Sojourn Time Distributions in the Queue Defined by a General QBD Process*

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

We consider a general QBD process as defining a FIFO queue and obtain the stationary distribution of the sojourn time of a customer in that queue as a matrix exponential distribution, which is identical to a phase-type distribution under a certain condition. Since QBD processes include many queueing models where the arrival and service process are dependent, these results form a substantial generalization of analogous results reported in the literature for queues such as the PH/PH/c queue. We also discuss asymptotic properties of the sojourn time distribution through its matrix exponential form.

Key wards: Quasi-birth-and-death process, queueing model, sojourn time distribution, matrix-exponential distribution

1 Introduction

Let $\mathcal{Z}_+$ denote the set of nonnegative integers and consider a continuous-time stochastic process $\{L(t)\}$ on $\mathcal{Z}_+$, where sample paths of the stochastic process are assumed to be right continuous. Then, letting $L(t)$ be the number of customers in the system at time t, we can regard $\{L(t)\}$ as a queueing model with a single-class of customers. If $L(t-) < L(t)$, then t is an arrival epoch of customer; if $L(t-) > L(t)$, then t is a departure epoch of customer. Here we assume that if $L(t-) = L(t)$, then any arrival or departure did not occur at time t. Let τ_n^A and τ_n^D be the nth arrival and departure epochs, where we number them so that $\tau_0^A \le 0$, $\tau_1^A > 0$, $\tau_0^D \le 0$, and $\tau_1^D > 0$. For $s, t \ge 0$, let $A(s,t)$ be the number of customers arriving in $(s,t]$ and $D(s,t)$ that of customers departing in the same interval. They are given by

$$A(s,t) = \sum_{n=1}^{\infty} 1_{\{s<\tau_n^A \le t\}} \left(L(\tau_n^A) - L(\tau_n^A -) \right),$$

$$D(s,t) = \sum_{n=1}^{\infty} 1_{\{s<\tau_n^D \le t\}} \left(L(\tau_n^D -) - L(\tau_n^D) \right),$$

where $1_{\{\cdot\}}$ is the indicator function. If $L(\tau_n^A) - L(\tau_n^A -) = 1$ for all n, then we say that the model has a single-arrival process; otherwise it has a batch-arrival process. If $L(\tau_n^D -) - L(\tau_n^D) = 1$ for all n, then we say that customers are served by single-service; otherwise they are served by batch-service. In the discussion hereafter, we assume single-arrival and single-service, and refer to the customer arriving at τ_n^A as customer n. In order to define sojourn times of customers

*This manuscript is a draft version of "T. Ozawa, Sojourn Time Distributions in the Queue Defined by a General QBD Process, Queueing Systems 53(4) (2006), 203-211."

in the system, we assume the first-in-first-out (FIFO) service discipline for the queueing model. Then the departure time of customer n, denoted by $\hat{\tau}_n^D$, and the sojourn time of the same customer, denoted by V_n, are given by

$$\begin{aligned}&\hat{\tau}_n^D = \inf\{t > \tau_n^A : D(\tau_n^A, t) = L(\tau_n^A)\},\\&V_n = \hat{\tau}_n^D - \tau_n^A.\end{aligned}$$

Note that we need more assumptions in order to define waiting times of customers for the queueing model. Assuming the existence of the stationary version of V_n, we denote it by V. Furthermore, we denote by L the stationary version of $L(t)$.

Next, we consider another stochastic process $\{J(t)\}$ on a state space $\mathcal{S}_0$ and assume that two-dimensional stochastic process $\{(L(t), J(t))\}$ is Markovian and time-homogeneous. If $\mathcal{S}_0$ is a finite set and the transition rate of the process $\{(L(t), J(t))\}$ does not depend on $L(t)$ when $L(t) > 0$, then $\{(L(t), J(t))\}$ becomes a quasi-birth-and-death (QBD) process. Our aim is to obtain the distribution of V for that QBD process. The simplest model among QBD processes is a birth-and-death (BD) process, where $\mathcal{S}_0$ is a singleton and the queueing model defined by the BD process is an M/M/1 queue. For the M/M/1 queue, the distribution of L is geometric and that of V exponential. In the case where $\mathcal{S}_0$ is not a singleton, it is well known that the distribution of L is given in a form of matrix-geometric [7, 10]. In this paper, we reveal that in that case the distribution of V can be represented in a form of matrix-exponential and, under a certain condition, the distribution is of phase-type. To our knowledge, it has not been known before. Of course, the stationary waiting time distribution of a GI/N/1 queue is matrix exponential [13, 14], but the class of queueing models defined by QBD processes includes not only MAP/PH/1 queues but also Markovian queueing models in which arrival and service processes are mutually dependent. One example of those Markovian queueing models is a MAP/MSP/1 queue [11], where MSP is the abbreviation of Markovian service process; the MSP is a model similar to a Markovian arrival process (MAP) [7, 8], 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 present some numerical examples for an N-policy model and an exceptional service model represented as MAP/MSP/1 queues. We also discuss asymptotic properties of the sojourn time distribution through the matrix exponential form.

The rest of the paper is organized as follows. In Sec. 2, the QBD process we consider is described in detail. In Sec. 3, the stationary sojourn time distribution as well as its asymptotic property is derived. In Sec. 4, we briefly explain a MAP/MSP/1 queue and give some numerical examples. The paper is concluded with a remark for a bath-service model in Sec. 5.

2 QBD process

Consider a QBD process $\{Y(t)\} = \{(L(t), J(t))\}$ on state space $\mathcal{S} = (\{0\} \times \mathcal{J}_B) \cup (\mathcal{N}_+ \times \mathcal{J}_A)$, where $L(t)$ and $J(t)$ are the level and phase at time t, $\mathcal{J}_A = \{1, 2, ..., s_A\}$ and $\mathcal{J}_B = \{1, 2, ..., s_B\}$ are its phase sets, and $\mathcal{N}_+$ is the set of positive integers. Let the infinitesimal generator of the QBD process be

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

where $\boldsymbol{A}(i)$, $i = 1, 2, 3$, are $s_A \times s_A$ matrices, $\boldsymbol{B}(0)$ an $s_B \times s_A$ matrix, $\boldsymbol{B}(1)$ an $s_B \times s_B$ matrix, and $\boldsymbol{B}(2)$ an $s_A \times s_B$ matrix. We assume that the QBD process has the stationary distribution, and denote it by $\boldsymbol{\pi} = (\, \boldsymbol{\pi}(0) \quad \boldsymbol{\pi}(1) \quad \boldsymbol{\pi}(2) \quad \cdots)$. Consider a period of time that begins when the process is in state (l, j) for some $l \in \{1, 2, ...\}$ and ends when it enters level $l-1$ for the first time. Let $n_{jj'}$ be the mean sojourn time of the process in state (l, j') during the period and let $\boldsymbol{N}$ be the $s_A \times s_A$ matrix defined by $\boldsymbol{N} = (n_{jj'})$. Let $\boldsymbol{R}$ $(= \boldsymbol{A}(0)\boldsymbol{N})$ denote the rate matrix of the QBD process, then the matrix geometric solution of the stationary distribution [7] is given by

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

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

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

where $\boldsymbol{I}$ is the identity matrix, $\boldsymbol{0}$ a column vector of 0's, and $\boldsymbol{e}$ as well as $\boldsymbol{e}'$ a column vector of 1's; the superscript $\top$ indicates the transpose.

Now we assume that the QBD process represents the behavior of a queue and regard the level as the number of customers in the system. An event that the level goes up by one corresponds to an arrival of customer and an event that the level goes down by one to a departure of customer. Assuming the FIFO discipline for the queue, we can define sojourn times of customers in the system in the same manner as in Section 1, without specifying service times of customers. Let V denote the sojourn time of a tagged customer in steady state, and without loss of generality we assume that the customer arrives at time 0. Then, V is equal to $\tau^D_{L(0)}$, which is the time epoch at which the $L(0)$th departure (downward jump of the level) counted from time 0 occurs.

3 Stationary Sojourn Time Distribution

3.1 Preliminaries

3.1.1 The steady state distribution just after an arrival epoch

Let $\hat{Y}_n = (\hat{L}_n, \hat{J}_n)$ denote the state of the QBD process just after the nth arrival epoch, i.e. $\hat{Y}_n = Y(\tau^A_n)$, then $\{\hat{Y}_n\}$ becomes a discrete-time GI/M-type Markov chain whose steady state distribution, denoted by $\hat{\boldsymbol{\pi}} = (\, \hat{\boldsymbol{\pi}}(1) \quad \hat{\boldsymbol{\pi}}(2) \quad \cdots)$, is given by the next theorem.

Theorem 1 *$\hat{\boldsymbol{\pi}}$ is given by*

$$\begin{aligned} &\hat{\boldsymbol{\pi}}(1) = \hat{\phi}\boldsymbol{\pi}(0)\boldsymbol{B}(0), \\ &\hat{\boldsymbol{\pi}}(l) = \hat{\phi}\boldsymbol{\pi}(l-1)\boldsymbol{A}(0) = \hat{\phi}\boldsymbol{\pi}(0)\boldsymbol{B}(0)\left(\boldsymbol{N}\boldsymbol{A}(0)\right)^{l-1} = \hat{\boldsymbol{\pi}}(1)\hat{\boldsymbol{R}}^{l-1},\ l \geq 2, \end{aligned} \tag{3}$$

where $\hat{\phi}$ is the normalizing constant given by

$$\hat{\phi} = \left\{\boldsymbol{\pi}(0)\boldsymbol{B}(0)(\boldsymbol{I} - \boldsymbol{N}\boldsymbol{A}(0))^{-1}\boldsymbol{e}\right\}^{-1}$$

and $\hat{\boldsymbol{R}} = \boldsymbol{N}\boldsymbol{A}(0)$ is the rate matrix of the discrete-time Markov chain $\{\hat{Y}_n\}$.

Proof. $[\boldsymbol{\pi}(0)\boldsymbol{B}(0)]_j$ is the rate that the QBD process enters state $(1, j)$ just after an arrival epoch in steady state and, for $l \in \{2, 3, ...\}$, $[\boldsymbol{\pi}(l-1)\boldsymbol{A}(0)]_j$ the rate that it enters state (l, j) just after an arrival epoch, where a rate means the expected number of events occurring in a unit time. Hence, by a standard argument for Markov chains [5], normalizing those rates, we obtain the stationary distribution of $\{\hat{Y}_n\}$ represented by equation (3).

Next, we shall show that $\hat{\boldsymbol{R}}$ is the rate matrix of the discrete-time Markov chain $\{\hat{Y}_n\}$. Suppose that, at the n_1th arrival epoch, $\tau^A_{n_1}$, the QBD process enters state (l, i) for some $l \in$

$\{1,2,...\}$, i.e. $\hat{Y}_{n_1} = Y(\tau^A_{n_1}) = (l,i)$, and consider the period of time that begins at $\tau^A_{n_1}$ and ends when the QBD process enters level $l-1$ for the first time. We denote by t' the ending time of the period and assume that just m customers arrive in $(\tau^A_{n_1}, t')$. From the definition of the period, we have $L(t) \geq l$ for all $t \in (\tau^A_{n_1}, t')$, and this implies that $\hat{L}_{n_1+k} \geq l+1$, $k = 1,2,...,m$. After ending the period, $L(t)$ becomes less than l and the level at the next arrival epoch is less than $l+1$, i.e. $\hat{L}_{n_1+m+1} < l+1$. Hence $[\boldsymbol{N}\boldsymbol{A}(0)]_{i,j}$ is the expected number of visits of the chain $\{\hat{Y}_n\}$ into state $(l+1,j)$ during the period of discrete-time that begins when the chain is in (l,i) and ends when it reenters level l or enters one of the levels less than l for the first time. This implies that $\hat{\boldsymbol{R}} = \boldsymbol{N}\boldsymbol{A}(0)$ is the rate matrix of $\{\hat{Y}_n\}$. □

We denote by $\hat{\boldsymbol{\eta}}$ the steady-state distribution of the phase just after an arrival epoch, which is given by

$$\hat{\boldsymbol{\eta}} = \sum_{l=1}^{\infty} \hat{\boldsymbol{\pi}}(l) = \hat{\boldsymbol{\pi}}(1)(\boldsymbol{I} - \hat{\boldsymbol{R}})^{-1}. \tag{4}$$

3.1.2 The number of departures in $(0,t]$

Let $O(t) = \max\{n : \tau^D_n \leq t\}$ denote the number of departures (downward jumps of the level) in $(0,t]$ and $\boldsymbol{P}(k,t)$ the matrix whose (i,j) element is $P(O(t) = k, J(t) = j \mid J(0) = i)$. Consider a MAP with representation $(\boldsymbol{C}, \boldsymbol{D})$, where $\boldsymbol{C} = (c_{ij}) = \boldsymbol{A}(0) + \boldsymbol{A}(1)$ and $\boldsymbol{D} = (d_{ij}) = \boldsymbol{A}(2)$. If the level of the QBD process does not become zero during $(0,t]$, the counting process $\{(O(t), J(t))\}$ can be represented by the MAP in stochastic sense. Hence, in that case, the matrices $\boldsymbol{P}(k,t)$, $k = 0,1,...,$ satisfy differential equations

$$\begin{aligned} &\frac{\partial}{\partial t}\boldsymbol{P}(0,t) = \boldsymbol{P}(0,t)\boldsymbol{C} = \boldsymbol{C}\boldsymbol{P}(0,t), \\ &\frac{\partial}{\partial t}\boldsymbol{P}(k,t) = \boldsymbol{P}(k,t)\boldsymbol{C} + \boldsymbol{P}(k-1,t)\boldsymbol{D} = \boldsymbol{C}\boldsymbol{P}(k,t) + \boldsymbol{D}\boldsymbol{P}(k-1,t),\ k \geq 1, \end{aligned} \tag{5}$$

and the initial conditions $\boldsymbol{P}(0,0) = \boldsymbol{I}$ and $\boldsymbol{P}(k,0) = \boldsymbol{O}$, $k \geq 1$, where $\boldsymbol{O}$ is a matrix of 0's.

3.2 Complementary distribution of V

In the following sections, we denote by $\otimes$ the Kronecker product operation and by $\oplus$ the Kronecker sum operation [4]. Let an $s_A^2 \times 1$ vector $\boldsymbol{\xi}$ be defined by

$$\boldsymbol{\xi} = \begin{pmatrix} \boldsymbol{e}_1 \\ \boldsymbol{e}_2 \\ \vdots \\ \boldsymbol{e}_{s_A} \end{pmatrix},$$

where $\boldsymbol{e}_k$ denote the $s_A \times 1$ vector whose kth element is 1 and whose other elements are all 0 (i.e. kth unit vector). This $\boldsymbol{\xi}$ has the following property.

Lemma 1 *For $1 \times s_A$ vectors $\boldsymbol{a} = (a_i)$ and $\boldsymbol{b} = (b_i)$, we have*

$$(\boldsymbol{a} \otimes \boldsymbol{b})\,\boldsymbol{\xi} = \boldsymbol{a}\boldsymbol{b}^\top = \boldsymbol{b}\boldsymbol{a}^\top.$$

Proof. This formula follows $(\boldsymbol{a} \otimes \boldsymbol{b})\,\boldsymbol{\xi} = \sum_{i=1}^{s_A} a_i \boldsymbol{b}\boldsymbol{e}_i = \sum_{i=1}^{s_A} a_i b_i$. □

The complementary distribution of V is given by the next theorem.

Theorem 2 *The complementary distribution of the sojourn time is given by*

$$P(V > t) = \left(\boldsymbol{e}^\top \otimes \hat{\boldsymbol{\eta}}\right) \exp\left(\left[\boldsymbol{C}^\top \otimes \boldsymbol{I} + \boldsymbol{D}^\top \otimes \hat{\boldsymbol{R}}\right] t\right) \boldsymbol{\xi}, \tag{6}$$

and the mean sojourn time by

$$E[V] = \int_0^\infty P(V > t)\, dt = \left(\boldsymbol{e}^\top \otimes \hat{\boldsymbol{\eta}}\right) \left(- \left[\boldsymbol{C}^\top \otimes \boldsymbol{I} + \boldsymbol{D}^\top \otimes \hat{\boldsymbol{R}}\right]\right)^{-1} \boldsymbol{\xi}. \tag{7}$$

Proof. If $O(t) < L(0)$, the tagged customer arriving at time 0 is still in the system at time t. Hence, in terms of $\hat{\boldsymbol{\pi}}(k)$ and $\boldsymbol{P}(k,t)$, the complementary probability of V is represented as

$$P(V > t) = \sum_{n=1}^{\infty} \hat{\boldsymbol{\pi}}(n) \sum_{k=0}^{n-1} \boldsymbol{P}(k,t)\boldsymbol{e} = \sum_{n=1}^{\infty} \hat{\boldsymbol{\pi}}(1)\hat{\boldsymbol{R}}^{n-1} \sum_{k=0}^{n-1} \boldsymbol{P}(k,t)\boldsymbol{e} = \sum_{k=0}^{\infty} \hat{\boldsymbol{\eta}}\hat{\boldsymbol{R}}^{k} \boldsymbol{P}(k,t)\boldsymbol{e}. \tag{8}$$

From Lemma 1, we therefore obtain

$$P(V > t) = \sum_{k=0}^{\infty} \left(\boldsymbol{e}^\top \boldsymbol{P}^\top(k,t) \otimes \hat{\boldsymbol{\eta}}\hat{\boldsymbol{R}}^{k}\right) \boldsymbol{\xi} = \left(\boldsymbol{e}^\top \otimes \hat{\boldsymbol{\eta}}\right) \sum_{k=0}^{\infty} \left(\boldsymbol{P}^\top(k,t) \otimes \hat{\boldsymbol{R}}^{k}\right) \boldsymbol{\xi}. \tag{9}$$

In addition, from the differential equations (5), we have

$$\begin{aligned} \frac{\partial}{\partial t}\boldsymbol{P}^\top(0,t) \otimes \boldsymbol{I} &= \left(\boldsymbol{P}^\top(0,t) \otimes \boldsymbol{I}\right)\left(\boldsymbol{C}^\top \otimes \boldsymbol{I}\right), \\ \frac{\partial}{\partial t}\boldsymbol{P}^\top(k,t) \otimes \hat{\boldsymbol{R}}^{k} &= \left(\boldsymbol{P}^\top(k,t) \otimes \hat{\boldsymbol{R}}^{k}\right)\left(\boldsymbol{C}^\top \otimes \boldsymbol{I}\right) \\ &\qquad + \left(\boldsymbol{P}^\top(k-1,t) \otimes \hat{\boldsymbol{R}}^{k-1}\right)\left(\boldsymbol{D}^\top \otimes \hat{\boldsymbol{R}}\right),\ k \geq 1. \end{aligned} \tag{10}$$

Because the original QBD process is assumed to have a stationary distribution, $\sum_{k=0}^{\infty} \hat{\boldsymbol{R}}^{k} = \boldsymbol{I} + \left(\boldsymbol{N} \sum_{k=1}^{\infty} \boldsymbol{R}^{k-1}\right) \boldsymbol{A}(0)$ is finite and $\sum_{k=0}^{\infty} \boldsymbol{P}^\top(k,t) \otimes \hat{\boldsymbol{R}}^{k}$ uniformly converges elementwise. Hence we can exchange $\sum$ and $\partial/\partial t$ on the left hand side of the sum of equations (10) and obtain the following matrix-differential equation with the initial condition $\sum_{k=0}^{\infty} \boldsymbol{P}^\top(k,0) \otimes \hat{\boldsymbol{R}}^{k} = \boldsymbol{I} \otimes \boldsymbol{I}$.

$$\frac{d}{dt}\sum_{k=0}^{\infty} \left(\boldsymbol{P}^\top(k,t) \otimes \hat{\boldsymbol{R}}^{k}\right) = \sum_{k=0}^{\infty} \left(\boldsymbol{P}^\top(k,t) \otimes \hat{\boldsymbol{R}}^{k}\right) \left(\boldsymbol{C}^\top \otimes \boldsymbol{I} + \boldsymbol{D}^\top \otimes \hat{\boldsymbol{R}}\right) \tag{11}$$

This implies

$$\sum_{k=0}^{\infty} \left(\boldsymbol{P}^\top(k,t) \otimes \hat{\boldsymbol{R}}^{k}\right) = \exp\left(\left[\boldsymbol{C}^\top \otimes \boldsymbol{I} + \boldsymbol{D}^\top \otimes \hat{\boldsymbol{R}}\right] t\right). \tag{12}$$

Equation (6) follows this and equation (9). □

Remark 1 *Denoting $\sum_{k=0}^{\infty} \hat{\boldsymbol{R}}^{k} \boldsymbol{P}(k,t)$ by $\boldsymbol{X}(t)$ and substituting it for equation (8), we obtain $P(V > t) = \hat{\boldsymbol{\eta}}\boldsymbol{X}(t)\boldsymbol{e}$. Furthermore, through a procedure similar to that used for deriving equation (11), we obtain the differential equation of $\boldsymbol{X}(t)$ as follows:*

$$\frac{d}{dt}\boldsymbol{X}(t) = \boldsymbol{X}(t)\boldsymbol{C} + \hat{\boldsymbol{R}}\boldsymbol{X}(t)\boldsymbol{D}, \quad \boldsymbol{X}(0) = \boldsymbol{I}. \tag{13}$$

If $\hat{\boldsymbol{R}}$ commutes with $\boldsymbol{X}(t)$, the matrix-differential equation (13) has the matrix-exponential solution $\boldsymbol{X}(t) = \exp([\boldsymbol{C} + \hat{\boldsymbol{R}}\boldsymbol{D}]t)$. An example where commutability does hold is the M/PH/1 queue. However, in our model that commutability does not always hold true. Because $\left(\boldsymbol{e}^\top \otimes \hat{\boldsymbol{\eta}}\right) = \hat{\boldsymbol{\eta}}\left(\boldsymbol{e}^\top \otimes \boldsymbol{I}\right)$, it can be seen that $\boldsymbol{X}(t)\boldsymbol{e}$ is given by

$$\boldsymbol{X}(t)\boldsymbol{e} = \left(\boldsymbol{e}^\top \otimes \boldsymbol{I}\right) \exp\left(\left[\boldsymbol{C}^\top \otimes \boldsymbol{I} + \boldsymbol{D}^\top \otimes \hat{\boldsymbol{R}}\right] t\right) \boldsymbol{\xi}.$$

Remark 2 *Applying the uniformization technique [7] to* $\boldsymbol{C}$ *and* $\boldsymbol{D}$*, we obtain the following formula, which helps us compute the value of* $P(V > t)$*:*

$$\exp\Big(\Big[\boldsymbol{C}^\top \otimes \boldsymbol{I} + \boldsymbol{D}^\top \otimes \hat{\boldsymbol{R}}\Big]\, t\Big) = \sum_{k=0}^{\infty} e^{-\nu t}\frac{(\nu t)^k}{k!}\left(\boldsymbol{P}^\top(0) \otimes \boldsymbol{I} + \boldsymbol{P}^\top(1) \otimes \hat{\boldsymbol{R}}\right)^k, \tag{14}$$

where $\nu = \max_{1\le i\le s_A} |c_{ii}|$*,* $\boldsymbol{P}(0) = \boldsymbol{I} + \frac{1}{\nu}\boldsymbol{C}$*, and* $\boldsymbol{P}(1) = \frac{1}{\nu}\boldsymbol{D}$*. From the definitions, both* $\boldsymbol{P}(0)$ *and* $\boldsymbol{P}(1)$ *are nonnegative, and* $\boldsymbol{P}(0) + \boldsymbol{P}(1)$ *stochastic.*

Remark 3 *If* $\boldsymbol{C} + \boldsymbol{D} = \boldsymbol{A}(0) + \boldsymbol{A}(1) + \boldsymbol{A}(2)$ *is irreducible, then* $\boldsymbol{C} + \boldsymbol{D}$ *has the stationary probability vector, which is strictly positive. Let* $\boldsymbol{x}$ *denote the vector, and let* $\boldsymbol{\Delta}$ *the diagonal matrix defined by* $\boldsymbol{\Delta} = \mathrm{diag}(\boldsymbol{x})$*, where* $\mathrm{diag}(\boldsymbol{x})$ *denotes the diagonal matrix whose diagonal elements are the elements of vector* $\boldsymbol{x}$*. We define a reversed MAP of the MAP with representation* $(\boldsymbol{C}, \boldsymbol{D})$ *as that with representation* $(\tilde{\boldsymbol{C}}, \tilde{\boldsymbol{D}})$*, where* $\tilde{\boldsymbol{C}} = (\tilde{c}_{ij}) = \boldsymbol{\Delta}^{-1}\boldsymbol{C}^\top\boldsymbol{\Delta}$ *and* $\tilde{\boldsymbol{D}} = (\tilde{d}_{ij}) = \boldsymbol{\Delta}^{-1}\boldsymbol{D}^\top\boldsymbol{\Delta}$ *[12]. Equation (6) means that the distribution of* V *is a matrix-exponential distribution with representation* $(\boldsymbol{\alpha}, \boldsymbol{T}, \boldsymbol{s})$ *[1], where* $\boldsymbol{\alpha} = \boldsymbol{e}^\top \otimes \hat{\boldsymbol{\eta}}$*,* $\boldsymbol{T} = \boldsymbol{C}^\top \otimes \boldsymbol{I} + \boldsymbol{D}^\top \otimes \hat{\boldsymbol{R}}$*, and* $\boldsymbol{s} = \boldsymbol{\xi}$*. Let* $\boldsymbol{\alpha}'$*,* $\boldsymbol{T}'$*, and* $\boldsymbol{s}'$ *be defined as follows:*

$$\begin{aligned}
&\boldsymbol{\alpha}' = \boldsymbol{\alpha}\,(\boldsymbol{\Delta} \otimes \boldsymbol{I}) = (\boldsymbol{x} \otimes \hat{\boldsymbol{\eta}}),\\
&\boldsymbol{T}' = (\boldsymbol{\Delta} \otimes \boldsymbol{I})^{-1}\,\boldsymbol{T}\,(\boldsymbol{\Delta} \otimes \boldsymbol{I}) = \tilde{\boldsymbol{C}} \otimes \boldsymbol{I} + \tilde{\boldsymbol{D}} \otimes \hat{\boldsymbol{R}},\\
&\boldsymbol{s}' = (\boldsymbol{\Delta} \otimes \boldsymbol{I})^{-1}\,\boldsymbol{s} = \left(\boldsymbol{\Delta}^{-1} \otimes \boldsymbol{I}\right)\boldsymbol{\xi}.
\end{aligned}$$

Then, $(\boldsymbol{\alpha}', \boldsymbol{T}', \boldsymbol{s}')$ *becomes another representation for the matrix-exponential distribution of* V*, and we obtain the next formula.*

$$P(V > t) = (\boldsymbol{x} \otimes \hat{\boldsymbol{\eta}})\exp\Big(\Big[\tilde{\boldsymbol{C}} \otimes \boldsymbol{I} + \tilde{\boldsymbol{D}} \otimes \hat{\boldsymbol{R}}\Big]\, t\Big)\left(\boldsymbol{\Delta}^{-1} \otimes \boldsymbol{I}\right)\boldsymbol{\xi} \tag{15}$$

Remark 4 *If the time between two consecutive departure epochs corresponds to the service time of the customer departing at the latter departure epoch, we can define waiting times of customers on a QBD process and derive the stationary waiting time distribution in the same manner as that used for deriving equation (6). Let* W *be the waiting time of an arbitrary customer in steady state, then the complementary distribution of* W *is given by*

$$P(W > t) = \left(\boldsymbol{e}^\top \otimes \hat{\boldsymbol{\eta}}\hat{\boldsymbol{R}}\right)\exp\Big(\Big[\boldsymbol{C}^\top \otimes \boldsymbol{I} + \boldsymbol{D}^\top \otimes \hat{\boldsymbol{R}}\Big]\, t\Big)\,\boldsymbol{\xi}, \tag{16}$$

and from this we obtain $P(W = 0) = 1 - P(W > 0) = \hat{\boldsymbol{\eta}}\hat{\boldsymbol{R}}\boldsymbol{e}$*.*

When $\hat{\boldsymbol{\eta}}$ defined by equation (4) is strictly positive, the stationary sojourn time distribution is of phase type.

Corollary 1 *Assume that* $\hat{\boldsymbol{\eta}}$ *is strictly positive, and define a* $1 \times s_A^2$ *vector* $\boldsymbol{\kappa}$ *and an* $s_A^2 \times s_A^2$ *matrix* $\boldsymbol{K}$ *as* $\boldsymbol{\kappa} = \boldsymbol{\xi}^\top\left(\boldsymbol{I} \otimes \hat{\boldsymbol{\Delta}}\right)$ *and* $\boldsymbol{K} = \boldsymbol{C} \otimes \boldsymbol{I} + \boldsymbol{D} \otimes \hat{\boldsymbol{\Delta}}^{-1}\hat{\boldsymbol{R}}^\top\hat{\boldsymbol{\Delta}}$*, where* $\hat{\boldsymbol{\Delta}} = \mathrm{diag}(\hat{\boldsymbol{\eta}})$*. Then, the stationary sojourn time distribution is identical to the phase-type distribution with representation* $(\boldsymbol{K}, \boldsymbol{\kappa})$*.*

Proof. First, we shall prove that $\hat{\boldsymbol{\Delta}}^{-1}\hat{\boldsymbol{R}}^\top\hat{\boldsymbol{\Delta}}$ is substochastic. The nonnegativity of $\hat{\boldsymbol{\Delta}}^{-1}\hat{\boldsymbol{R}}^\top\hat{\boldsymbol{\Delta}}$ follows from that of $\hat{\boldsymbol{R}}$. In addition, we have

$$\boldsymbol{e} - \hat{\boldsymbol{\Delta}}^{-1}\hat{\boldsymbol{R}}^\top\hat{\boldsymbol{\Delta}}\,\boldsymbol{e} = \hat{\boldsymbol{\Delta}}^{-1}\hat{\boldsymbol{\eta}}^\top - \hat{\boldsymbol{\Delta}}^{-1}\hat{\boldsymbol{R}}^\top\hat{\boldsymbol{\eta}}^\top = \hat{\boldsymbol{\Delta}}^{-1}\left\{\hat{\boldsymbol{\eta}}(\boldsymbol{I} - \hat{\boldsymbol{R}})\right\}^\top = \hat{\boldsymbol{\Delta}}^{-1}\hat{\boldsymbol{\pi}}(1)^\top \ge \boldsymbol{0},$$

and from this we obtain $\hat{\boldsymbol{\Delta}}^{-1}\hat{\boldsymbol{R}}^{\top}\hat{\boldsymbol{\Delta}}\,\boldsymbol{e} \le \boldsymbol{e}$. Because of the stationarity assumption of the original QBD process, the spectral radius of $\hat{\boldsymbol{R}}$ is less than one and that of $\hat{\boldsymbol{\Delta}}^{-1}\hat{\boldsymbol{R}}^{\top}\hat{\boldsymbol{\Delta}}$ is also less than one. This implies that $\hat{\boldsymbol{\Delta}}^{-1}\hat{\boldsymbol{R}}^{\top}\hat{\boldsymbol{\Delta}}$ is strictly substochastic.

Next, we shall prove that $\boldsymbol{\kappa}$ is a probability vector and $\boldsymbol{K}$ is a phase generator. The non-negativity of $\boldsymbol{\kappa}$ is obvious, and the next formula indicates that the sum of $\boldsymbol{\kappa}$'s elements is one.

$$\boldsymbol{\kappa}\,(\boldsymbol{e}\otimes\boldsymbol{e}) = \boldsymbol{\xi}^{\top}\left(\boldsymbol{I}\otimes\hat{\boldsymbol{\Delta}}\right)(\boldsymbol{e}\otimes\boldsymbol{e}) = \boldsymbol{\xi}^{\top}\left(\boldsymbol{e}\otimes\hat{\boldsymbol{\eta}}^{\top}\right) = \left(\boldsymbol{e}^{\top}\otimes\hat{\boldsymbol{\eta}}\right)\boldsymbol{\xi} = \hat{\boldsymbol{\eta}}\boldsymbol{e} = 1$$

In this calculation, we use Lemma 1. Since $\boldsymbol{C}+\boldsymbol{D}$ has negative diagonal elements and non-negative off-diagonal elements and $\hat{\boldsymbol{\Delta}}^{-1}\hat{\boldsymbol{R}}^{\top}\hat{\boldsymbol{\Delta}}$ is substochastic, $\boldsymbol{K} = \boldsymbol{C}\otimes\boldsymbol{I}+\boldsymbol{D}\otimes\hat{\boldsymbol{\Delta}}^{-1}\hat{\boldsymbol{R}}^{\top}\hat{\boldsymbol{\Delta}}$ has negative diagonal elements and non-negative off-diagonal elements. Furthermore, we have

$$\boldsymbol{K}\,(\boldsymbol{e}\otimes\boldsymbol{e}) = \boldsymbol{C}\boldsymbol{e}\otimes\boldsymbol{e}+\boldsymbol{D}\boldsymbol{e}\otimes\hat{\boldsymbol{\Delta}}^{-1}\hat{\boldsymbol{R}}^{\top}\hat{\boldsymbol{\Delta}}\boldsymbol{e} \le \boldsymbol{C}\boldsymbol{e}\otimes\boldsymbol{e}+\boldsymbol{D}\boldsymbol{e}\otimes\boldsymbol{e} = (\boldsymbol{C}+\boldsymbol{D})\boldsymbol{e}\otimes\boldsymbol{e} = \boldsymbol{0}.$$

From this and that $\hat{\boldsymbol{\Delta}}^{-1}\hat{\boldsymbol{R}}^{\top}\hat{\boldsymbol{\Delta}}$ is substochastic, $\boldsymbol{K}$ is a nonsingular phase generator.

Finally, the next formula shows that the distribution of V is of phase type with representation $(\boldsymbol{K},\boldsymbol{\kappa})$ [12].

$$\begin{aligned}
P(V>t) &= \left\{\left(\boldsymbol{e}^{\top}\otimes\hat{\boldsymbol{\eta}}\right)\exp\!\left(\left[\boldsymbol{C}^{\top}\otimes\boldsymbol{I}+\boldsymbol{D}^{\top}\otimes\hat{\boldsymbol{R}}\right]t\right)\boldsymbol{\xi}\right\}^{\top}\\
&= \boldsymbol{\xi}^{\top}\exp\!\left(\left[\boldsymbol{C}\otimes\boldsymbol{I}+\boldsymbol{D}\otimes\hat{\boldsymbol{R}}^{\top}\right]t\right)\left(\boldsymbol{e}\otimes\hat{\boldsymbol{\eta}}^{\top}\right)\\
&= \boldsymbol{\xi}^{\top}\left(\boldsymbol{I}\otimes\hat{\boldsymbol{\Delta}}\right)\exp\!\left(\left[\boldsymbol{C}\otimes\boldsymbol{I}+\boldsymbol{D}\otimes\hat{\boldsymbol{\Delta}}^{-1}\hat{\boldsymbol{R}}^{\top}\hat{\boldsymbol{\Delta}}\right]t\right)\left(\boldsymbol{I}\otimes\hat{\boldsymbol{\Delta}}^{-1}\right)\left(\boldsymbol{e}\otimes\hat{\boldsymbol{\eta}}^{\top}\right)\\
&= \boldsymbol{\kappa}\exp(\boldsymbol{K}t)\,(\boldsymbol{e}\otimes\boldsymbol{e})
\end{aligned}$$

□

3.3 Asymptotic properties of $P(V>0)$

For a square matrix $\boldsymbol{H}$, let $\mathrm{sp}(\boldsymbol{H})$ denote the set of eigenvalues of $\boldsymbol{H}$ and $\mathrm{spr}(\boldsymbol{H})$ the spectral radius of $\boldsymbol{H}$, i.e. $\mathrm{spr}(\boldsymbol{H}) = \max\{|\lambda| : \lambda\in\mathrm{sp}(\boldsymbol{H})\}$ [2]. From the Perron-Frobenius theorem, the rate matrix $\boldsymbol{R}$ has the real eigenvalue that equals $\mathrm{spr}(\boldsymbol{R})$, denoted by γ_R, and has the corresponding nonnegative left and right eigenvectors, denoted by $\boldsymbol{u}_R$ and $\boldsymbol{v}_R$. Since the QBD process we consider is assumed to have the stationary distribution, we have $0<\gamma_R<1$. Let ν be defined by $\nu = \max_{1\le i\le s_A}|c_{ii}|$ and consider nonnegative matrix $\boldsymbol{P}(x) = \boldsymbol{I}+\frac{1}{\nu}(\boldsymbol{C}+x\boldsymbol{D})$ for $x\in[0,1]$. Let $\lambda(x)$ be the maximum eigenvalue of $\boldsymbol{P}(x)$, i.e. $\lambda(x) = \mathrm{spr}(\boldsymbol{P}(x))$, and let $\boldsymbol{u}(x)$ and $\boldsymbol{v}(x)$ be the corresponding nonnegative left and right eigenvectors, where we assume that $\boldsymbol{u}(x)$ and $\boldsymbol{v}(x)$ are normalized by $\boldsymbol{u}(x)\boldsymbol{v}(x) = 1$. Then we obtain $\lambda(x) = \boldsymbol{u}(x)\boldsymbol{P}(x)\boldsymbol{v}(x)$ and

$$\frac{d}{dx}\lambda(x) = \boldsymbol{u}(x)\boldsymbol{D}\boldsymbol{v}(x) \ge 0.$$

Thus, when x varies from 0 to γ_R, $\lambda(x)$ takes the maximum value at $x=\gamma_R$. Let γ^* be given by $\gamma^* = -\nu(1-\lambda(\gamma_R))$, then γ^* is the real eigenvalue of $\boldsymbol{C}+\gamma_R\boldsymbol{D}$ that has the maximum real part. The assumption that the QBD process has the stationary distribution implies that $\lambda(\gamma_R) = \mathrm{spr}\Big(\boldsymbol{I}+\frac{1}{\nu}(\boldsymbol{C}+\gamma_R\boldsymbol{D})\Big) < \mathrm{spr}\Big(\boldsymbol{I}+\frac{1}{\nu}(\boldsymbol{C}+\boldsymbol{D})\Big) = \lambda(1) = 1$. Hence we have $\gamma^*<0$. Let us define some vectors as follows:

$$\boldsymbol{u}^* = \boldsymbol{u}(\gamma_R),\ \boldsymbol{v}^* = \boldsymbol{v}(\gamma_R),\ \boldsymbol{u}_{\hat{R}} = \boldsymbol{u}_R\boldsymbol{A}(0),\ \boldsymbol{v}_{\hat{R}} = \boldsymbol{N}\boldsymbol{v}_R.$$

The next lemma asserts that the maximum eigenvalue of $\boldsymbol{C}^{\top}\otimes\boldsymbol{I}+\boldsymbol{D}^{\top}\otimes\hat{\boldsymbol{R}}$ is given by γ^*.

Lemma 2 *γ^* is the eigenvalue of $\boldsymbol{C}^\top \otimes \boldsymbol{I} + \boldsymbol{D}^\top \otimes \hat{\boldsymbol{R}}$ that has the maximum real part, and the corresponding left and right eigenvectors are given by $(\boldsymbol{v}^*)^\top \otimes \boldsymbol{u}_{\hat{R}}$ and $(\boldsymbol{u}^*)^\top \otimes \boldsymbol{v}_{\hat{R}}$, respectively.*

Proof. Since $\boldsymbol{R} = \boldsymbol{A}(0)\boldsymbol{N}$ and $\hat{\boldsymbol{R}} = \boldsymbol{N}\boldsymbol{A}(0)$, $\mathrm{sp}(\boldsymbol{R}) = \mathrm{sp}(\hat{\boldsymbol{R}})$ and γ_R is also the maximum eigenvalue of $\hat{\boldsymbol{R}}$, where the corresponding left and right eigenvectors are given by $\boldsymbol{u}_{\hat{R}}$ and $\boldsymbol{v}_{\hat{R}}$, respectively. For $z \in \mathrm{sp}(\hat{\boldsymbol{R}})$, let $\boldsymbol{\zeta}(z)$ denote the left eigenvector of $\hat{\boldsymbol{R}}$ corresponding to the eigenvalue z. For a $1 \times s_A$ vector $\boldsymbol{a}$, we have

$$(\boldsymbol{a} \otimes \boldsymbol{\zeta}(z)) \left(\boldsymbol{C}^\top \otimes \boldsymbol{I} + \boldsymbol{D}^\top \otimes \hat{\boldsymbol{R}}\right) = \boldsymbol{a} \left(\boldsymbol{C}^\top + z\boldsymbol{D}^\top\right) \otimes \boldsymbol{\zeta}(z).$$

From this, we obtain

$$\mathrm{sp}\Big(\boldsymbol{C}^\top \otimes \boldsymbol{I} + \boldsymbol{D}^\top \otimes \hat{\boldsymbol{R}}\Big) = \bigcup_{z \in \mathrm{sp}(\hat{\boldsymbol{R}})} \mathrm{sp}\Big(\boldsymbol{C}^\top + z\boldsymbol{D}^\top\Big) = \bigcup_{z \in \mathrm{sp}(\boldsymbol{R})} \mathrm{sp}(\boldsymbol{C} + z\boldsymbol{D})\,.$$

In addition, since $\left(\boldsymbol{I} + \frac{1}{\nu}\boldsymbol{C}\right)^\top \otimes \boldsymbol{I} + \left(\frac{1}{\nu}\boldsymbol{D}\right)^\top \otimes \hat{\boldsymbol{R}}$ is nonnegative, it has the real eigenvalue that equals $\lambda^\dagger = \mathrm{spr}\left(\left(\boldsymbol{I} + \frac{1}{\nu}\boldsymbol{C}\right)^\top \otimes \boldsymbol{I} + \left(\frac{1}{\nu}\boldsymbol{D}\right)^\top \otimes \hat{\boldsymbol{R}}\right)$ and has the corresponding left eigenvector in real values. Denoting that eigenvector by $\boldsymbol{u}^\dagger \otimes \boldsymbol{\zeta}(z^\dagger)$, we have

$$\boldsymbol{u}^\dagger \left(\boldsymbol{I} + \frac{1}{\nu}\left(\boldsymbol{C} + z^\dagger \boldsymbol{D}\right)\right) = \boldsymbol{u}^\dagger \left(\boldsymbol{I} + \frac{1}{\nu}\boldsymbol{C}\right) + \frac{1}{\nu} z^\dagger \boldsymbol{u}^\dagger \boldsymbol{D} = \lambda^\dagger \boldsymbol{u}^\dagger.$$

This means that $z^\dagger$ is a real number or $\boldsymbol{u}^\dagger \boldsymbol{D}$ is equal to $\boldsymbol{0}^\top$. In the former case, $\lambda^\dagger$ has to be $\lambda(\gamma_R)$, since $z^\dagger \le \gamma_R$ and $\lambda(x)$ is nondecreasing with respect to x. In the latter case, we have $\boldsymbol{u}^\dagger \left(\boldsymbol{I} + \frac{1}{\nu}\boldsymbol{C}\right) = \boldsymbol{u}^\dagger \boldsymbol{P}(0) = \lambda^\dagger \boldsymbol{u}^\dagger$ and hence we obtain $\lambda^\dagger = \lambda(0) \le \lambda(\gamma_R)$. As a result, in both the cases, we obtain $\lambda^\dagger = \lambda(\gamma_R)$ and $\boldsymbol{u}^\dagger = \boldsymbol{u}(\gamma_R) = \boldsymbol{u}^*$. In the same manner, it can be seen that the corresponding right eigenvector is given by $\boldsymbol{v}(\gamma_R) = \boldsymbol{v}^*$. The assertion of the lemma follows $\gamma^* = -\nu(1 - \lambda(\gamma_R))$. □

Considering the Jordan canonical form of $\boldsymbol{C}^\top \otimes \boldsymbol{I} + \boldsymbol{D}^\top \otimes \hat{\boldsymbol{R}}$, we obtain the next lemma.

Lemma 3 *For some $k \in \{0, 1, ...\}$, we have*

$$\exp\Big(\Big[\boldsymbol{C}^\top \otimes \boldsymbol{I} + \boldsymbol{D}^\top \otimes \hat{\boldsymbol{R}}\Big]\, t\Big) = O(t^k e^{-|\gamma^*|t}). \tag{17}$$

In order to obtain an asymptotic constant for $P(V > t)$, we set the following assumption.

Assumption 1 *γ^* is simple, i.e. in the characteristic polynomial of $\boldsymbol{C}^\top \otimes \boldsymbol{I} + \boldsymbol{D}^\top \otimes \hat{\boldsymbol{R}}$, the multiplicity of γ^* is 1.*

Under this assumption, the dimension of the eigenspace corresponding to γ^* becomes 1, and every eigenvalue of $\boldsymbol{C}^\top \otimes \boldsymbol{I} + \boldsymbol{D}^\top \otimes \hat{\boldsymbol{R}}$ except for γ^* has a real part less than γ^*. Hence we obtain the next theorem.

Theorem 3 *Under Assumption 1, we have*

$$\lim_{t \to \infty} e^{|\gamma^*|t} P(V > t) = \hat{\boldsymbol{\eta}} \boldsymbol{N} \boldsymbol{v}_R \boldsymbol{u}_R \boldsymbol{A}(0) \boldsymbol{v}^* \boldsymbol{u}^* \boldsymbol{e}. \tag{18}$$

Proof. From Lemma 3 and Assumption 1, we obtain

$$\lim_{t\to\infty} e^{|\gamma^*|t} \exp\Big(\Big[\boldsymbol{C}^\top \otimes \boldsymbol{I} + \boldsymbol{D}^\top \otimes \hat{\boldsymbol{R}}\Big]\, t\Big) = \Big((\boldsymbol{u}^*)^\top \otimes \boldsymbol{v}_{\hat{R}}\Big)\,\Big((\boldsymbol{v}^*)^\top \otimes \boldsymbol{u}_{\hat{R}}\Big),$$

where $\boldsymbol{v}_{\hat{R}} = \boldsymbol{N}\boldsymbol{u}_R$ and $\boldsymbol{u}_{\hat{R}} = \boldsymbol{u}_R\boldsymbol{A}(0)$. Hence, through Lemma 1, the assertion of the theorem follows

$$\lim_{t\to\infty} e^{|\gamma^*|t} P(V > t) = \Big(\boldsymbol{e}^\top \otimes \hat{\boldsymbol{\eta}}\Big)\,\Big((\boldsymbol{v}^*\boldsymbol{u}^*)^\top \otimes \boldsymbol{N}\boldsymbol{v}_R\boldsymbol{u}_R\boldsymbol{A}(0)\Big)\,\boldsymbol{\xi}.$$

□

Remark 5 *The size of the matrices used for computing γ^* and the right hand side of equation (18) is $s_A \times s_A$. This is much smaller than the size of $\exp\Big(\Big[\boldsymbol{C}^\top \otimes \boldsymbol{I} + \boldsymbol{D}^\top \otimes \hat{\boldsymbol{R}}\Big]\, t\Big)$, which is $s_A^2 \times s_A^2$.*

Remark 6 *For a nonnegative $d \times d$ matrix $\boldsymbol{H} = (h_{ij})$, let the set of states for $\boldsymbol{H}$ be defined by $S = \{1, 2, ..., d\}$ and say that states i and j communicate if $h_{ij}^{(k)} > 0$ for some $k \in \{0, 1, ...\}$ and $h_{ji}^{(l)} > 0$ for some $l \in \{0, 1, ...\}$, where $\boldsymbol{H}^n = (h_{ij}^{(n)})$. Using this relation, we can define communication classes and irreducibility for nonnegative matrices in the same manner as that for stochastic matrices [2]. In our framework, the nonnegative matrix $\Big(\boldsymbol{I} + \frac{1}{\nu}\boldsymbol{C}^\top\Big) \otimes \boldsymbol{I} + \frac{1}{\nu}\boldsymbol{D}^\top \otimes \hat{\boldsymbol{R}}$ may not be irreducible and it may have several communication classes, say m communication classes. In that case, for $k \in \{1, 2, ..., m\}$, let square matrix $\boldsymbol{P}_k$ be the restriction of $\Big(\boldsymbol{I} + \frac{1}{\nu}\boldsymbol{C}^\top\Big) \otimes \boldsymbol{I} + \frac{1}{\nu}\boldsymbol{D}^\top \otimes \hat{\boldsymbol{R}}$ to communication class k, then $\boldsymbol{P}_k$ is irreducible and $\lambda_k = \mathrm{spr}(\boldsymbol{P}_k)$ becomes a simple eigenvalue of $\boldsymbol{P}_k$. Furthermore, we have*

$$\mathrm{sp}\Big(\Big(\boldsymbol{I} + \frac{1}{\nu}\boldsymbol{C}^\top\Big) \otimes \boldsymbol{I} + \frac{1}{\nu}\boldsymbol{D}^\top \otimes \hat{\boldsymbol{R}}\Big) = \bigcup_{k=1}^{m} \mathrm{sp}(\boldsymbol{P}_k).$$

Assumption 1 is, therefore, equivalent to that $\lambda(\gamma_R) = \max_k \mathrm{spr}(\boldsymbol{P}_k)$ is attained by only one communication class or that the matrix $\Big(\boldsymbol{I} + \frac{1}{\nu}\boldsymbol{C}^\top\Big) \otimes \boldsymbol{I} + \frac{1}{\nu}\boldsymbol{D}^\top \otimes \hat{\boldsymbol{R}}$ itself is irreducible.

4 MAP/MSP/1 queue and numerical examples

4.1 MAP/MSP/1 queue

At first, we briefly explain Markovian service process (MSP) [11]. Consider two sets of states (phases) for the server, $\mathcal{J}_1 = \{1, 2, ..., s_1\}$ and $\mathcal{J}_1' = \{1, 2, ..., s_1'\}$, and assume that when the system is not empty, the server's state is in $\mathcal{J}_1$; otherwise it is in $\mathcal{J}_1'$. Let $\boldsymbol{S}$ and $\boldsymbol{T}$ denote $s_1 \times s_1$ matrices and let $\boldsymbol{S} + \boldsymbol{T}$ be the infinitesimal generator of the continuous-time Markov chain that governs 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 $s_1 \times s_1$ transition probability matrix $\boldsymbol{U}$ which governs state transition of the server at customer arrival epochs. This $\boldsymbol{U}$ enables us to represent models in which the server changes its state at customer arrival epochs. An $s_1' \times s_1'$ matrix $\boldsymbol{S}'$, an $s_1 \times s_1'$ matrix $\boldsymbol{T}'$ and an $s_1' \times s_1$ matrix $\boldsymbol{U}'$ are similarly defined in the case where the system is empty. The MSP is represented by these six elements $(\boldsymbol{S}, \boldsymbol{T}, \boldsymbol{U}, \boldsymbol{S}', \boldsymbol{T}', \boldsymbol{U}')$.

Next, we consider a MAP with representation $(\bar{\boldsymbol{C}}, \bar{\boldsymbol{D}})$, where the phase set of the MAP is given by $\mathcal{I} = \{1, 2, ..., s_2\}$. We define the state of the system at time t by $Y(t) = (L(t), J(t), I(t))$, where $L(t)$ is the number of customers in the system, $J(t)$ the phase of the MSP, and $I(t)$ the phase of the MAP, and call it a MAP/MSP/1 queue. $\{Y(t)\}$ is the continuous-time Markov chain whose infinitesimal generator $\boldsymbol{Q}$ is given by the block tri-diagonal matrix

$$\boldsymbol{Q} = \begin{pmatrix} \boldsymbol{S}' \oplus \bar{\boldsymbol{C}} & \boldsymbol{U}' \otimes \bar{\boldsymbol{D}} & & & \\ \boldsymbol{T}' \otimes \boldsymbol{I} & \boldsymbol{S} \oplus \bar{\boldsymbol{C}} & \boldsymbol{U} \otimes \bar{\boldsymbol{D}} & & \\ & \boldsymbol{T} \otimes \boldsymbol{I} & \boldsymbol{S} \oplus \bar{\boldsymbol{C}} & \boldsymbol{U} \otimes \bar{\boldsymbol{D}} & \\ & & \ddots & \ddots & \ddots \end{pmatrix}. \tag{19}$$

This means that $\{Y(t)\}$ is a QBD process and that we can compute the sojourn time distribution by using the results in the previous section. In numerical examples, we deal with well-known two kinds of service models below, which are represented in block forms:

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

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

N-policy model [6]: N-policy is a service discipline in which once the server becomes idle, it does not begin service until more than N new customers arrive. Let the service time distribution be of phase type with representation $(\bar{\boldsymbol{B}}, \bar{\boldsymbol{\beta}})$. The representation of an MSP with N-policy is given by

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

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

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

where $\bar{\boldsymbol{b}} = -\bar{\boldsymbol{B}}\boldsymbol{e}$. $\boldsymbol{S}_{11}$, $\boldsymbol{T}_{11}$, and $\boldsymbol{U}_{11}$ are $N \times N$ matrices, i.e. $s_1 = N$. The MSP with N-policy is an example of service models in which the service process depends on the arrival process.

Exceptional service model [3, 15]: Exceptional service is a service model in which at most K customers firstly arriving in each busy period receive different service from that received by other customers. For $k \in \{1, 2, ..., K\}$, let the service time distribution of the kth customer arriving in each busy period be of phase type with representation $(\bar{\boldsymbol{B}}_{1k}, \bar{\boldsymbol{\beta}}_{1k})$. Let the service time distribution of other customers be of phase type with representation $(\bar{\boldsymbol{B}}_2, \bar{\boldsymbol{\beta}}_2)$. Then the representation of an MSP with exceptional service is given by

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

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

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

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

where $\bar{\boldsymbol{b}}_{1k} = -\bar{\boldsymbol{B}}_{1k}\boldsymbol{e}$, $k = 1, 2, ..., K$, and $\bar{\boldsymbol{b}}_2 = -\bar{\boldsymbol{B}}_2\boldsymbol{e}$. $\mathrm{diag}(\bar{\boldsymbol{B}}_{11}, ..., \bar{\boldsymbol{B}}_{1K})$ denotes the block diagonal matrix whose block diagonal elements are $\bar{\boldsymbol{B}}_{11}$, ..., $\bar{\boldsymbol{B}}_{1\,K-1}$, and $\bar{\boldsymbol{B}}_{1K}$.

4.2 Numerical examples

We show some numerical examples for the N-policy model and for the exceptional service model described in the previous subsection. Consider a MAP whose representation $(\bar{\boldsymbol{C}}, \bar{\boldsymbol{D}})$ is given by

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

This MAP is a Markov modulated Poisson process (MMPP) and its mean arrival rate λ is given by $\lambda = (\gamma_1\lambda_2 + \gamma_2\lambda_1)/(\gamma_1 + \gamma_2)$. We use this MAP in both the models. In the N-policy model, let service times be subject to a 2-Erlang distribution with mean h. Then we have

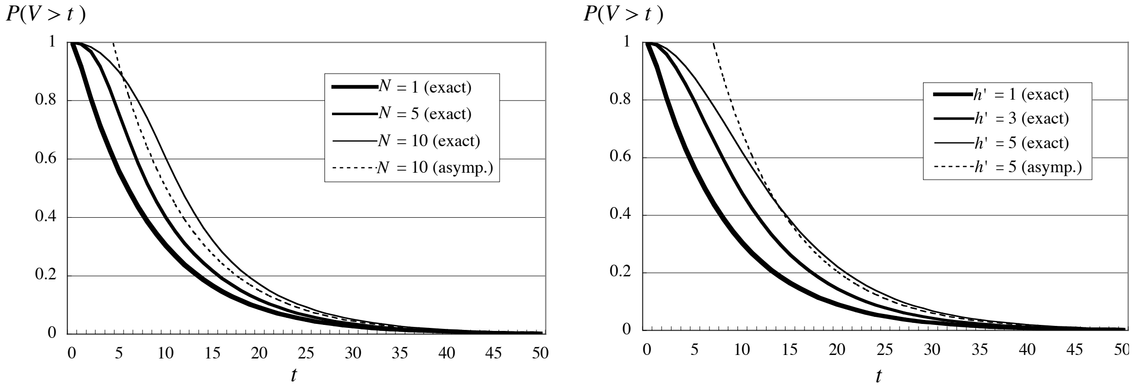
$$\bar{\boldsymbol{B}} = \begin{pmatrix} -2/h & 2/h \\ 0 & -2/h \end{pmatrix}, \quad \bar{\boldsymbol{\beta}} = (\,1 \quad 0\,).$$

For 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'. Then we have

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

Let the service times of other customers be subject to the same distribution as that used for the N-policy model, i.e. $\bar{\boldsymbol{B}}_2 = \bar{\boldsymbol{B}}$.

Figure 1 shows complementary distributions of sojourn time for the models, where *exact* indicates that numerical results are obtained through formula (6) and *asymp.* indicates that those are obtained through formula (18). The parameters of the MAP are set as $\gamma_1 = 1/2$, $\gamma_2 = 4/5$, $\lambda_1 = 1/2$ and $\lambda_2 = 3/2$, and we obtain $\lambda = 23/26$. The mean of ordinary service times is set as $h = 1$. Traffic intensity ρ defined by $\rho = \lambda h$ is equal to $23/26 \approx 0.88$. In the N-policy model, N takes values in $\{1, 5, 10\}$. From Fig. 1 (a), we can see how the value of N influences the sojourn time distribution. In the exceptional service model, the mean service time of the first and second customers, h', takes values in $\{1, 3, 5\}$. When $h' = 1$, the model corresponds to an ordinary MMPP/E$_2$/1 queue. From Fig. 1 (b), we can also see how the value of h'_1 influences the sojourn time distribution.



(a) N-policy model

(b) Exceptional service model

Figure 1: Complimentary distributions of sojourn time.

5 Concluding Remarks

The results in Section 3 can be extended to GI/M-type Markov chains [7, 10], which represent single-arrival batch-service queueing models. Let the infinitesimal generator of a GI/M-type Markov chain be

$$\boldsymbol{Q} = \begin{pmatrix} \boldsymbol{B}(1) & \boldsymbol{B}(0) & & & & \\ \boldsymbol{B}(2) & \boldsymbol{A}(1) & \boldsymbol{A}(0) & & & \\ \boldsymbol{B}(3) & \boldsymbol{A}(2) & \boldsymbol{A}(1) & \boldsymbol{A}(0) & & \\ \boldsymbol{B}(4) & \boldsymbol{A}(3) & \boldsymbol{A}(2) & \boldsymbol{A}(1) & \boldsymbol{A}(0) & \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}, \tag{20}$$

then the complementary distribution of the stationary sojourn time is obtained through the same arguments as in Section 3, as follows:

$$P(V > t) = \left(\boldsymbol{e}^{\top} \otimes \hat{\boldsymbol{\eta}}\right) \exp\left(\left[\sum_{l=0}^{\infty} \boldsymbol{D}_l^{\top} \otimes \hat{\boldsymbol{R}}^l\right] t\right) \boldsymbol{\xi}, \tag{21}$$

where $\boldsymbol{D}_0 = \boldsymbol{A}(0) + \boldsymbol{A}(1)$ and $\boldsymbol{D}_l = \boldsymbol{A}(l-1)$, $l \geq 1$; $\hat{\boldsymbol{\eta}}$ and $\hat{\boldsymbol{R}}$ are defined in the same manner as in Section 3, and $(\boldsymbol{D}_l,\, l \geq 0)$ corresponds to the representation of a batch Markovian arrival process (BMAP) [7, 9].

Acknowledgment

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

References

[1] S. Asmussen and M. Bladt, Renewal Theory and Queueing Algorithms for Matrix-Exponential Distributions, in: *Matrix-Analytic Methods in Stochastic Models*, eds. Chakravarthy and Alfa, Marcel Dekker, New York, 1997, pp. 313–341.

[2] S. Asmussen, *Applied Probability and Queues*, Springer-Verlag, New York, 2003.

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

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

[5] P. Brémaud, *Markov Chains Gibbs Fields, Monte Carlo Simulation, and Queues*, Springer-Verlag, New York, 1999.

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

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

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

[9] D. M. Lucantoni, New Results on the Single Server Queue with a Batch Markovian Arrival Process, Stochastic Models 7(1) (1991) 1–46.

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

[11] T. Ozawa, Analysis of Queues with Markovian Service Processes, Stochastic Models 20(4) (2004) 391–413.

[12] V. Ramaswami, A Duality Theorem for the Matrix Paradigms in Queueing Theory, Stochastic Models 6(1) (1990) 151–161.

[13] V. Ramaswami, From the Matrix-Geometric to the Matrix-Exponential, Queueing Systems 6 (1990) 229–260.

[14] 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(1) (1989) 159–180.

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