. Neuts, *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach* (Johns Hopkins University Press, Baltimore, 1981).

[Neut89] M. F. Neuts, *Structured Stochastic Matrices of M/G/1 Type and Their Applications* (Marcel Dekker, New York, 1989).

[Ott84] T. J. Ott, The Sojourn-Time Distribution in the M/G/1 Queue with Processor Sharing, *Journal of Applied Probability*, 21 (1984) 360–378.

[Rama78] V. Ramaswami, The N/G/∞ Queue, Technical Report, Department of Mathematics, Drexel University, Philadelphia (1978).

[Rama80] V. Ramaswami and M. F. Neuts, Some Explicit Formulas and Computational Methods for Infinite-Server Queues with Phase-Type Arrivals, *Journal of Applied Probability*, 17 (1980) 498–514.

[Rama84] V. Ramaswami, The Sojourn Time in the GI/M/1 Queue with Processor Sharing, *Journal of Applied Probability*, 21 (1984) 437–442.

[Rama88] V. Ramaswami, A Stable Recursion for the Steady State Vector in Markov Chains of M/G/1 Type, *Stochastic Models*, 4 (1988) 183–188.

[Regt86] G. J. K. Regterschot and J. H. A. de Smit, The Queue M|G|1 with Markov Modulated Arrivals and Services, *Mathematics of Operations Research*, 11 (1986) 465–483.

[Sche90] H. Schellhaas, On Ramaswami's Algorithm for the Computation of the Steady State Vector in Markov Chains of M/G/1-Type, *Stochastic Models*, 6 (1990) 541–550.

[Seng85] B. Sengupta and D. L. Jagerman, A Conditional Response Time of the M/M/1 Processor-Sharing Queue, *AT & T Technical Journal*, 64 (1985) 409–421.

[Seng89] B. Sengupta, Markov Processes Whose Steady State Distribution Is Matrix-Exponential with an Application to the GI/PH/1 Queue, *Advances in Applied Probability*, 21 (1989) 159–180.

[Seng90] B. Sengupta, A Queue with Service Interruptions in an Alternating Random Environment, *Operations Research*, 38 (1990) 308–318.

[Seng92] B. Sengupta, An Approximation for the Sojourn-Time Distribution for the GI/G/1 Processor-Sharing Queue, *Stochastic Models*, 8 (1992) 35–57.

[Shan66] D. N. Shanbhag, On Infinite Server Queues with Batch Arrivals, *Journal of Applied Probability*, 3 (1966) 274–279.

[Smit72] W. Smith, The Infinitely-Many-Server Queue with Semi-Markovian Arrivals and Customer-Dependent Exponential Service Times, *Operations Research*, 22 (1972) 907–912.

[Taka62] L. Takács, *Introduction to the Theory of Queues* (Oxford University Press, New York, 1962).

[Taki94a] T. Takine and T. Hasegawa, The Workload in the MAP/G/1 Queue with State-Dependent Services: Its Application to a Queue with Preemptive Resume Priority, *Stochastic Models*, 10 (1994) 183–204.

[Taki94b] T. Takine, Y. Matsumoto, T. Suda and T. Hasegawa, Mean Waiting Times in Nonpreemptive Priority Queues with Markovian Arrival and I.I.D. Service Processes, *Performance Evaluation*, 20 (1994) 131–149.

[Taki95] T. Takine, B. Sengupta and R. W. Yeung, A Generalization of the Matrix M/G/1 Paradigm for Markov Chains with a Tree Structure, *Stochastic Models*, 11 (1995) 411–421.

[Taki96] T. Takine, A Continuous Version of Matrix-Analytic Methods with the Skip-Free to the Left Property, *Stochastic Models*, 12 (1996) 673–682.

[Taki97] T. Takine and B. Sengupta, A Single Server Queue with Service Interruptions, *Queueing Systems*, 26 (1997) 285-300.

[Taki99] T. Takine, The Nonpreemptive Priority MAP/G/1 Queue, *Operations Research*, 47 (1999) 917–927.

[Taki00] T. Takine, A New Recursion for the Queue Length Distribution in the Stationary BMAP/G/1 Queue, *Stochastic Models*, 16 (2000) 335–341.

[Taki01a] T. Takine, Distributional Form of Little's Law for FIFO Queues with Multiple Markovian Arrival Streams and Its Application to Queues with Vacations, *Queueing Systems*, 37 (2001) 31–63.

[Taki01b] T. Takine, Queue Length Distribution in a FIFO Single-Server Queue with Multiple Arrival Streams Having Different Service Time Distributions, *Queueing Systems*, 39 (2001) 349–375.

[Taki01c] T. Takine, A Recent Progress in Algorithmic Analysis of FIFO Queues with Markovian Arrival Streams, *Journal of the Korean Mathematical Society*, 38 (2001) 807–842.

[Taki02] T. Takine, Matrix Product-Form Solution for an LCFS-PR Single-Server Queue with Multiple Arrival Streams Governed by a Markov Chain, *Queueing Systems*, 42 (2002) 131–151.

[Tijm94] H. C. Tijms, *Stochastic Models: An Algorithmic Approach* (John Wiley & Sons, Chichester, 1994).

[Yash83] S. F. Yashkov, A Derivation of Response Time Distribution for a M/G/1 Processor-Sharing Queue, *Problems of Control and Information Theory*, 12 (1983) 133–148.

[Yeun94] R. W. Yeung and B. Sengupta, Matrix Product-Form Solutions for Markov Chains with a Tree Structure, *Advances in Applied Probability*, 26 (1994) 965–987.

[Zhu91] Y. Zhu and N. U. Prabhu, Markov-Modulated PH/G/1 Queueing Systems, *Queueing Systems*, 9 (1991) 313–322.