Analysis of Queues with Markovian Service Processes*

Toshihisa OZAWA
Department of Business Administration, Komazawa University
1-23-1 Komazawa, Setagaya-ku, Tokyo 154-8525, Japan
E-mail: toshi@komazawa-u.ac.jp

Abstract

Markovian service process (MSP) is a model similar to the Markovian arrival process (MAP), where arrivals are replaced with service completions. The MSP can represent various queueing models such as vacation models, N-policy models and exceptional service models. We analyze MAP/MSP/1 queues and obtain a new sort of matrix-type factorization of the vector generating function for the stationary queue length. The MAP/MSP/1 queue is a very tractable model since its behavior is represented as a quasi-birth-and-death process.

1 Introduction

In the area of queueing theory, various service disciplines have been proposed for analyzing computer communications systems, production systems and other kinds of systems. Typical examples are vacation [5], N-policy [10] and exceptional service [1, 12, 21]. According to the disciplines, a server is engaged in two kinds of work: one is ordinary service for customers and the other is secondary work corresponding to vacation or exceptional service, for example. We represent the behavior of the server as a continuous-time Markov chain whose state indicates which kind of work the server is engaged in. This service process is similar to a Markovian arrival process (MAP) [15], where arrivals are replaced with service completions. We refer to the service process as a Markovian service process (MSP) and study MAP/MSP/1 queues. In Refs [19, 20], the idea of using two types of states (phases) was used in analyzing queueing models with vacations. This idea is extended to a general service discipline in our model. The behavior of a MAP/MSP/1 queue is represented as a quasi-birth-and-death (QBD) process, which is analyzed by using matrix analytic methods [16].

One of the most interesting properties in the M/G/1 queue with multiple vacations is the stochastic decomposition for the stationary queue length [5]; it is the property that the stationary queue length is distributed as the independent sum of the stationary queue length in an M/G/1 queue without vacations and the number of arrivals during a residual vacation time. (The property of stochastic decomposition also holds true for other random variables such as the stationary waiting time [4, 5]. However, we will focus only on the stationary queue length in this paper.) The aim of studying such a stochastic decomposition is to understand stochastic characteristics of the behavior of queueing models, and this understanding is expected to help analyze the queueing models. For the purpose, appropriate stochastic interpretations should be given to the obtained stochastic decompositions.

The property of stochastic decomposition has been extended to some directions. Reference [9] deals with an extended version of ordinary vacation, called generalized vacation. The class

*This manuscript is a draft version of "T. Ozawa, Analysis of queues with Markovian Service Processes, Stochastic Models, 20(4) (2004), 391-413."

of generalized vacation includes almost all nonpreemptive-type service disciplines forcing the server to do other work than service for customers. In the M/G/1 queue with generalized vacations, the stationary queue length is distributed as the independent sum of the stationary queue length in an M/G/1 queue without vacations and the number of customers in the system at an arbitrary time during the vacation. References [7, 15] show that the vector generating function (v.g.f.) for the stationary queue length in the MAP/G/1 queue with multiple vacations is given by the product of the v.g.f. for the stationary queue length in a MAP/G/1 queue without vacations and the matrix generating function (m.g.f.) for the number of arrivals during a residual vacation time. This result relies on a certain characteristic of the multiple-vacation model, and in general it seems not to hold in MAP/G/1 queues with other kinds of service disciplines such as the MAP/G/1 queue with setup times. Reference [14] shows that a sort of matrix-type factorization holds in the MAP/G/1 queue under multiple and single vacations with N-policy. In this model, the v.g.f. for the stationary queue length is factorized into two parts: the v.g.f. for the queue length at an arbitrary time when the server is not in service and the rest matrix part. However, it seems difficult to give a stochastic interpretation to the matrix part. Reference [3] shows that a similar matrix-type factorization also holds in the BMAP/G/1 queue with generalized vacations.

In this paper, focusing on the block forms of the MSP, we obtain a new sort of matrix-type factorization of the v.g.f. for the stationary queue length in MAP/MSP/1 queues, where the v.g.f. is factorized into three parts. This matrix-type factorization has a certain stochastic interpretation in some cases. For example, in the MAP/MSP/1 queue corresponding to a MAP/PH/1 queue with multiple vacations, the first part is the v.g.f. for the queue length at an arbitrary time during the vacation, the second part a matrix of phase transition rates and the third part an un-normalized m.g.f. for the queue length at an arbitrary time during the busy period in a MAP/PH/1 queue without vacations. Furthermore, in a certain type of MAP/MSP/1 queue, the first and third parts can be independently obtained from two submodels and the second part is given in terms of a matrix of the transition probabilities connecting the two submodels. We also obtain a new sort of stochastic decomposition for the stationary queue length in the M/PH/1 queue in which the service speed may vary depending on the server's state. A typical example is a working vacation model [17, 22], where the service speed falls down when the server is on vacation. In the M/PH/1 queue with working vacations, the stationary queue length is distributed as the conditional sum of the stationary queue length in an M/PH/1 queue without vacations and the number of customers in the system at an arbitrary time during the vacation. Note that, the block forms of the MSP are based on two types of phases of the MSP and similar factorizations (decompositions) have been obtained for vacation models in Refs [19, 20].

The rest of the paper is organized as follows. In Section 2, the MAP/MSP/1 queue is described in detail and some examples of the MSP are presented. In Section 3, we obtain a sort of matrix-type factorization for a general QBD process. In Section 4, matrix-type factorizations for MAP/MSP/1 queues are derived from the results in Section 3. Some special cases are discussed in Section 5. In Section 6, we explain computation of the stationary queue-length distribution and give some numerical examples.

2 Queueing Models with Markovian Service Processes

2.1 Markovian Service Process

Consider a queueing model in which the server behaves in a different manner when the system is empty. For example, in a multiple-vacation model, the server keeps on taking new vacations whenever the system is empty while it begins service for customers when the system is not empty after a vacation. In order to represent this behavior of the server, we consider two different sets of states (phases) for the server, $\mathcal{J}$ and $\mathcal{J}'$, defined as $\mathcal{J} = \{1, 2, ..., s\}$ and $\mathcal{J}' = \{1, 2, ..., s'\}$, where s is the number of states in $\mathcal{J}$ and s' that of states in $\mathcal{J}'$. When the system is not empty, the server's state is in $\mathcal{J}$; when the system is empty, the server's state is in $\mathcal{J}'$. Let $\boldsymbol{S}$ and $\boldsymbol{T}$ denote $s \times s$ matrices and let $\boldsymbol{S} + \boldsymbol{T}$ be the infinitesimal generator of the continuous-time Markov chain that governs the state transition of the server when the system is not empty. The elements of $\boldsymbol{S}$ are state transition rates without service completions, and those of $\boldsymbol{T}$ are state transition rates with service completions. Furthermore, we introduce an $s \times s$ transition probability matrix $\boldsymbol{U}$ which governs the state transition of the server at customer arrival points. This $\boldsymbol{U}$ enables us to represent the models in which the server changes its state at customer arrival points; for example, we can represent an N-policy model as a MSP. An $s' \times s'$ matrix $\boldsymbol{S}'$, an $s \times s'$ matrix $\boldsymbol{T}'$ and an $s' \times s$ matrix $\boldsymbol{U}'$ are similarly defined in the case where the system is empty. Note that $\boldsymbol{S}'$ itself is an infinitesimal generator. The MSP is represented by these six elements $(\boldsymbol{S}, \boldsymbol{T}, \boldsymbol{U}, \boldsymbol{S}', \boldsymbol{T}', \boldsymbol{U}')$.

2.2 MAP/MSP/1 Queue

Using the MSP, a MAP/MSP/1 queue is constructed as follows. Let s_A denote the number of phases of the MAP and $\mathcal{I}$ the phase set defined by $\mathcal{I} = \{1, 2, ..., s_A\}$. Let $(\boldsymbol{C}, \boldsymbol{D})$ be the representation of the MAP. Let $L(t)$ denote the number of customers in the system at time t, $J(t)$ the state (phase) of the MSP and $I(t)$ the phase of the MAP. We define the state of the system at time t by $Y(t) = (L(t), J(t), I(t))$. $\{Y(t)\}$ is the continuous-time Markov chain whose infinitesimal generator $\boldsymbol{Q}$ is given by the block tri-diagonal matrix

$$\boldsymbol{Q} = \begin{pmatrix} \boldsymbol{S}' \oplus \boldsymbol{C} & \boldsymbol{U}' \otimes \boldsymbol{D} & & & \\ \boldsymbol{T}' \otimes \boldsymbol{I} & \boldsymbol{S} \oplus \boldsymbol{C} & \boldsymbol{U} \otimes \boldsymbol{D} & & \\ & \boldsymbol{T} \otimes \boldsymbol{I} & \boldsymbol{S} \oplus \boldsymbol{C} & \boldsymbol{U} \otimes \boldsymbol{D} & \\ & & \ddots & \ddots & \ddots \end{pmatrix}, \tag{1}$$

where $\boldsymbol{I}$ is the identity matrix of suitable dimension. We denote by $\otimes$ the Kronecker product operation and by $\oplus$ the Kronecker sum operation [2]. (For some notations, we may use primes or subscripts for indicating their dimensions. For example, $\boldsymbol{I}'$ is the identity matrix of dimension s', and $\boldsymbol{I}_A$ is that of dimension s_A.) From the block form of $\boldsymbol{Q}$, $\{Y(t)\}$ is a quasi-birth-and-death (QBD) process.

For $l \geq 1$, we define $\mathcal{L}(l)$ by $\mathcal{L}(l) = \{(l, j, i);\ j \in \mathcal{J},\ i \in \mathcal{I}\}$ and refer to it as level l. We also define $\mathcal{L}(0)$ by $\mathcal{L}(0) = \{(0, j, i);\ j \in \mathcal{J}',\ i \in \mathcal{I}\}$ and refer to it as level 0. Let $\pi(l, j, i)$ be the steady state probability that the process is in the state (l, j, i). $\boldsymbol{\pi}(l) = (\pi(l, j, i),\ j \in \mathcal{J},\ i \in \mathcal{I})$ (resp. $\boldsymbol{\pi}(0) = (\pi(0, j, i),\ j \in \mathcal{J}',\ i \in \mathcal{I})$) is the row vector of the probabilities that the process is in level l (resp. level 0), and $\boldsymbol{\pi} = (\boldsymbol{\pi}(l),\ l \geq 0)$ is the stationary distribution of the process. We assume that the stationary distribution $\boldsymbol{\pi}$ always exists. Consider a period of time that begins when the process is in the state (l, j, i) for some $l \geq 1$ and ends when it enters level $l-1$ for the first time. Let $n_{(j,i)(j',i')}$ be the mean sojourn time of the process in the state (l, j', i') during the period and let $\boldsymbol{N}$ be the $s\,s_A \times s\,s_A$ matrix defined by $\boldsymbol{N} = (n_{(j,i)(j',i')})$. This $\boldsymbol{N}$ will play a crucial role in analyzing the model. Since $\{Y(t)\}$ is a QBD process, $\boldsymbol{\pi}$ is given by the matrix geometric solution [13, 16]

$$\boldsymbol{\pi}(l) = \boldsymbol{\pi}(0)(\boldsymbol{U}' \otimes \boldsymbol{D})\boldsymbol{N}\boldsymbol{R}^{l-1},\ l \geq 1, \tag{2}$$

where $\boldsymbol{R}$ is the rate matrix of $\{Y(t)\}$ and it is given by

$$\boldsymbol{R} = (\boldsymbol{U} \otimes \boldsymbol{D})\boldsymbol{N}. \tag{3}$$

$\boldsymbol{\pi}(0)$ is given by the non-negative row vector that satisfies

$$\boldsymbol{\pi}(0)\left\{\boldsymbol{S}' \oplus \boldsymbol{C} + (\boldsymbol{U}' \otimes \boldsymbol{D})\boldsymbol{N}(\boldsymbol{T}' \otimes \boldsymbol{I})\right\} = \boldsymbol{0}^\top \tag{4}$$

and the normalizing condition

$$\boldsymbol{\pi}(0)\left\{\boldsymbol{e}' + (\boldsymbol{U}' \otimes \boldsymbol{D})\boldsymbol{N}(\boldsymbol{I} - \boldsymbol{R})^{-1}\boldsymbol{e}\right\} = 1, \tag{5}$$

where $\boldsymbol{0}$ is a column vector of 0's, $\boldsymbol{e}$ is that of 1's and the superscript $\top$ indicates the transpose.

2.3 Block Forms of the MSP

In order to explore matrix-type factorizations in the next section, we divide the phase set $\mathcal{J}$ into two exclusive subsets: the set of secondary service phases, denoted by $\mathcal{J}_1$, and that of primary service phases, denoted by $\mathcal{J}_2$. Secondary service and primary service will be specified in each model. For example, in a multiple vacation model, the secondary service corresponds to vacation and the primary service corresponds to service for customers. Without loss of generality, we assume that $\mathcal{J}_1 = \{1, 2, ..., s_1\}$ and $\mathcal{J}_2 = \{s_1+1, s_1+2, ..., s_1+s_2\}$, where $s_1+s_2 = s$. According to this partition of $\mathcal{J}$, the elements $\boldsymbol{S}$, $\boldsymbol{T}$ and $\boldsymbol{U}$ of the MSP are given in the following block forms:

$$\boldsymbol{S} = \begin{pmatrix} \boldsymbol{S}_{11} & \boldsymbol{S}_{12} \\ \boldsymbol{S}_{21} & \boldsymbol{S}_{22} \end{pmatrix}, \quad \boldsymbol{T} = \begin{pmatrix} \boldsymbol{T}_{11} & \boldsymbol{T}_{12} \\ \boldsymbol{T}_{21} & \boldsymbol{T}_{22} \end{pmatrix}, \quad \boldsymbol{U} = \begin{pmatrix} \boldsymbol{U}_{11} & \boldsymbol{U}_{12} \\ \boldsymbol{U}_{21} & \boldsymbol{U}_{22} \end{pmatrix}.$$

For example, $\boldsymbol{S}_{12}$ is the $s_1 \times s_2$ matrix whose elements are state transition rates from $\mathcal{J}_1$ to $\mathcal{J}_2$. The other elements of the MSP except for $\boldsymbol{S}'$ are also given in the block forms

$$\boldsymbol{T}' = \begin{pmatrix} \boldsymbol{T}'_{11} \\ \boldsymbol{T}'_{21} \end{pmatrix}, \quad \boldsymbol{U}' = \begin{pmatrix} \boldsymbol{U}'_{11} & \boldsymbol{U}'_{12} \end{pmatrix}.$$

The infinitesimal generator $\boldsymbol{Q}$ of $\{Y(t)\}$ is given by the block tri-diagonal matrix

$$\boldsymbol{Q} = \begin{pmatrix} \boldsymbol{S}' \oplus \boldsymbol{C} & \begin{pmatrix} \boldsymbol{U}'_{11} \otimes \boldsymbol{D} & \boldsymbol{U}'_{12} \otimes \boldsymbol{D} \end{pmatrix} & & \\ \begin{pmatrix} \boldsymbol{T}'_{11} \otimes \boldsymbol{I} \\ \boldsymbol{T}'_{21} \otimes \boldsymbol{I} \end{pmatrix} & \begin{pmatrix} \boldsymbol{S}_{11} \oplus \boldsymbol{C} & \boldsymbol{S}_{12} \otimes \boldsymbol{I} \\ \boldsymbol{S}_{21} \otimes \boldsymbol{I} & \boldsymbol{S}_{22} \oplus \boldsymbol{C} \end{pmatrix} & \begin{pmatrix} \boldsymbol{U}_{11} \otimes \boldsymbol{D} & \boldsymbol{U}_{12} \otimes \boldsymbol{D} \\ \boldsymbol{U}_{21} \otimes \boldsymbol{D} & \boldsymbol{U}_{22} \otimes \boldsymbol{D} \end{pmatrix} & \\ & \begin{pmatrix} \boldsymbol{T}_{11} \otimes \boldsymbol{I} & \boldsymbol{T}_{12} \otimes \boldsymbol{I} \\ \boldsymbol{T}_{21} \otimes \boldsymbol{I} & \boldsymbol{T}_{22} \otimes \boldsymbol{I} \end{pmatrix} & \begin{pmatrix} \boldsymbol{S}_{11} \oplus \boldsymbol{C} & \boldsymbol{S}_{12} \otimes \boldsymbol{I} \\ \boldsymbol{S}_{21} \otimes \boldsymbol{I} & \boldsymbol{S}_{22} \oplus \boldsymbol{C} \end{pmatrix} & \begin{pmatrix} \boldsymbol{U}_{11} \otimes \boldsymbol{D} & \boldsymbol{U}_{12} \otimes \boldsymbol{D} \\ \boldsymbol{U}_{21} \otimes \boldsymbol{D} & \boldsymbol{U}_{22} \otimes \boldsymbol{D} \end{pmatrix} \\ & \ddots & \ddots & \ddots \end{pmatrix},$$

and $\boldsymbol{R}$ and $\boldsymbol{N}$ are given in the block forms

$$\boldsymbol{R} = \begin{pmatrix} \boldsymbol{R}_{11} & \boldsymbol{R}_{12} \\ \boldsymbol{R}_{21} & \boldsymbol{R}_{22} \end{pmatrix}, \quad \boldsymbol{N} = \begin{pmatrix} \boldsymbol{N}_{11} & \boldsymbol{N}_{12} \\ \boldsymbol{N}_{21} & \boldsymbol{N}_{22} \end{pmatrix}.$$

Furthermore, the matrix $\boldsymbol{N}$ is represented as

$$\boldsymbol{N} = \begin{pmatrix} \boldsymbol{N}_{11} & \boldsymbol{K}_{12}\boldsymbol{N}_{22} \\ \boldsymbol{K}_{21}\boldsymbol{N}_{11} & \boldsymbol{N}_{22} \end{pmatrix},$$

where $\boldsymbol{K}_{12}$ is an $s_1 s_A \times s_2 s_A$ substochastic matrix and $\boldsymbol{K}_{21}$ is an $s_2 s_A \times s_1 s_A$ substochastic matrix. We say that a matrix $\boldsymbol{A}$ is substochastic (resp. stochastic) if $\boldsymbol{A} \geq \boldsymbol{O}$ and $\boldsymbol{A}\boldsymbol{e} \leq \boldsymbol{e}$ (resp. $\boldsymbol{A}\boldsymbol{e} = \boldsymbol{e}$), where $\boldsymbol{O}$ is a matrix of 0's, and use the terms "substochastic" and "stochastic" not only for square matrices but also for non-square matrices as well as column vectors. $\boldsymbol{K}_{12}$ and $\boldsymbol{K}_{21}$ have the following meanings. For $l \geq 1$, let $\mathcal{L}_1(l) \subset \mathcal{L}(l)$ be defined by $\mathcal{L}_1(l) = \{(l, j, i);\ j \in \mathcal{J}_1,\ i \in \mathcal{I}\}$ and $\mathcal{L}_2(l)$ by $\mathcal{L}_2(l) = \mathcal{L}(l) - \mathcal{L}_1(l)$. $[\boldsymbol{K}_{12}]_{(j,i),(j',i')}$ is the probability that a sample path of $\{Y(t)\}$ starting in the state $(l, j, i) \in \mathcal{L}_1(l)$ for some $l \geq 1$ visits $\mathcal{L}_2(l)$ before entering $\mathcal{L}(l-1)$ and

(l,j',i') is the first state visited in $\mathcal{L}_2(l)$. $[\boldsymbol{K}_{21}]_{(j,i),(j',i')}$ is characterized in the same way. The stationary queue length distribution is given in the block form

$$\boldsymbol{\pi}(l) = (\,\boldsymbol{\pi}_1(l) \quad \boldsymbol{\pi}_2(l)\,)\,,\ l \geq 1,$$

where $\boldsymbol{\pi}_k(l) = (\pi(l,j,i),\ j \in \mathcal{J}_k,\ i \in \mathcal{I})$, $k = 1,2$.

In the following sections, we will obtain matrix-type factorizations related to the partition of $\mathcal{J}$ above. In preparation for it, we here classify MSPs as follows.

- *Exhaustive service type (EX-type)*: $\boldsymbol{S}_{21} = \boldsymbol{O}$, $\boldsymbol{T}_{21} = \boldsymbol{O}$ and $\boldsymbol{U}_{21} = \boldsymbol{O}$.
- *Non-preemptive service type (NP-type)*: $\boldsymbol{S}_{21} = \boldsymbol{O}$, $\boldsymbol{U}_{21} = \boldsymbol{O}$ and $\boldsymbol{U}_{22} = \boldsymbol{I}$.
- *Vacation type (VA-type)*: $\boldsymbol{T}_{11} = \boldsymbol{O}$ and $\boldsymbol{T}_{12} = \boldsymbol{O}$. (This implies $\boldsymbol{T}'_{11} = \boldsymbol{O}$.)

In an MSP of EX-type, once the server visits one of the primary service phases, it stays among the primary service phases until the system becomes empty. In an MSP of NP-type, once the server visits one of the primary service phases, it stays among the primary service phases independently of the arrival process until one customer departs from the system. In an MSP of VA-type, no departures occur when the server is in the secondary service phases.

Let $(\boldsymbol{B}_2, \boldsymbol{\beta}_2)$ be the representation of a PH-type distribution. For MSPs of EX-type and those of NP-type, we further define a special class of MSP as follows.

- *I.i.d. PH-type service type (IID-type)*:

$$\begin{aligned}
&\boldsymbol{S} = \begin{pmatrix} \boldsymbol{S}_{11} & \boldsymbol{s}_{12}\boldsymbol{\beta}_2 \\ \boldsymbol{O} & \boldsymbol{B}_2 \end{pmatrix}, \quad \boldsymbol{T} = \begin{pmatrix} \boldsymbol{T}_{11} & \boldsymbol{t}_{12}\boldsymbol{\beta}_2 \\ \boldsymbol{T}_{21} & \boldsymbol{t}_{22}\boldsymbol{\beta}_2 \end{pmatrix}, \quad \boldsymbol{U} = \begin{pmatrix} \boldsymbol{U}_{11} & \boldsymbol{u}_{12}\boldsymbol{\beta}_2 \\ \boldsymbol{O} & \boldsymbol{U}_{22} \end{pmatrix}, \\
&\boldsymbol{U}' = (\,\boldsymbol{U}'_{11} \quad \boldsymbol{u}'_{12}\boldsymbol{\beta}_2\,),
\end{aligned} \tag{6}$$

where $\boldsymbol{u}_{12} = \boldsymbol{e} - \boldsymbol{U}_{11}\boldsymbol{e}$, $\boldsymbol{u}'_{12} = \boldsymbol{e} - \boldsymbol{U}'_{11}\boldsymbol{e}$, and $\boldsymbol{s}_{12}$ and $\boldsymbol{t}_{12}$ are non-negative column vectors satisfying $\boldsymbol{S}_{11}\boldsymbol{e} + \boldsymbol{s}_{12} + \boldsymbol{T}_{11}\boldsymbol{e} + \boldsymbol{t}_{12} = \boldsymbol{0}$.

2.4 Examples of the MSP

Multiple vacation and setup time [4]: Let the service time distribution be of phase type with representation $(\boldsymbol{B}_2, \boldsymbol{\beta}_2)$ and let the vacation time distribution be of phase type with representation $(\boldsymbol{B}_1, \boldsymbol{\beta}_1)$. The representation of an MSP with multiple vacations is given by

$$\boldsymbol{S} = \begin{pmatrix} \boldsymbol{B}_1 & \boldsymbol{b}_1\boldsymbol{\beta}_2 \\ \boldsymbol{O} & \boldsymbol{B}_2 \end{pmatrix}, \quad \boldsymbol{T} = \begin{pmatrix} \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{b}_2\boldsymbol{\beta}_2 \end{pmatrix}, \quad \boldsymbol{U} = \begin{pmatrix} \boldsymbol{I} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{I} \end{pmatrix},$$

$$\boldsymbol{S}' = \boldsymbol{B}_1 + \boldsymbol{b}_1\boldsymbol{\beta}_1, \quad \boldsymbol{T}' = \begin{pmatrix} \boldsymbol{O} \\ \boldsymbol{b}_2\boldsymbol{\beta}_1 \end{pmatrix}, \quad \boldsymbol{U}' = (\,\boldsymbol{I} \quad \boldsymbol{O}\,),$$

where $\boldsymbol{b}_1 = -\boldsymbol{B}_1\boldsymbol{e}$ and $\boldsymbol{b}_2 = -\boldsymbol{B}_2\boldsymbol{e}$. Consider an MSP with setup times in which the setup-time distribution is of phase type with the representation $(\boldsymbol{B}_2, \boldsymbol{\beta}_2)$. Then, the representation of the MSP with setup times becomes the same as that of the MSP with multiple vacations except for $\boldsymbol{S}'$; $\boldsymbol{S}'$ is equal to $\boldsymbol{O}$ in the setup time model. The multiple-vacation model and the setup time model are of EX, NP, VA and IID-type.

Bernoulli vacation [11]: Let the service time distribution be of phase type with representation $(\boldsymbol{B}_2, \boldsymbol{\beta}_2)$ and let the vacation time distribution be of phase type with representation

$(\boldsymbol{B}_1, \boldsymbol{\beta}_1)$. Let p be the probability that the server begins a vacation after completing a service. The representation of an MSP with Bernoulli vacations is given by

$$\boldsymbol{S} = \begin{pmatrix} \boldsymbol{B}_1 & \boldsymbol{b}_1\boldsymbol{\beta}_2 \\ \boldsymbol{O} & \boldsymbol{B}_2 \end{pmatrix}, \quad \boldsymbol{T} = \begin{pmatrix} \boldsymbol{O} & \boldsymbol{O} \\ p\boldsymbol{b}_2\boldsymbol{\beta}_1 & (1-p)\boldsymbol{b}_2\boldsymbol{\beta}_2 \end{pmatrix}, \quad \boldsymbol{U} = \begin{pmatrix} \boldsymbol{I} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{I} \end{pmatrix},$$

$$\boldsymbol{S}' = \boldsymbol{B}_1 + \boldsymbol{b}_1\boldsymbol{\beta}_1, \quad \boldsymbol{T}' = \begin{pmatrix} \boldsymbol{O} \\ \boldsymbol{b}_2\boldsymbol{\beta}_1 \end{pmatrix}, \quad \boldsymbol{U}' = \begin{pmatrix} \boldsymbol{I} & \boldsymbol{O} \end{pmatrix},$$

where $\boldsymbol{b}_1 = -\boldsymbol{B}_1\boldsymbol{e}$ and $\boldsymbol{b}_2 = -\boldsymbol{B}_2\boldsymbol{e}$. In this model, when the system is empty, the server takes multiple vacations. The Bernoulli vacation model is of NP, VA and IID-type.

On-off service [8]: Let the service time distribution be of phase type with representation $(\boldsymbol{B}_2, \boldsymbol{\beta}_2)$ and let the on and off time distributions for the server be of phase type with representations $(\boldsymbol{B}_{on}, \boldsymbol{\beta}_{on})$ and $(\boldsymbol{B}_{off}, \boldsymbol{\beta}_{off})$, respectively. The representation of an MSP with on-off service is given by

$$\boldsymbol{S} = \begin{pmatrix} \boldsymbol{B}_{off} \otimes \boldsymbol{I} & \boldsymbol{b}_{off}\boldsymbol{\beta}_{on} \otimes \boldsymbol{I} \\ \boldsymbol{b}_{on}\boldsymbol{\beta}_{off} \otimes \boldsymbol{I} & \boldsymbol{B}_{on} \oplus \boldsymbol{B}_2 \end{pmatrix}, \quad \boldsymbol{T} = \begin{pmatrix} \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{I} \otimes \boldsymbol{b}_2\boldsymbol{\beta}_2 \end{pmatrix}, \quad \boldsymbol{U} = \begin{pmatrix} \boldsymbol{I} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{I} \end{pmatrix},$$

$$\boldsymbol{S}' = \begin{pmatrix} \boldsymbol{B}_{off} \otimes \boldsymbol{I} & \boldsymbol{b}_{off}\boldsymbol{\beta}_{on} \otimes \boldsymbol{I} \\ \boldsymbol{b}_{on}\boldsymbol{\beta}_{off} \otimes \boldsymbol{I} & \boldsymbol{B}_{on} \otimes \boldsymbol{I} \end{pmatrix}, \quad \boldsymbol{T}' = \boldsymbol{T}, \quad \boldsymbol{U}' = \boldsymbol{U},$$

where $\boldsymbol{b}_1 = -\boldsymbol{B}_1\boldsymbol{e}$, $\boldsymbol{b}_{on} = -\boldsymbol{B}_{on}\boldsymbol{e}$ and $\boldsymbol{b}_{off} = -\boldsymbol{B}_{off}\boldsymbol{e}$. The on-off service model is of VA-type.

Working vacation [17]: Let the service time distribution be of phase type with representation $(\boldsymbol{B}_2, \boldsymbol{\beta}_2)$ and let the vacation time distribution be of phase type with representation $(\boldsymbol{B}_1, \boldsymbol{\beta}_1)$. Let $c \geq 0$ be the ratio of service speed when the server is on vacation. The representation of an MSP with preemptive-repeat working vacations is given by

$$\boldsymbol{S} = \begin{pmatrix} \boldsymbol{B}_1 \oplus c\boldsymbol{B}_2 & \boldsymbol{b}_1 \otimes \boldsymbol{e}\boldsymbol{\beta}_2 \\ \boldsymbol{O} & \boldsymbol{B}_2 \end{pmatrix}, \quad \boldsymbol{T} = \begin{pmatrix} \boldsymbol{I} \otimes c\boldsymbol{b}_2\boldsymbol{\beta}_2 & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{b}_2\boldsymbol{\beta}_2 \end{pmatrix}, \quad \boldsymbol{U} = \begin{pmatrix} \boldsymbol{I} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{I} \end{pmatrix},$$

$$\boldsymbol{S}' = (\boldsymbol{B}_1 + \boldsymbol{b}_1\boldsymbol{\beta}_1) \otimes \boldsymbol{I}, \quad \boldsymbol{T}' = \begin{pmatrix} \boldsymbol{I} \otimes c\boldsymbol{b}_2\boldsymbol{\beta}_2 \\ \boldsymbol{\beta}_1 \otimes \boldsymbol{b}_2\boldsymbol{\beta}_2 \end{pmatrix}, \quad \boldsymbol{U}' = \begin{pmatrix} \boldsymbol{I} & \boldsymbol{O} \end{pmatrix},$$

where $\boldsymbol{b}_1 = -\boldsymbol{B}_1\boldsymbol{e}$ and $\boldsymbol{b}_2 = -\boldsymbol{B}_2\boldsymbol{e}$. In this model, if a vacation ends during a service for a customer, the service is repeated from the beginning. On the other hand, if the service is resumed, the model is called a preemptive-resume working vacation model and $\boldsymbol{S}_{12} = \boldsymbol{b}_1 \otimes \boldsymbol{e}\boldsymbol{\beta}_2$ is replaced with $\boldsymbol{S}_{12} = \boldsymbol{b}_1 \otimes \boldsymbol{I}$. While the preemptive-repeat working vacation model is of EX, NP and IID-type, the preemptive-resume working vacation model is of EX and NP-type.

***N*-policy** [10]: Let the service time distribution be of phase type with representation $(\boldsymbol{B}_2, \boldsymbol{\beta}_2)$. The representation of an MSP with N-policy is given by

$$\boldsymbol{S}_{11} = \boldsymbol{O},\ \boldsymbol{S}_{12} = \boldsymbol{O},\ \boldsymbol{S}_{21} = \boldsymbol{O},\ \boldsymbol{S}_{22} = \boldsymbol{B}_2, \quad \boldsymbol{T}_{11} = \boldsymbol{O},\ \boldsymbol{T}_{12} = \boldsymbol{O},\ \boldsymbol{T}_{21} = \boldsymbol{O},\ \boldsymbol{T}_{22} = \boldsymbol{b}_2\boldsymbol{\beta}_2,$$

$$\boldsymbol{U}_{11} = \begin{pmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & 0 & 1 \\ & & & 0 \end{pmatrix}, \quad \boldsymbol{U}_{12} = \begin{pmatrix} \boldsymbol{0}^\top \\ \vdots \\ \boldsymbol{0}^\top \\ \boldsymbol{\beta}_2 \end{pmatrix}, \quad \boldsymbol{U}_{21} = \boldsymbol{O}, \quad \boldsymbol{U}_{22} = \boldsymbol{I},$$

$$\boldsymbol{S}' = \boldsymbol{O}, \quad \boldsymbol{T}'_{11} = \boldsymbol{O}, \quad \boldsymbol{T}'_{21} = \begin{pmatrix} \boldsymbol{b}_2 & \boldsymbol{0} & \cdots & \boldsymbol{0} \end{pmatrix}, \quad \boldsymbol{U}'_{11} = \boldsymbol{U}_{11},\ \boldsymbol{U}'_{12} = \boldsymbol{U}_{12},$$

where $\boldsymbol{b}_2 = -\boldsymbol{B}_2\boldsymbol{e}$. $\boldsymbol{S}_{11}$, $\boldsymbol{T}_{11}$ and $\boldsymbol{U}_{11}$ are $N \times N$ matrices, i.e., $s_1 = N$. This representation makes the Markov chain $\{Y(t)\}$ reducible, but it satisfies $\boldsymbol{U}'_{11} = \boldsymbol{U}_{11}$ and $\boldsymbol{U}'_{12} = \boldsymbol{U}_{12}$. Hence we have $s' = s_1$ and this makes some formulas derived in the following sections simple. The N-policy model is of EX, NP, VA and IID-type.

Exceptional service [1]: Consider a model in which at most k customers firstly arriving in each busy period receive different services from those received by other customers. For $i \in \{1, 2, ..., k\}$, let the service time distribution of the ith customer arriving in each busy period be of phase type with representation $(\boldsymbol{B}_{1i}, \boldsymbol{\beta}_{1i})$. The service time distribution of other customers is of phase type with representation $(\boldsymbol{B}_2, \boldsymbol{\beta}_2)$. The representation of an MSP with exceptional service is given by

$$\boldsymbol{S}_{11} = \mathrm{diag}(\boldsymbol{B}_{11}, ..., \boldsymbol{B}_{1k}),\ \boldsymbol{S}_{12} = \boldsymbol{O},\ \boldsymbol{S}_{21} = \boldsymbol{O},\ \boldsymbol{S}_{22} = \boldsymbol{B}_2,$$

$$\boldsymbol{T}_{11} = \begin{pmatrix} \boldsymbol{O} & \boldsymbol{b}_{11}\boldsymbol{\beta}_{12} & & \\ & \ddots & \ddots & \\ & & \boldsymbol{O} & \boldsymbol{b}_{1k-1}\boldsymbol{\beta}_{1k} \\ & & & \boldsymbol{O} \end{pmatrix},\ \boldsymbol{T}_{12} = \begin{pmatrix} \boldsymbol{O} \\ \vdots \\ \boldsymbol{O} \\ \boldsymbol{b}_{1k}\boldsymbol{\beta}_2 \end{pmatrix},\ \boldsymbol{T}_{21} = \boldsymbol{O},\ \boldsymbol{T}_{22} = \boldsymbol{b}_2\boldsymbol{\beta}_2,$$

$$\boldsymbol{U}_{11} = \boldsymbol{I},\ \boldsymbol{U}_{12} = \boldsymbol{O},\ \boldsymbol{U}_{21} = \boldsymbol{O},\ \boldsymbol{U}_{22} = \boldsymbol{I},$$

$$\boldsymbol{S}' = \boldsymbol{O},\ \boldsymbol{T}'_{11} = \begin{pmatrix} \boldsymbol{b}_{11}\boldsymbol{\beta}_{11} & \boldsymbol{O} \\ \vdots & \vdots \\ \boldsymbol{b}_{1k}\boldsymbol{\beta}_{11} & \boldsymbol{O} \end{pmatrix},\ \boldsymbol{T}'_{21} = \begin{pmatrix} \boldsymbol{b}_2\boldsymbol{\beta}_{11} & \boldsymbol{O} \end{pmatrix},\ \boldsymbol{U}'_{11} = \boldsymbol{U}_{11},\ \boldsymbol{U}'_{12} = \boldsymbol{U}_{12},$$

where $\boldsymbol{b}_{1i} = -\boldsymbol{B}_{1i}\boldsymbol{e}$, $i = 1, 2, ..., k$, and $\boldsymbol{b}_2 = -\boldsymbol{B}_2\boldsymbol{e}$. $\mathrm{diag}(\boldsymbol{A}_1, ..., \boldsymbol{A}_k)$ denotes the block diagonal matrix whose block diagonal elements are $\boldsymbol{A}_1$, ..., $\boldsymbol{A}_{k-1}$ and $\boldsymbol{A}_k$. This representation makes the Markov chain $\{Y(t)\}$ reducible, but it satisfies $\boldsymbol{U}'_{11} = \boldsymbol{U}_{11}$ and $\boldsymbol{U}'_{12} = \boldsymbol{U}_{12}$. The exceptional service model is of EX, NP and IID-type.

3 Matrix-Type Factorizations for QBD Processes

In the following sections, we will derive matrix-type factorizations for various MAP/MSP/1 queues, which are based on the partition of the phase set $\mathcal{J}$ described in Subsection 2.3. Here we show that a sort of matrix-type factorization can be obtained for a general QBD process. The matrix-type factorizations for the MAP/MSP/1 queues are directly derived from the result.

We define some notations for a QBD process, but they are used in this section only. Hence we denote the notations with bars so that they can easily be distinguished from notations for MAP/MSP/1 queues. Consider a QBD process $\{\bar{Y}(t)\} = \{(\bar{L}(t), \bar{J}(t))\}$ on state space $\bar{\mathcal{S}} = (\{0\} \times \bar{\mathcal{J}}_B) \cup (\mathcal{N}_+ \times \bar{\mathcal{J}}_A)$, where $\bar{L}(t)$ and $\bar{J}(t)$ are the level and the phase at time t, $\bar{\mathcal{J}}_A = \{1, 2, ..., \bar{s}_A\}$ and $\bar{\mathcal{J}}_B = \{1, 2, ..., \bar{s}_B\}$ are the phase sets, and $\mathcal{N}_+$ is the set of positive integers. Let the infinitesimal generator of the QBD process be

$$\bar{\boldsymbol{Q}} = \begin{pmatrix} \bar{\boldsymbol{B}}(1) & \bar{\boldsymbol{B}}(0) & & & \\ \bar{\boldsymbol{B}}(2) & \bar{\boldsymbol{A}}(1) & \bar{\boldsymbol{A}}(0) & & \\ & \bar{\boldsymbol{A}}(2) & \bar{\boldsymbol{A}}(1) & \bar{\boldsymbol{A}}(0) & \\ & & \ddots & \ddots & \ddots \end{pmatrix}, \tag{7}$$

where $\bar{\boldsymbol{A}}(i)$, $i = 1, 2, 3$, are $\bar{s}_A \times \bar{s}_A$ matrices, $\bar{\boldsymbol{B}}(0)$ an $\bar{s}_B \times \bar{s}_A$ matrix, $\bar{\boldsymbol{B}}(1)$ an $\bar{s}_B \times \bar{s}_B$ matrix and $\bar{\boldsymbol{B}}(2)$ an $\bar{s}_A \times \bar{s}_B$ matrix. We assume that the QBD process has the stationary distribution and denote it by $\bar{\boldsymbol{\pi}} = \begin{pmatrix} \bar{\boldsymbol{\pi}}(0) & \bar{\boldsymbol{\pi}}(1) & \bar{\boldsymbol{\pi}}(2) & ... \end{pmatrix}$. Consider a period of time that begins when the process is in the state (l, j) for some $l \geq 1$ and ends when it enters level $l - 1$ for the first time. Let $\bar{n}_{jj'}$ be the mean sojourn time of the process in the state (l, j') during the period and let $\bar{\boldsymbol{N}}$ be the $\bar{s}_A \times \bar{s}_A$ matrix defined by $\bar{\boldsymbol{N}} = (\bar{n}_{jj'})$. Let $\bar{\boldsymbol{R}}$ denote the rate matrix of the QBD process, then the matrix geometric solution of the stationary distribution is given by

$$\bar{\boldsymbol{\pi}}(i) = \bar{\boldsymbol{\pi}}(1)\bar{\boldsymbol{R}}^{i-1},\ i \geq 2, \tag{8}$$

where $\bar{\boldsymbol{\pi}}(1) = \bar{\boldsymbol{\pi}}(0)\bar{\boldsymbol{B}}(0)\bar{\boldsymbol{N}}$, and $\bar{\boldsymbol{\pi}}(0)$ is given by the non-negative vector that satisfies

$$\bar{\boldsymbol{\pi}}(0)\left\{\bar{\boldsymbol{B}}(1) + \bar{\boldsymbol{B}}(0)\bar{\boldsymbol{N}}\bar{\boldsymbol{B}}(2)\right\} = \boldsymbol{0}^\top \quad \text{and} \quad \bar{\boldsymbol{\pi}}(0)\left\{\boldsymbol{I} + \bar{\boldsymbol{B}}(0)\bar{\boldsymbol{N}}(\boldsymbol{I} - \bar{\boldsymbol{R}})^{-1}\right\}\boldsymbol{e} = 1.$$

Now we arbitrarily divide $\bar{\mathcal{J}}_A$ into two subsets $\bar{\mathcal{J}}_{A1}$ and $\bar{\mathcal{J}}_{A2}$. Without loss of generality, we assume that $\bar{\mathcal{J}}_{A1} = \{1, 2, ..., \bar{s}_{A1}\}$ and $\bar{\mathcal{J}}_{A2} = \{\bar{s}_{A1}+1, \bar{s}_{A1}+2, ..., \bar{s}_{A1}+\bar{s}_{A2}\}$, where $\bar{s}_{A1}+\bar{s}_{A2} = \bar{s}_A$. Then, the following block forms of $\bar{\boldsymbol{A}}(k)$, $\bar{\boldsymbol{B}}(0)$, $\bar{\boldsymbol{B}}(2)$, $\bar{\boldsymbol{R}}$ and $\bar{\boldsymbol{\pi}}(l)$ are obtained.

$$\bar{\boldsymbol{A}}(k) = \begin{pmatrix} \bar{\boldsymbol{A}}_{11}(k) & \bar{\boldsymbol{A}}_{12}(k) \\ \bar{\boldsymbol{A}}_{21}(k) & \bar{\boldsymbol{A}}_{22}(k) \end{pmatrix},\ k = 0, 1, 2, \quad \bar{\boldsymbol{B}}(0) = \begin{pmatrix} \bar{\boldsymbol{B}}_{11}(0) & \bar{\boldsymbol{B}}_{12}(0) \end{pmatrix},$$

$$\bar{\boldsymbol{B}}(2) = \begin{pmatrix} \bar{\boldsymbol{B}}_{11}(2) \\ \bar{\boldsymbol{B}}_{21}(2) \end{pmatrix}, \quad \bar{\boldsymbol{R}} = \begin{pmatrix} \bar{\boldsymbol{R}}_{11} & \bar{\boldsymbol{R}}_{12} \\ \bar{\boldsymbol{R}}_{21} & \bar{\boldsymbol{R}}_{22} \end{pmatrix}, \quad \bar{\boldsymbol{\pi}}(l) = \begin{pmatrix} \bar{\boldsymbol{\pi}}_1(l) & \bar{\boldsymbol{\pi}}_2(l) \end{pmatrix},\ l \geq 0.$$

Matrix-type factorizations for the QBD process are given by the next lemma.

Lemma 1 *For $l \geq 1$, $\bar{\boldsymbol{\pi}}_1(l)$ is represented in terms of $\bar{\boldsymbol{\pi}}_2(k)$ as follows.*

$$\bar{\boldsymbol{\pi}}_1(l) = \bar{\boldsymbol{\pi}}_1(1)\bar{\boldsymbol{R}}_{11}^{l-1} + \sum_{k=1}^{l-1} \bar{\boldsymbol{\pi}}_2(k)\bar{\boldsymbol{R}}_{21}\bar{\boldsymbol{R}}_{11}^{l-1-k}. \tag{9}$$

For $l \geq 1$, $\bar{\boldsymbol{\pi}}_2(l)$ is also represented in terms of $\bar{\boldsymbol{\pi}}_1(k)$ as follows.

$$\bar{\boldsymbol{\pi}}_2(l) = \bar{\boldsymbol{\pi}}_2(1)\bar{\boldsymbol{R}}_{22}^{l-1} + \sum_{k=1}^{l-1} \bar{\boldsymbol{\pi}}_1(k)\bar{\boldsymbol{R}}_{12}\bar{\boldsymbol{R}}_{22}^{l-1-k}. \tag{10}$$

From these formulae, the transform of $\bar{\boldsymbol{\pi}}(l)$, $\tilde{\boldsymbol{\pi}}^*(z)$, defined as

$$\tilde{\boldsymbol{\pi}}^*(z) \equiv \sum_{l=1}^{\infty} z^l \begin{pmatrix} \bar{\boldsymbol{\pi}}_1(l) & \bar{\boldsymbol{\pi}}_2(l) \end{pmatrix} = \begin{pmatrix} \tilde{\boldsymbol{\pi}}_1^*(z) & \tilde{\boldsymbol{\pi}}_2^*(z) \end{pmatrix}$$

is given by

$$\tilde{\boldsymbol{\pi}}_1^*(z) = \left(z\bar{\boldsymbol{\pi}}_1(1) + \tilde{\boldsymbol{\pi}}_2^*(z)z\bar{\boldsymbol{R}}_{21}\right)\left(\boldsymbol{I} - z\bar{\boldsymbol{R}}_{11}\right)^{-1}, \tag{11}$$

$$\tilde{\boldsymbol{\pi}}_2^*(z) = \left(z\bar{\boldsymbol{\pi}}_2(1) + \tilde{\boldsymbol{\pi}}_1^*(z)z\bar{\boldsymbol{R}}_{12}\right)\left(\boldsymbol{I} - z\bar{\boldsymbol{R}}_{22}\right)^{-1}. \tag{12}$$

Proof: From the matrix geometric solution, we have $\bar{\boldsymbol{\pi}}(l+1) = \bar{\boldsymbol{\pi}}(l)\bar{\boldsymbol{R}}$ for $l \geq 1$. Hence we obtain

$$\bar{\boldsymbol{\pi}}_1(l+1) = \bar{\boldsymbol{\pi}}_1(l)\bar{\boldsymbol{R}}_{11} + \bar{\boldsymbol{\pi}}_2(l)\bar{\boldsymbol{R}}_{21},\ l \geq 1,$$

$$\bar{\boldsymbol{\pi}}_2(l+1) = \bar{\boldsymbol{\pi}}_2(l)\bar{\boldsymbol{R}}_{22} + \bar{\boldsymbol{\pi}}_1(l)\bar{\boldsymbol{R}}_{12},\ l \geq 1.$$

Equations (9) and (10) are derived from these recursive formulae. □

The reason why we use the transform $\tilde{\boldsymbol{\pi}}^*(z)$ in this lemma will be explained in Remark 1 in the next section. Here we do not give any interpretations to matrix-type factorizations (11) and (12) since the partition of the phase set of the QBD process is arbitrary and it has no meanings. However, we have to say that the assertion of this lemma is essential and matrix-type factorizations for MAP/MSP/1 queues can be directly obtained from these factorizations.

4 Matrix-Type Factorizations for MAP/MSP/1 Queues - General Case -

Here we consider the MAP/MSP/1 queue described in Section 2. For $k \in \{1, 2\}$, let $\boldsymbol{\pi}_k(l)$ be the row vector of the probabilities that the queue length is l and the server's state is in $\mathcal{J}_k$ at an arbitrary time. In the case of an M/PH/1 queue with multiple vacations, $\boldsymbol{\pi}_1(l)$ is the row vector of the probabilities that the queue length is l and the server is on vacation at an arbitrary time; $\boldsymbol{\pi}_2(l)$ is the row vector of the probabilities that the queue length is l and the server is serving a customer at an arbitrary time. From the stochastic decomposition property for the M/G/1 queue with multiple vacations, we obtain

$$\boldsymbol{\pi}_2(l)\boldsymbol{e} = \boldsymbol{\pi}(l)\boldsymbol{e} - \boldsymbol{\pi}_1(l)\boldsymbol{e} = \sum_{k=1}^{l} \boldsymbol{\pi}_1(l-k)\boldsymbol{e}\, \frac{\pi_0(k)}{1-\rho_0},\ l \geq 1, \tag{13}$$

where $\boldsymbol{\pi}_0 = (\pi_0(l))$ is the stationary queue length distribution in an M/PH/1 queue without vacations and ρ_0 is the traffic intensity of the queue. In this formula, we assume that $\boldsymbol{\pi}_1(0) = \boldsymbol{\pi}(0)$. From equation (13), we obtain the next transform of $\boldsymbol{\pi}_2(l)\boldsymbol{e}$:

$$\tilde{\boldsymbol{\pi}}_2^*(z)\boldsymbol{e} \equiv \sum_{l=1}^{\infty} z^l \boldsymbol{\pi}_2(l)\boldsymbol{e} = (\boldsymbol{\pi}(0)\boldsymbol{e} + \tilde{\boldsymbol{\pi}}_1^*(z)\boldsymbol{e})\ \frac{\tilde{\pi}_0^*(z)}{1-\rho_0}, \tag{14}$$

where $\tilde{\boldsymbol{\pi}}_1^*(z)\boldsymbol{e} \equiv \sum_{l=1}^{\infty} z^l \boldsymbol{\pi}_1(l)\boldsymbol{e}$ and $\tilde{\pi}_0^*(z) \equiv \sum_{l=1}^{\infty} z^l \pi_0(l)$. For the MAP/MSP/1 queue, taking account of $\boldsymbol{\pi}(1) = \boldsymbol{\pi}(0)(\boldsymbol{U}' \otimes \boldsymbol{D})\boldsymbol{N}$, we directly derive the next lemma from Lemma 1, and this lemma gives an extension of formula (14).

Lemma 2 *In the MAP/MSP/1 queue, for $l \geq 1$, $\boldsymbol{\pi}_1(l)$ is represented in terms of $\boldsymbol{\pi}_2(k)$ as follows:*

$$\boldsymbol{\pi}_1(l) = \boldsymbol{\pi}(0)(\boldsymbol{I} \otimes \boldsymbol{D})\hat{\boldsymbol{K}}_{21}'\boldsymbol{N}_{11}\boldsymbol{R}_{11}^{l-1} + \sum_{k=1}^{l-1} \boldsymbol{\pi}_2(k)(\boldsymbol{I} \otimes \boldsymbol{D})\hat{\boldsymbol{K}}_{21}\boldsymbol{N}_{11}\boldsymbol{R}_{11}^{l-1-k}, \tag{15}$$

where $\hat{\boldsymbol{K}}_{21}' = (\boldsymbol{U}_{11}' \otimes \boldsymbol{I}) + (\boldsymbol{U}_{12}' \otimes \boldsymbol{I})\boldsymbol{K}_{21}$ and $\hat{\boldsymbol{K}}_{21} = (\boldsymbol{U}_{21} \otimes \boldsymbol{I}) + (\boldsymbol{U}_{22} \otimes \boldsymbol{I})\boldsymbol{K}_{21}$. Both $\hat{\boldsymbol{K}}_{21}'$ and $\hat{\boldsymbol{K}}_{21}$ are substochastic. For $l \geq 1$, $\boldsymbol{\pi}_2(l)$ is also represented in terms of $\boldsymbol{\pi}_1(k)$ as follows:

$$\boldsymbol{\pi}_2(l) = \boldsymbol{\pi}(0)(\boldsymbol{I} \otimes \boldsymbol{D})\hat{\boldsymbol{K}}_{12}'\boldsymbol{N}_{22}\boldsymbol{R}_{22}^{l-1} + \sum_{k=1}^{l-1} \boldsymbol{\pi}_1(k)(\boldsymbol{I} \otimes \boldsymbol{D})\hat{\boldsymbol{K}}_{12}\boldsymbol{N}_{22}\boldsymbol{R}_{22}^{l-1-k}, \tag{16}$$

where $\hat{\boldsymbol{K}}_{12}' = (\boldsymbol{U}_{11}' \otimes \boldsymbol{I})\boldsymbol{K}_{12} + (\boldsymbol{U}_{12}' \otimes \boldsymbol{I})$ and $\hat{\boldsymbol{K}}_{12} = (\boldsymbol{U}_{11} \otimes \boldsymbol{I})\boldsymbol{K}_{12} + (\boldsymbol{U}_{12} \otimes \boldsymbol{I})$. Both $\hat{\boldsymbol{K}}_{12}'$ and $\hat{\boldsymbol{K}}_{12}$ are substochastic. From these formulae, $\tilde{\boldsymbol{\pi}}^*(z) \equiv \sum_{l=1}^{\infty} z^l \left(\boldsymbol{\pi}_1(l) \quad \boldsymbol{\pi}_2(l) \right) = \left(\tilde{\boldsymbol{\pi}}_1^*(z) \quad \tilde{\boldsymbol{\pi}}_2^*(z) \right)$ is given by

$$\tilde{\boldsymbol{\pi}}_1^*(z) = \left\{ \boldsymbol{\pi}(0)(\boldsymbol{I} \otimes \boldsymbol{D})\hat{\boldsymbol{K}}_{21}' + \tilde{\boldsymbol{\pi}}_2^*(z)(\boldsymbol{I} \otimes \boldsymbol{D})\hat{\boldsymbol{K}}_{21} \right\} z\boldsymbol{N}_{11}\left(\boldsymbol{I} - z\boldsymbol{R}_{11}\right)^{-1}, \tag{17}$$

$$\tilde{\boldsymbol{\pi}}_2^*(z) = \left\{ \boldsymbol{\pi}(0)(\boldsymbol{I} \otimes \boldsymbol{D})\hat{\boldsymbol{K}}_{12}' + \tilde{\boldsymbol{\pi}}_1^*(z)(\boldsymbol{I} \otimes \boldsymbol{D})\hat{\boldsymbol{K}}_{12} \right\} z\boldsymbol{N}_{22}\left(\boldsymbol{I} - z\boldsymbol{R}_{22}\right)^{-1}. \tag{18}$$

Remark 1 *In our model, $\mathcal{J}'$, which is the phase set of the MSP when the system is empty, can be given independently of $\mathcal{J}_1$ and $\mathcal{J}_2$, which are the phase sets of the MSP when the system is not empty. Hence we can not always define the ordinary v.g.f. for the stationary queue length. Here the ordinary v.g.f. $\boldsymbol{\pi}^*(z)$ is defined as $\boldsymbol{\pi}^*(z) \equiv \sum_{l=0}^{\infty} z^l \left(\boldsymbol{\pi}_1(l) \quad \boldsymbol{\pi}_2(l) \right) = \left(\boldsymbol{\pi}_1^*(z) \quad \boldsymbol{\pi}_2^*(z) \right)$. This is the reason why we use the transforms such as $\tilde{\boldsymbol{\pi}}_1^*(z)$ in Lemma 2. Typical cases where the ordinary v.g.f. can be defined are Case J1: $\mathcal{J}' = \mathcal{J}_1 + \mathcal{J}_2 = \mathcal{J}$, Case J2: $\mathcal{J}' = \mathcal{J}_1$ and Case J3: $\mathcal{J}' = \mathcal{J}_2$. In each case, $\boldsymbol{\pi}_1(0)$ and $\boldsymbol{\pi}_2(0)$ are defined as follows.*

- **Case J1**: *We assume that* $\boldsymbol{\pi}(0)$ *has the block form* $\boldsymbol{\pi}(0) = (\,\boldsymbol{\pi}_1(0) \quad \boldsymbol{\pi}_2(0)\,)$.
- **Case J2**: *We assume that* $\boldsymbol{\pi}_1(0) = \boldsymbol{\pi}(0)$ *and* $\boldsymbol{\pi}_2(0) = \mathbf{0}^\top$. *This means that when the system is empty, the server's state is always in the secondary service phases.*
- **Case J3**: *We assume that* $\boldsymbol{\pi}_1(0) = \mathbf{0}^\top$ *and* $\boldsymbol{\pi}_2(0) = \boldsymbol{\pi}(0)$. *This means that when the system is empty, the server's state is always in the primary service phases.*

Almost all the examples in Section 2 are included in Case J2. Only one exception is the on-off model, which is included in Case J1. Hence we will mainly investigate Case J2 in the next section. □

In Lemma 2, equation (18) represents a sort of factorization of the v.g.f. for the steady state queue length in the MAP/MSP/1 queue. In order to see this point more clearly, we define a new type of MSP in the case of $\mathcal{J}' = \mathcal{J}_1$ (Case J2 in Remark 1).

- *Homogenous type (HM-type)*: $\boldsymbol{U}'_{11} = \boldsymbol{U}_{11}$ and $\boldsymbol{U}'_{12} = \boldsymbol{U}_{12}$.

In an MSP of HM-type, the dimension of $\boldsymbol{S}'$ is equal to that of $\boldsymbol{S}_{11}$, and we obtain $\hat{\boldsymbol{K}}'_{12} = \hat{\boldsymbol{K}}_{12}$. In the MAP/MSP/1 queue of HM-type, we can, therefore, define the ordinary v.g.f. for the stationary queue length, $\boldsymbol{\pi}^*(z) = (\,\boldsymbol{\pi}^*_1(z) \quad \boldsymbol{\pi}^*_2(z)\,)$, where $\boldsymbol{\pi}^*_2(z)$ is given by

$$\boldsymbol{\pi}^*_2(z) = \boldsymbol{\pi}^*_1(z)(\boldsymbol{I} \otimes \boldsymbol{D})\hat{\boldsymbol{K}}_{12} z \boldsymbol{N}_{22} \left(\boldsymbol{I} - z\boldsymbol{R}_{22}\right)^{-1}. \tag{19}$$

This equation shows that $\boldsymbol{\pi}^*_2(z)$ can be factorized into three parts: $\boldsymbol{\pi}^*_1(z)$, which is the v.g.f. for the queue length at an arbitrary time during the secondary service, $(\boldsymbol{I} \otimes \boldsymbol{D})\hat{\boldsymbol{K}}_{12}$, which is a matrix of phase transition rates, and the rest matrix part $z\boldsymbol{N}_{22} \left(\boldsymbol{I} - z\boldsymbol{R}_{22}\right)^{-1}$. In the next section, we consider two special cases where the rest matrix part has an obvious stochastic interpretation. An interpretation for the second part is given by the next remark.

Remark 2 *As mentioned in Subsection 2.3,* $[\boldsymbol{K}_{12}]_{(j,i),(j',i')}$ *is the probability that a sample path of* $\{Y(t)\}$ *starting in the state* $(l, j, i) \in \mathcal{L}_1(l)$ *for some* $l \geq 1$ *visits* $\mathcal{L}_2(l)$ *before entering* $\mathcal{L}(l-1)$ *and* (l, j', i') *is the first state visited in* $\mathcal{L}_2(l)$. $\boldsymbol{U}_{11} \otimes \boldsymbol{D}$ *is the matrix of the phase transition rates that an arrival occurs and the phases of the MAP and the MSP change in* $\mathcal{J}_1 \times \mathcal{I}$, *and* $\boldsymbol{U}_{12} \otimes \boldsymbol{D}$ *is the matrix of the phase transition rates that an arrival occurs and the phases change from* $\mathcal{J}_1 \times \mathcal{I}$ *to* $\mathcal{J}_2 \times \mathcal{I}$. *Hence,* $[(\boldsymbol{I} \otimes \boldsymbol{D})\hat{\boldsymbol{K}}_{12}]_{(j,i),(j',i')}$ *is the phase transition rate that the process* $\{Y(t)\}$ *is in the state* $(l-1, j, i) \in \mathcal{L}_1(l-1)$ *for some* $l \geq 2$ *just before an arrival point of time and, after the point of time, it visits* $\mathcal{L}_2(l)$ *before entering* $\mathcal{L}(l-1)$ *and* (l, j', i') *is the first state visited in* $\mathcal{L}_2(l)$. *Here a " rate" means the expected number of events occurring in a unit time.*

5 Matrix-type Factorizations for MAP/MSP/1 Queues - Two Special Cases -

5.1 Preliminaries

Consider a MAP/MSP/1 queue, in which the representation of the MSP is $(\boldsymbol{S}, \boldsymbol{T}, \boldsymbol{U}, \boldsymbol{S}', \boldsymbol{T}', \boldsymbol{U}')$ and that of the MAP is $(\boldsymbol{C}, \boldsymbol{D})$. We use the same partition of the phase space $\mathcal{J}$ as used in Section 2. In this section, we study two special cases where $\boldsymbol{R}_{22}$ becomes the rate matrix of another MAP/MSP/1 queue, called a base model: one is the case where the MSP is of EX-type and the other is the case where it is of NP-type. In these cases, an obvious interpretation is obtained for the matrix part $z\boldsymbol{N}_{22}(\boldsymbol{I} - z\boldsymbol{R}_{22})^{-1}$ in the equation (18).

Let the representation of the MSP of the base model be $(\boldsymbol{S}_0, \boldsymbol{T}_0, \boldsymbol{U}_0, \boldsymbol{O}, \boldsymbol{T}_0, \boldsymbol{U}_0)$. $\boldsymbol{S}_0$, $\boldsymbol{T}_0$ and $\boldsymbol{U}_0$ will be specified in each case. The representation of the MAP in the base model is the same as that of the original model. Let $\boldsymbol{\pi}_0 = (\boldsymbol{\pi}_0(l))$ denote the stationary queue length distribution in the base model and $\boldsymbol{R}_0$ the rate matrix. $\boldsymbol{R}_0$ is given by $\boldsymbol{R}_0 = (\boldsymbol{U}_0 \otimes \boldsymbol{D})\boldsymbol{N}_0$, where $\boldsymbol{N}_0$ is the non-negative matrix corresponding to $\boldsymbol{N}$ of the original model. $\boldsymbol{\pi}_0$ is given by the matrix geometric solution

$$\boldsymbol{\pi}_0(l) = \boldsymbol{\pi}_0(0)(\boldsymbol{U}_0 \otimes \boldsymbol{D})\boldsymbol{N}_0\boldsymbol{R}_0^{l-1},\ l \geq 1, \tag{20}$$

and its transform is given by

$$\tilde{\boldsymbol{\pi}}_0^*(z) \equiv \sum_{l=1}^{\infty} z^l \boldsymbol{\pi}_0(l) = \boldsymbol{\pi}_0(0)(\boldsymbol{I} \otimes \boldsymbol{D})(\boldsymbol{U}_0 \otimes \boldsymbol{I}_A)z\boldsymbol{N}_0(\boldsymbol{I} - z\boldsymbol{R}_0)^{-1}. \tag{21}$$

For the base model, the ordinary v.g.f. of the stationary queue length can be defined, and it is given by $\boldsymbol{\pi}_0^*(z) \equiv \sum_{l=0}^{\infty} z^l \boldsymbol{\pi}_0(l) = \boldsymbol{\pi}_0(0) + \tilde{\boldsymbol{\pi}}_0^*(z)$.

5.2 The Case of Exhaustive Service Type (EX-Type)

5.2.1 General model

Here we consider the MSP of EX-type, i.e., $\boldsymbol{S}_{21} = \boldsymbol{O}$, $\boldsymbol{T}_{21} = \boldsymbol{O}$ and $\boldsymbol{U}_{21} = \boldsymbol{O}$. In this case, how to specify the base model is trivial; it is the MAP/MSP/1 queue in which $\boldsymbol{S}_0$, $\boldsymbol{T}_0$ and $\boldsymbol{U}_0$ are given by $\boldsymbol{S}_0 = \boldsymbol{S}_{22}$, $\boldsymbol{T}_0 = \boldsymbol{T}_{22}$ and $\boldsymbol{U}_0 = \boldsymbol{U}_{22}$. If the original model is stable, the base model is also stable. From the definition of EX-type, it is obvious that $\boldsymbol{K}_{21} = \boldsymbol{O}$ and this leads us to $\hat{\boldsymbol{K}}_{21}' = \boldsymbol{U}_{11}' \otimes \boldsymbol{I}$ and $\hat{\boldsymbol{K}}_{21} = \boldsymbol{O}$. Furthermore, once the server's state enters one of the primary service phases, the state transition of the server is governed by $\boldsymbol{S}_{22}$, $\boldsymbol{T}_{22}$ and $\boldsymbol{U}_{22}$ until the system becomes empty. Hence we obtain $\boldsymbol{N}_{22} = \boldsymbol{N}_0$ and this leads us to $\boldsymbol{R}_{22} = \boldsymbol{R}_0$. As a result, the non-negative matrix $\boldsymbol{N}$ and the rate matrix $\boldsymbol{R}$ are given by

$$\boldsymbol{N} = \begin{pmatrix} \boldsymbol{N}_{11} & \boldsymbol{K}_{12}\boldsymbol{N}_0 \\ \boldsymbol{O} & \boldsymbol{N}_0 \end{pmatrix}, \quad \boldsymbol{R} = \begin{pmatrix} \boldsymbol{R}_{11} & (\boldsymbol{I} \otimes \boldsymbol{D})\hat{\boldsymbol{K}}_{12}\boldsymbol{N}_0 \\ \boldsymbol{O} & \boldsymbol{R}_0 \end{pmatrix}, \tag{22}$$

where $\boldsymbol{R}_{11} = (\boldsymbol{U}_{11} \otimes \boldsymbol{D})\boldsymbol{N}_{11}$, $\boldsymbol{R}_0 = (\boldsymbol{U}_0 \otimes \boldsymbol{D})\boldsymbol{N}_0$ and $\hat{\boldsymbol{K}}_{12} = (\boldsymbol{U}_{11} \otimes \boldsymbol{I})\boldsymbol{K}_{12} + (\boldsymbol{U}_{12} \otimes \boldsymbol{I})$. Let $\boldsymbol{\pi} = (\, \boldsymbol{\pi}(0) \quad \boldsymbol{\pi}(1) \quad \dots\,)$ be the stationary distribution of the original model, where $\boldsymbol{\pi}(l) = (\, \boldsymbol{\pi}_1(l) \quad \boldsymbol{\pi}_2(l)\,)$, $l \geq 1$. From Lemma 2, we obtain the next lemma.

Lemma 3 *In the MAP/MSP/1 queue of EX-type, $\tilde{\boldsymbol{\pi}}_1^*(z)$ is given by*

$$\tilde{\boldsymbol{\pi}}_1^*(z) = \boldsymbol{\pi}(0)(\boldsymbol{U}_{11}' \otimes \boldsymbol{D})z\boldsymbol{N}_{11}(\boldsymbol{I} - z\boldsymbol{R}_{11})^{-1}, \tag{23}$$

and $\tilde{\boldsymbol{\pi}}_2^*(z)$ is given by

$$\tilde{\boldsymbol{\pi}}_2^*(z) = \left\{\boldsymbol{\pi}(0)(\boldsymbol{I} \otimes \boldsymbol{D})\hat{\boldsymbol{K}}_{12}' + \tilde{\boldsymbol{\pi}}_1^*(z)(\boldsymbol{I} \otimes \boldsymbol{D})\hat{\boldsymbol{K}}_{12}\right\} z\boldsymbol{N}_0(\boldsymbol{I} - z\boldsymbol{R}_0)^{-1}, \tag{24}$$

where $\hat{\boldsymbol{K}}_{12}' = (\boldsymbol{U}_{11}' \otimes \boldsymbol{I})\boldsymbol{K}_{12} + (\boldsymbol{U}_{12}' \otimes \boldsymbol{I})$. □

Remark 3 *The matrix part $z\boldsymbol{N}_0(\boldsymbol{I} - z\boldsymbol{R}_0)^{-1}$ in equation (24) is represented as*

$$z\boldsymbol{N}_0(\boldsymbol{I} - z\boldsymbol{R}_0)^{-1} = \sum_{l=1}^{\infty} z^l \boldsymbol{N}_0 \boldsymbol{R}_0^{l-1}.$$

Let $\{Y_0(t)\} = \{(L_0(t), J_0(t), I_0(t))\}$ be the QBD process representing the behavior of the base model, where $L_0(t)$ is the number of customers in the system at time t, $J_0(t)$ the phase of the

MSP and $I_0(t)$ the phase of the MAP. $[\boldsymbol{N}_0\boldsymbol{R}_0^{l-1}]_{(j,i)(j',i')}$ *is the mean sojourn time of the process* $\{Y_0(t)\}$ *in the state* (l,j',i') *during a busy period starting in the state* $(1,j,i)$*. This leads us to a stochastic interpretation for the matrix part in the equation (24). Let* $\boldsymbol{A}$ *be defined by* $\boldsymbol{A} = diag(\boldsymbol{N}_0(\boldsymbol{I}-\boldsymbol{R})^{-1}\boldsymbol{e})$*, where* $diag(\boldsymbol{a})$ *is the diagonal matrix whose diagonal elements are those of vector* $\boldsymbol{a}$*. Then,* $[\boldsymbol{A}^{-1}\boldsymbol{N}_0\boldsymbol{R}_0^{l-1}]_{(j,i)(j',i')}$ *is the conditional probability that the queue length in the base model is* l *and the phase processes are in the state* (j',i') *at an arbitrary time during a busy period given that the busy period has started in the state* $(1,j,i)$*. Hence the matrix* $\boldsymbol{A}^{-1}z\boldsymbol{N}_0(\boldsymbol{I}-z\boldsymbol{R}_0)^{-1}$ *is the m.g.f. for the queue length at an arbitrary time during the busy period.* □

Remark 4 *If the MSP is of VA-type, the number of customers in the system never decreases whenever the server's state is in the secondary service phases. Hence,* $\hat{\boldsymbol{K}}_{12}$ *and* $\hat{\boldsymbol{K}}'_{12}$ *become stochastic.* □

If the MSP is of HM-type, i.e., $\boldsymbol{U}'_{11} = \boldsymbol{U}_{11}$ and $\boldsymbol{U}'_{12} = \boldsymbol{U}_{12}$, then it is included in Case J2 in Remark 1 and we have $\hat{\boldsymbol{K}}'_{12} = \hat{\boldsymbol{K}}_{12}$. Hence we obtain

$$\boldsymbol{\pi}_1^*(z) = \boldsymbol{\pi}(0)(\boldsymbol{I}-z\boldsymbol{R}_{11})^{-1}, \tag{25}$$

$$\boldsymbol{\pi}_2^*(z) = \tilde{\boldsymbol{\pi}}_2^*(z) = \boldsymbol{\pi}_1^*(z)(\boldsymbol{I}\otimes\boldsymbol{D})\hat{\boldsymbol{K}}_{12}z\boldsymbol{N}_0(\boldsymbol{I}-z\boldsymbol{R}_0)^{-1}. \tag{26}$$

Let ρ_2 be defined by $\rho_2 = \boldsymbol{\pi}_2^*(1)\boldsymbol{e}$, $\boldsymbol{Y}_1(z)$ by $\boldsymbol{Y}_1(z) = \frac{\boldsymbol{\pi}_1^*(z)}{1-\rho_2}$, and $\boldsymbol{Y}_2(z)$ by $\boldsymbol{Y}_2(z) = \frac{\boldsymbol{\pi}_2^*(z)}{\rho_2}$. $\boldsymbol{Y}_1(z)$ is the v.g.f. for the stationary queue length when the server's state is in the secondary service phases, and $\boldsymbol{Y}_2(z)$ is that for the stationary queue length when the server's state is in the primary service phases. From equation (26), we also obtain the following matrix-type factorization.

$$\boldsymbol{Y}_2(z) = \boldsymbol{Y}_1(z)(\boldsymbol{I}\otimes\boldsymbol{D})\hat{\boldsymbol{K}}_{12}\frac{1-\rho_2}{\rho_2}z\boldsymbol{N}_0(\boldsymbol{I}-z\boldsymbol{R}_0)^{-1} \tag{27}$$

Remark 5 *Comparing equation (26) with equation (21), we can obtain an interpretation for equation (26) as follows: equation (26) has the same form as equation (21), where* $\boldsymbol{\pi}_0(0)$ *in equation (21) is replaced with* $\boldsymbol{\pi}_1^*(z)$ *in equation (26) and* $(\boldsymbol{U}_0\otimes\boldsymbol{I}_A)$ *with* $\hat{\boldsymbol{K}}_{12}$*.* □

5.2.2 I.i.d. PH-type (IID-type) service model

We further assume that the MSP of the original model is also of IID-type, where the representation of the service time distribution is given by $(\boldsymbol{B}_2, \boldsymbol{\beta}_2)$ (see the block forms of (6)). Let $\tilde{\boldsymbol{K}}_{12}$ be defined by $\tilde{\boldsymbol{K}}_{12} = \boldsymbol{K}_{12}(\boldsymbol{e}_2\otimes\boldsymbol{I}_A)$. $[\tilde{\boldsymbol{K}}_{12}]_{(j,i),i'}$ is the probability that a sample path of $\{Y(t)\}$ starting in the state $(l,j,i)\in\mathcal{L}_1(l)$ for some $l\geq 1$ visits $\mathcal{L}_2(l)$ before entering $\mathcal{L}(l-1)$ and the first state visited in $\mathcal{L}_2(l)$ is in $\{(l,j',i');\ j'\in\mathcal{J}_2\}$. Since every service begins with the initial distribution $\boldsymbol{\beta}_2$, $\boldsymbol{K}_{12}$ is given by $\boldsymbol{K}_{12} = \tilde{\boldsymbol{K}}_{12}(\boldsymbol{\beta}_2\otimes\boldsymbol{I})$. Hence we obtain

$$\tilde{\boldsymbol{\pi}}_2^*(z) = \left\{\boldsymbol{\pi}(0)(\boldsymbol{I}\otimes\boldsymbol{D})\tilde{\boldsymbol{A}}' + \tilde{\boldsymbol{\pi}}_1^*(z)(\boldsymbol{I}\otimes\boldsymbol{D})\tilde{\boldsymbol{A}}\right\}(\boldsymbol{\beta}_2\otimes\boldsymbol{I}_A)z\boldsymbol{N}_0(\boldsymbol{I}-z\boldsymbol{R}_0)^{-1}, \tag{28}$$

where $\tilde{\boldsymbol{A}} = (\boldsymbol{U}_{11}\otimes\boldsymbol{I})\tilde{\boldsymbol{K}}_{12} + (\boldsymbol{u}_{11}\otimes\boldsymbol{I})$ and $\tilde{\boldsymbol{A}}' = (\boldsymbol{U}'_{11}\otimes\boldsymbol{I})\tilde{\boldsymbol{K}}_{12} + (\boldsymbol{u}'_{11}\otimes\boldsymbol{I})$. In this case, the base model becomes a MAP/PH/1 queue.

5.2.3 Poisson arrival and i.i.d. PH-type (IID-type) service model

Here we further assume that the arrival process is Poissonian with intensity λ. In this case, $\boldsymbol{K}_{12}$ is given by $\boldsymbol{K}_{12} = \boldsymbol{k}_{12}\boldsymbol{\beta}_2$, where $\boldsymbol{k}_{12}$ is a substochastic column vector. The base model becomes an M/PH/1 queue. From equations (23) and (28), we obtain

$$\tilde{\boldsymbol{\pi}}_1^*(z) = \boldsymbol{\pi}(0)z\lambda\boldsymbol{U}'_{11}\boldsymbol{N}_{11}(\boldsymbol{I} - z\boldsymbol{R}_{11})^{-1}, \tag{29}$$

$$\tilde{\boldsymbol{\pi}}_2^*(z) = \{\boldsymbol{\pi}(0)(\boldsymbol{U}'_{11}\boldsymbol{k}_{12} + \boldsymbol{u}'_{11}) + \tilde{\boldsymbol{\pi}}_1^*(z)(\boldsymbol{U}_{11}\boldsymbol{k}_{12} + \boldsymbol{u}_{11})\}\,\boldsymbol{\beta}_2 z\boldsymbol{R}_0(\boldsymbol{I} - z\boldsymbol{R}_0)^{-1}, \tag{30}$$

where $\boldsymbol{R}_{11} = \lambda\boldsymbol{U}_{11}\boldsymbol{N}_{11}$ and $\boldsymbol{R}_0 = \lambda\boldsymbol{N}_0$. Let ρ_0 denote the traffic intensity of the base model; ρ_0 is given by $\rho_0 = \lambda\boldsymbol{\beta}_2(-\boldsymbol{B}_2)^{-1}\boldsymbol{e}$. $\tilde{\boldsymbol{\pi}}_0^*(z)$ and $\boldsymbol{\pi}_0^*(z)$ are given by

$$\begin{aligned}
\tilde{\boldsymbol{\pi}}_0^*(z) &= (1-\rho_0)\boldsymbol{\beta}_2 z\boldsymbol{R}_0(\boldsymbol{I} - z\boldsymbol{R}_0)^{-1}, \\
\boldsymbol{\pi}_0^*(z) &= (1-\rho_0)\boldsymbol{\beta}_2(\boldsymbol{I} - z\boldsymbol{R}_0)^{-1}.
\end{aligned}$$

In terms of $\tilde{\boldsymbol{\pi}}_0^*(z)$, $\tilde{\boldsymbol{\pi}}_2^*(z)$ is given by

$$\tilde{\boldsymbol{\pi}}_2^*(z) = \{\boldsymbol{\pi}(0)(\boldsymbol{U}'_{11}\boldsymbol{k}_{12} + \boldsymbol{u}'_{11}) + \tilde{\boldsymbol{\pi}}_1^*(z)(\boldsymbol{U}_{11}\boldsymbol{k}_{12} + \boldsymbol{u}_{11})\}\,\frac{\tilde{\boldsymbol{\pi}}_0^*(z)}{1-\rho_0}. \tag{31}$$

If the MSP is of HM-type, i.e., $\boldsymbol{U}'_{11} = \boldsymbol{U}_{11}$, then we obtain

$$\begin{aligned}
\boldsymbol{\pi}_1^*(z) &= \boldsymbol{\pi}(0)(\boldsymbol{I} - z\boldsymbol{R}_{11})^{-1}, \\
\boldsymbol{\pi}_2^*(z) &= \boldsymbol{\pi}_1^*(z)(\boldsymbol{U}_{11}\boldsymbol{k}_{12} + \boldsymbol{u}_{11})\frac{\tilde{\boldsymbol{\pi}}_0^*(z)}{1-\rho_0}.
\end{aligned}$$

Furthermore, if $\boldsymbol{U}_{11} = \boldsymbol{I}$ then $\boldsymbol{u}_{11} = \boldsymbol{0}$, and we obtain

$$\boldsymbol{\pi}_2^*(z) = \boldsymbol{\pi}_1^*(z)\boldsymbol{k}_{12}\frac{\tilde{\boldsymbol{\pi}}_0^*(z)}{1-\rho_0}.$$

Let L be a random variable being subject to the stationary queue length distribution at an arbitrary time in the original model and let $\pi^*(z)$ be defined by $\pi^*(z) = \mathrm{E}[z^L]$. Then, we obtain

$$\pi^*(z) = \boldsymbol{\pi}_1^*(z)\boldsymbol{e}_1 + \boldsymbol{\pi}_2^*(z)\boldsymbol{e}_2 = \boldsymbol{\pi}_1^*(z)(\boldsymbol{e}_1 - \boldsymbol{k}_{12}) + \boldsymbol{\pi}_1^*(z)\boldsymbol{k}_{12}\frac{\boldsymbol{\pi}_0^*(z)\boldsymbol{e}_2}{1-\rho_0}. \tag{32}$$

In this formula, $\boldsymbol{k}_{12}$ is the column vector of the probabilities that the server can begin primary service in a non-empty period. Here we define a non-empty period as the interval between a time point when a customer arrives at the system of empty state and the time point when the system becomes empty again. Equation (32) can, therefore, be recognized as a conditional stochastic decomposition for the stationary queue length. Note that the case of $\boldsymbol{U}'_{11} = \boldsymbol{U}_{11} = \boldsymbol{I}$ includes the M/PH/1 queue with exceptional services and that with preemptive-repeat working vacations. If the MSP of the original model is of VA-type, $\boldsymbol{k}_{12}$ becomes stochastic, i.e., $\boldsymbol{k}_{12} = \boldsymbol{e}$, and we obtain the ordinary (unconditional) stochastic decomposition.

5.3 The Case of Nonpreemptive Service Type (NP-Type)

Here we consider the MSP of NP-type, i.e., $\boldsymbol{S}_{21} = \boldsymbol{O}$, $\boldsymbol{U}_{21} = \boldsymbol{O}$ and $\boldsymbol{U}_{22} = \boldsymbol{I}$. In this case, how to specify the base model is not trivial. Let $\boldsymbol{S}_0$ and $\boldsymbol{U}_0$ be set at $\boldsymbol{S}_{22}$ and $\boldsymbol{I}$, respectively, and let $\boldsymbol{T}_0$ be temporally set at an arbitrary non-negative matrix satisfying $(\boldsymbol{S}_0 + \boldsymbol{T}_0)\boldsymbol{e} = \boldsymbol{0}$. In general, since $\boldsymbol{K}_{21}$ is nonzero, we cannot take the same approach as used in the previous subsection.

Let $\boldsymbol{G}$ be the fundamental matrix of the original model, i.e., $[\boldsymbol{G}]_{(j,i)(j',i')}$ is the probability that a sample path of $\{Y(t)\}$ starting in the state $(l+1, j, i)$ for some $l \geq 1$ visits $\mathcal{L}(l)$ and

(l, j', i') is the first state visited in $\mathcal{L}(l)$. Since $\{Y(t)\}$ is a QBD process, $\boldsymbol{G}$ is stochastic and is given by $\boldsymbol{G} = \boldsymbol{N}(\boldsymbol{T} \otimes \boldsymbol{I})$. According to the partition of $\mathcal{J}$, $\boldsymbol{G}$ has the block form

$$\boldsymbol{G} = \begin{pmatrix} \boldsymbol{G}_{11} & \boldsymbol{G}_{12} \\ \boldsymbol{G}_{21} & \boldsymbol{G}_{22} \end{pmatrix}.$$

Let $\boldsymbol{G}'_{22}$ be defined as $\boldsymbol{G}'_{22} = \boldsymbol{G}_{22} + \boldsymbol{G}_{21}\boldsymbol{K}_{12}$. $[\boldsymbol{G}'_{22}]_{(j,i)(j',i')}$, $j, j' \in \mathcal{J}_2$, is the probability that a sample path of $\{Y(t)\}$ starting in the state $(l+1, j, i)$ for some $l \geq 1$ visits $\mathcal{L}_2(l)$ before entering $\mathcal{L}(l-1)$ and (l, j', i') is the first state visited in $\mathcal{L}_2(l)$. $\boldsymbol{N}_{22}$ satisfies

$$\boldsymbol{N}_{22} = (-\boldsymbol{S}_{22} \oplus \boldsymbol{C})^{-1} + (-\boldsymbol{S}_{22} \oplus \boldsymbol{C})^{-1}(\boldsymbol{I} \oplus \boldsymbol{D})\boldsymbol{G}'_{22}\boldsymbol{N}_{22}. \tag{33}$$

On the other hand, in the base model, $\boldsymbol{N}_0$ satisfies

$$\boldsymbol{N}_0 = (-\boldsymbol{S}_{22} \oplus \boldsymbol{C})^{-1} + (-\boldsymbol{S}_{22} \oplus \boldsymbol{C})^{-1}(\boldsymbol{I} \oplus \boldsymbol{D})\boldsymbol{G}_0\boldsymbol{N}_0, \tag{34}$$

where $\boldsymbol{G}_0$ is the fundamental matrix of the base model. Hence, if $\boldsymbol{G}'_{22} = \boldsymbol{G}_0$ then $\boldsymbol{N}_{22} = \boldsymbol{N}_0$. It seems difficult to obtain general conditions on which $\boldsymbol{G}'_{22} = \boldsymbol{G}_0$ hold. One sufficient condition but trivial one is that the arrival process is Poissonian with intensity λ and the MSP of the original model is of NP, IID and VA-type. In that case, letting $\boldsymbol{T}_0$ be set at $\boldsymbol{b}_2\boldsymbol{\beta}_2$, we obtain $\boldsymbol{G}'_{22} = \boldsymbol{G}_0 = \boldsymbol{e}\boldsymbol{\beta}_2$. The base model becomes an M/PH/1 queue, and we obtain the next lemma.

Lemma 4 *If the arrival process is Poissonian and the MSP of the original model is NP, IID and VA-type, we obtain*

$$\tilde{\boldsymbol{\pi}}^*_2(z) = \frac{\pi^*_1(z)}{1-\rho_0}\tilde{\boldsymbol{\pi}}^*_0(z), \tag{35}$$

where $\pi^*_1(z) = \boldsymbol{\pi}(0)\boldsymbol{e}' + \tilde{\boldsymbol{\pi}}^*_1(z)\boldsymbol{e}_1$. *Let L be a random variable being subject to the stationary queue length distribution at an arbitrary time in the original model, and let $\pi^*(z)$ be defined by* $\pi^*(z) = \mathrm{E}[z^L]$. *From equation (35), $\pi^*(z)$ is given by*

$$\pi^*(z) = \boldsymbol{\pi}(0)\boldsymbol{e}' + \tilde{\boldsymbol{\pi}}^*_1(z)\boldsymbol{e}_1 + \tilde{\boldsymbol{\pi}}^*_2(z)\boldsymbol{e}_2 = \frac{\pi^*_1(z)}{1-\rho_0}\boldsymbol{\pi}^*_0(z)\boldsymbol{e}_2. \tag{36}$$

□

Remark 6 *In Lemma 4, $\boldsymbol{T}_{21}$ can be arbitrary set as long as it satisfies $\boldsymbol{T}_{21}\boldsymbol{e} + \boldsymbol{t}_{22} = \boldsymbol{b}_2$. This means that the length of a vacation may depend on the service phase in which the preceding service has just ended.* □

Equation (36) corresponds to the stochastic decomposition for the stationary queue length in the M/G/1 queue with generalized vacations, derived by Fuhrmann & Cooper [9].

6 Computation of Stationary Queue Length Distribution and Numerical Examples

In Section 4 we showed that, in the MAP/MSP/1 queue of HM-type, $\boldsymbol{\pi}^*_2(z)$ was factorized into three parts (cf. equation (19)), and in Section 5 we presented two special cases where the third part of $\boldsymbol{\pi}^*_2(z)$ was given by another MAP/MSP/1 queue called a base model. Taking account of those results, we here discuss computation of the stationary queue length distribution in the MAP/MSP/1 queue of EX-type and give some numerical examples. Since the stochastic decomposition that we obtained for a MAP/MSP/1 queue of NP-type (cf. Lemma 4) corresponds to that derived in Ref. [9], we do not mention it in this section. Note that, since the behavior of a MAP/MSP/1 queue is represented as a QBD process, the stationary queue length distribution of the MAP/MSP/1 queue is given by the matrix-geometric solution and the rate matrix can be directly calculated by using the existing algorithms [13]. However, for the MAP/MSP/1 queue of EX-type, we can separately calculate each block of the rate matrix according to the matrix type factorization.

6.1 Computation for the MAP/MSP/1 Queue of EX-Type

For simplicity, we here deal with a MAP/MSP/1 queue in which the MSP is of EX and HM-type, and use notations described in Subsection 5.2. The results below can easily be extended to a MAP/MSP/1 queue whose MSP is of EX-type but not HM-type. When the MSP is of EX and HM-type, $\boldsymbol{\pi}_1^*(z)$ is given by equation (25) and $\boldsymbol{\pi}_2^*(z)$ by equation (26), where $\boldsymbol{\pi}_2^*(z)$ is factorized into $\boldsymbol{\pi}_1^*(z)$, $(\boldsymbol{I}\otimes\boldsymbol{D})\hat{\boldsymbol{K}}_{12}$ and $z\boldsymbol{N}_0(\boldsymbol{I}-z\boldsymbol{R}_0)^{-1}$. To apply these equations, we need to know the values of $\boldsymbol{\pi}(0)$, $\boldsymbol{N}_{11}$, $\boldsymbol{N}_0$ and $\boldsymbol{K}_{12}$; the values of $\boldsymbol{R}_{11}$, $\boldsymbol{R}_0$ and $\hat{\boldsymbol{K}}_{12}$ are obtained from those of $\boldsymbol{N}_{11}$, $\boldsymbol{N}_0$ and $\boldsymbol{K}_{12}$, respectively. $\boldsymbol{N}_0$ is obtained from the base model of the original MAP/MSP/1 queue. Since the base model is also a MAP/MSP/1 queue, the value of $\boldsymbol{N}_0$ $(=\boldsymbol{N}_{22})$ is given by the minimal non-negative solution of the following equation of $\boldsymbol{X}$.

$$(\boldsymbol{U}_{22}\otimes\boldsymbol{D})+(\boldsymbol{U}_{22}\otimes\boldsymbol{D})\boldsymbol{X}(\boldsymbol{S}_{22}\oplus\boldsymbol{C})+\Big((\boldsymbol{U}_{22}\otimes\boldsymbol{D})\boldsymbol{X}\Big)^2(\boldsymbol{T}_{22}\otimes\boldsymbol{I})=\boldsymbol{O}$$

Hence we can use existing algorithms for calculating the value of $\boldsymbol{N}_0$. $\boldsymbol{N}_{11}$ is obtained by analyzing a certain Markov chain of GI/M/1-type (referred as Submodel 1) and its value is also given by the minimal non-negative solution of the following equation of $\boldsymbol{X}$.

$$(\boldsymbol{U}_{11}\otimes\boldsymbol{D})+(\boldsymbol{U}_{11}\otimes\boldsymbol{D})\boldsymbol{X}(\boldsymbol{S}_{11}\oplus\boldsymbol{C})+\Big((\boldsymbol{U}_{11}\otimes\boldsymbol{D})\boldsymbol{X}\Big)^2(\boldsymbol{T}_{11}\otimes\boldsymbol{I})=\boldsymbol{O}$$

As a result, the values of $\boldsymbol{N}_0$ and $\boldsymbol{N}_{11}$ can independently obtained from the base model and Submodel 1, respectively. However, $\boldsymbol{K}_{12}$ is the matrix of the probabilities that connect the base model and Submodel 1 (cf. Remark 2) and we have to consider them at one time in order to obtain the value of $\boldsymbol{K}_{12}$.

The value of $\boldsymbol{K}_{12}$ can be obtained as follows. Consider the fundamental matrix $\boldsymbol{G}$ of the original model, which has the block form

$$\boldsymbol{G}=\begin{pmatrix}\boldsymbol{G}_{11} & \boldsymbol{G}_{12}\\ \boldsymbol{O} & \boldsymbol{G}_{22}\end{pmatrix}.$$

Since $\boldsymbol{G}$ is represented as $\boldsymbol{G}=\boldsymbol{N}(\boldsymbol{T}\otimes\boldsymbol{I})$, its blocks are given by

$$\boldsymbol{G}_{11}=\boldsymbol{N}_{11}(\boldsymbol{T}_{11}\otimes\boldsymbol{I}),\quad \boldsymbol{G}_{22}=\boldsymbol{N}_0(\boldsymbol{T}_{22}\otimes\boldsymbol{I}),\quad \boldsymbol{G}_{12}=\boldsymbol{N}_{11}(\boldsymbol{T}_{12}\otimes\boldsymbol{I})+\boldsymbol{K}_{12}\boldsymbol{G}_{22}.$$

Considering the first transition of $Y(t)$ from $\mathcal{L}_1(l)$ to the outside of $\mathcal{L}_1(l)$, we obtain the following equation of $\boldsymbol{K}_{12}$:

$$\begin{aligned}\boldsymbol{K}_{12} &= \boldsymbol{P}_1+\boldsymbol{P}_2\boldsymbol{G}_{22}+\boldsymbol{P}_3\Big(\boldsymbol{G}_{11}\boldsymbol{K}_{12}+\boldsymbol{G}_{12}\Big)\\ &= \boldsymbol{P}_1+\boldsymbol{P}_2\boldsymbol{G}_{22}+\boldsymbol{P}_3\boldsymbol{N}_{11}(\boldsymbol{T}_{12}\otimes\boldsymbol{I})+\boldsymbol{P}_3\Big(\boldsymbol{G}_{11}\boldsymbol{K}_{12}+\boldsymbol{K}_{12}\boldsymbol{G}_{22}\Big),\end{aligned}\tag{37}$$

where $\boldsymbol{P}_1=(-\boldsymbol{S}_{11}\oplus\boldsymbol{C})^{-1}(\boldsymbol{S}_{12}\otimes\boldsymbol{I})$, $\boldsymbol{P}_2=(-\boldsymbol{S}_{11}\oplus\boldsymbol{C})^{-1}(\boldsymbol{U}_{12}\otimes\boldsymbol{D})$ and $\boldsymbol{P}_3=(-\boldsymbol{S}_{11}\oplus\boldsymbol{C})^{-1}(\boldsymbol{U}_{11}\otimes\boldsymbol{D})$. On the right hand side of the first line of equation (37), the first term corresponds to the case where a sample path of $\{Y(t)\}$ starting in $\mathcal{L}_1(l)$ leaves $\mathcal{L}_1(l)$ and the first state visited in the outside of $\mathcal{L}_1(l)$ is in $\mathcal{L}_2(l)$, the second term to the case where that state is in $\mathcal{L}_2(l+1)$ and the third term to the case where that state is in $\mathcal{L}_1(l+1)$. From equation (37), the value of $\boldsymbol{K}_{12}$ is given by the minimal non-negative solution of the following equation of $\boldsymbol{X}$.

$$\boldsymbol{X}=\boldsymbol{P}_1+\boldsymbol{P}_2\boldsymbol{G}_{22}+\boldsymbol{P}_3\boldsymbol{N}_{11}(\boldsymbol{T}_{12}\otimes\boldsymbol{I})+\boldsymbol{P}_3\Big(\boldsymbol{G}_{11}\boldsymbol{X}+\boldsymbol{X}\boldsymbol{G}_{22}\Big)$$

The solution can be numerically obtained by iteration with the initial value of $\boldsymbol{X} = \boldsymbol{O}$. The value of $\boldsymbol{\pi}(0)$ is obtained by solving equations (4) and (5).

Remark 7 *Since $\boldsymbol{G}_{11} + \boldsymbol{G}_{12}$ and $\boldsymbol{G}_{22}$ are stochastic, we obtain*

$$(\boldsymbol{G}_{11} + \boldsymbol{G}_{12})\boldsymbol{e} = \boldsymbol{G}_{11}\boldsymbol{e} + \boldsymbol{N}_{11}(\boldsymbol{T}_{12} \otimes \boldsymbol{I})\boldsymbol{e} + \boldsymbol{K}_{12}\boldsymbol{e} = \boldsymbol{e}.$$

This leads us to

$$\boldsymbol{K}_{12}\boldsymbol{e} = \Big(\boldsymbol{I} - \boldsymbol{G}_{11} - \boldsymbol{N}_{11}(\boldsymbol{T}_{12} \otimes \boldsymbol{I})\Big)\boldsymbol{e},$$

and hence $\boldsymbol{K}_{12}\boldsymbol{e}$ is represented in terms of $\boldsymbol{N}_{11}$. As mentioned in Subsection 5.2, in the case where the arrival process is Poissonian and the MSP is of EX and IID-type, $\boldsymbol{K}_{12}$ is given as $\boldsymbol{K}_{12} = \boldsymbol{K}_{12}\boldsymbol{e}\boldsymbol{\beta}_2 = \boldsymbol{k}_{12}\boldsymbol{\beta}_2$. Hence, in that case, the values of $\boldsymbol{K}_{12}$ is given by the product of two terms, $\boldsymbol{k}_{12}$ and $\boldsymbol{\beta}_2$, which are independently obtained from Submodel 1 and the base model, respectively.

6.2 Numerical Examples

Here we show some numerical examples for a preemptive-resume working vacation model and for an exceptional service model presented in Subsection 2.4. Since the MSPs of both the models are of EX and HM-type, we can numerically obtain the stationary queue length distribution by using the method described in the previous subsection.

Consider a MAP whose representation $(\boldsymbol{C}, \boldsymbol{D})$ is given by

$$\boldsymbol{C} = \begin{pmatrix} -(\gamma_1 + \lambda_1) & \gamma_1 \\ \gamma_2 & -(\gamma_2 + \lambda_2) \end{pmatrix}, \quad \boldsymbol{D} = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}.$$

This MAP is a Markov modulated Poisson process (MMPP) and its mean arrival rate $\bar{\lambda}$ is given by $\bar{\lambda} = (\gamma_1\lambda_2 + \gamma_2\lambda_1)/(\gamma_1 + \gamma_2)$. We use this MAP in both the models. In the working vacation model, let vacation times be subject to a 2-Erlang distribution with mean h_1 and ordinary service times to another 2-Erlang distribution with mean h_2. Hence we have

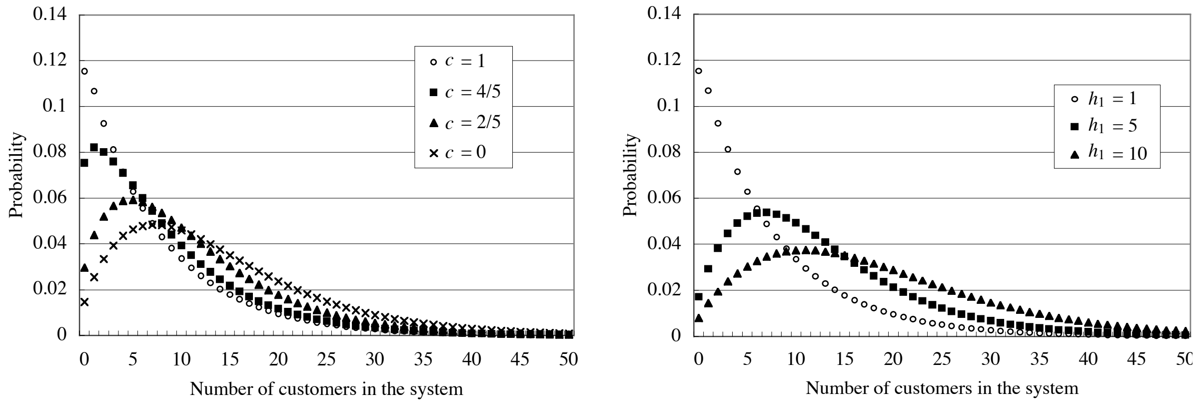
$$\boldsymbol{B}_i = \begin{pmatrix} -2/h_i & 2/h_i \\ 0 & -2/h_i \end{pmatrix}, \ i = 1, 2, \quad \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = (\,1 \quad 0\,).$$

We denote by c the ratio of service speed in the working vacation model. In the exceptional service model, we assume that at most two customers firstly arriving in each busy period receive different service from that received by other customers (i.e. $k = 2$). Let the service times of the first and second customers be subject to a common 2-Erlang distribution with mean h'_1. Then we have

$$\boldsymbol{B}_{1i} = \begin{pmatrix} -2/h'_1 & 2/h'_1 \\ 0 & -2/h'_1 \end{pmatrix}, \ i = 1, 2, \quad \boldsymbol{\beta}_{11} = \boldsymbol{\beta}_{12} = (\,1 \quad 0\,).$$

Let the service times of other customers be subject to the same distribution as that used for ordinary service times in the working vacation model.

Figure 1 shows the queue length distributions of the models, where the values of $\boldsymbol{\pi}(l)\boldsymbol{e}$, $l = 0, 1, ..., 50$, are plotted on each graph. The parameters of the MAP are set as $\gamma_1 = \frac{1}{2}$, $\gamma_2 = \frac{4}{5}$, $\lambda_1 = \frac{1}{2}$ and $\lambda_2 = \frac{3}{2}$, and we obtain $\bar{\lambda} = \frac{23}{26}$. The mean ordinary service time is set as $h_2 = 1$. The traffic intensity ρ defined as $\rho = \bar{\lambda} h_1$ is equal to $\frac{23}{26}$. In the working vacation model, h_1 is set at 10 and the ratio of service speed, c, takes values in $\{0, \frac{2}{5}, \frac{4}{5}, 1\}$. When $c = 0$, the model becomes an MMPP/E_2/1 queue with multiple vacations; when $c = 1$, it becomes an MMPP/E_2/1 queue without vacations. From Fig. 1 (a), we can see how the ratio of service speed influences the queue length distribution. In the exceptional service model, the mean service time of the first and second customers, h'_1, takes values in $\{1, 5, 10\}$. When $h'_1 = 1$, the model becomes an ordinary MMPP/E_2/1 queue. From Fig. 1 (b), we can also see how the value of h'_1 influences the queue length distribution.



(a) Working vacation model

(b) Exceptional service model

Figure 1: Queue length distributions.

7 Conclusions

We studied MAP/MSP/1 queues and obtained a new sort of matrix-type factorization of the vector generating function for the stationary queue length. The MAP/MSP/1 queue is a very tractable model since its behavior is represented as a quasi-birth-and-death process. Furthermore, the MAP/MSP/1 queue can represent various queueing models such as vacation models, N-policy models and exceptional service models. Hence there is a great advantage in using it as a fundamental model for analyzing queueing models with various service disciplines. One extension of the MAP/MSP/1 queue is a model in which the arrival process is governed by another MAP when the system is empty. In the case of Poisson arrival, such a model was studied in Refs [6, 18]. Our results would also hold in that extended model.

Acknowledgment

The author would like to thank the anonymous referees for their valuable comments.

References

[1] Baba, Y. On M/G/1 Queues with the First N Customers of Each Busy Period Receiving Exceptional Services. J. of Operations Research Society of Japan **1999**, *42* (4), 490–500.

[2] Bellman, R. *Introduction to Matrix Analysis*, 2nd Ed.; SIAM: Philadelphia, 1997.

[3] Chang, S.H.; Takine, T.; Chae, K.C.; Lee, H.W. A Unified Queue Length Formula for BMAP/G/1 Queue with Generalized Vacations. Stochastic Models **2002**, *18* (3), 369–386.

[4] Doshi, B.T. A Note on Stochastic Decomposition in a GI/G/1 Queue with Vacations or Set-Up Times. J. of Applied Probability **1985**, *22*, 419–428.

[5] Doshi, B.T. Queueing Systems with Vacations – A Survey. Queueing Systems **1986**, *1*, 29–66.

[6] Doshi, B.T. Conditional and Unconditional Distributions for M/G/1 Type Queues with Server Vacations. Queueing Systems **1990**, *7*, 229–252.

[7] Doshi, B.T. Generalizations of the Stochastic Decomposition Results for Single Server Queues with Vacations. Stochastic Models **1990**, *6* (2), 307–333.

[8] Federgruen, A.; Green, L. Queueing Systems with Service Interruptions. Operations Research **1986**, *34*, 752–768.

[9] Fuhrmann, S.W.; Cooper, R.B. Stochastic Decompositions in the M/G/1 Queue with Generalized Vacations. Operations Research **1985**, *33* (5), 1117–1129.

[10] Heyman, D.P. Optimal Operating Policies for M/G/1 Queueing Systems. Operations Research **1968**, *16*, 362–382.

[11] Keilson, J.; Servi, L.D. Oscillating Random Walk Models for GI/G/1 Vacation Systems with Bernoulli Schedules. J. of Applied Probability **1986**, *23*, 790–802.

[12] Kengaku, H.; Miyazawa, M. A Regenerative Cycle Approach to an M/G/1 Queue with Exceptional Service. J. of Operations Research Society of Japan **2000**, *43* (4), 486–504.

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

[14] Lee, H.W.; Ahn, B.Y.; Park, N.I. Decompositions of the Queue Length Distributions in the MAP/G/1 Queue under Multiple and Single Vacations with N-Policy. Stochastic Models **2001**, *17* (2), 157–190.

[15] Lucantoni, D.M.; Meier-Hellstern, K.S.; Neuts, M. A Single-Server Queue with Server Vacations and a Class of Non-Renewal Arrival Processes. Advances in Applied Probability **1990**, *22*, 676–705.

[16] Neuts, M.F. *Structured Stochastic Matrices of M/G/1 Type and Their Applications*; Marcel Dekker, Inc.: New York, 1989.

[17] Servi, L.D.; Finn, S.G. M/M/1 Queues with Working Vacations. Performance Evaluation **2002**, *50* (1), 41–52.

[18] Shanthikumar, J.G. On Stochastic Decomposition in M/G/1 Type Queues with Generalized Server Vacations. Operations Research **1988**, *36* (4), 566–569.

[19] Tian, N.; Zhang, D.; Cao, C. The GI/M/1 Queue with Exponential Vacations. Queueing Systems **1989**, *5*, 331–344.

[20] Tian, N.; Li, Q.-L.; Cao, J. Conditional Stochastic Decompositions in the M/M/c Queue with Server Vacations. Stochastic Models **1999**, *15* (2), 367–377.

[21] Welch, P.D. On a Generalized M/G/1 Queueing Process in Which the First Customer of Each Busy Period Receives Exceptional Service. Operations Research **1964**, *12*, 736–752.

[22] Wu, D.; Takagi, H. M/G/1 Queue with Multiple Working Vacations. Proceedings of the Queueing Symposium: Stochastic Models and Their Applications, Kakegawa, Japan, Jan 20–22, 2003, 51–60.