Journal of the Operations Research Society of Japan
Vol. 54, No. 1, March 2011, pp. 46–64

ASYMPTOTIC BEHAVIOR FOR MArP/PH/2 QUEUE WITH JOIN THE SHORTEST QUEUE DISCIPLINE

Yutaka Sakuma
Tokyo University of Science

(Received September 4, 2008; Revised October 25, 2010)

Abstract This paper considers a parallel queueing model with two servers, where arriving customers join the shortest queue. In [14], we studied a similar queueing model, and obtained the tail decay rate of the stationary distribution by using the matrix analytic approach. The main objectives of this paper are to extend the result in [14] to a more general model, and to clarify difficulty when we apply similar techniques as in [14]. In this paper, we study an MArP/PH/2 queue with join the shortest queue discipline, and show that the geometric tail asymptotics of the stationary distribution is obtained under a certain condition on the service time distribution.

Keywords: Queue, Markovian arrival process, join the shortest queue discipline, phase type distribution, tail decay rate, quasi-birth-and-death process

1. Introduction

We consider a parallel queueing model with two servers, where each server has a single waiting line with infinite capacity. Arriving customers are assumed to be assigned to the shortest queue, which is referred to as join the shortest queue discipline. There are many studies on the parallel queueing model with join the shortest queue discipline (see, e.g., [4, 5, 7, 9, 16] and references therein). Even for the case of the M/M/2 queue, the stationary distribution of the parallel queueing model is difficult to obtain in analytically tractable form since there is a correlation between queues. Hence many researchers study its stationary tail asymptotics.

To the best of our knowledge, most of studies on the parallel queueing model with join the shortest queue discipline assume that the service time is exponentially distributed, the arrival process is Poisson, and the number of servers is two. It seems that the assumption on the exponential service time may be unrealistic, and that the arrivals of customers may be correlated. Furthermore, there seems to be no study on the parallel queueing model with generally distributed service times. Therefore we are interested in a queueing model which has more general service time distributions and arrival process.

In [14], we studied a PH/M/2 queue with join the shortest queue discipline, and obtained the tail decay rate of the stationary distribution by using the matrix analytic approach. Although the preceding result is expected for a more general model, it is not trivial to solve this problem as noted in [14].

The main objectives of this paper are to extend the result in [14] to a more general model, and to clarify where difficulty arises when we use similar techniques as in [14]. Specifically, we generalize both the arrival process and the service time distributions to a Markovian arrival process and phase type distributions, respectively, while still keeping two servers. This queueing model is referred to as an MArP/PH/2 queue with join the shortest queue

discipline. We note that our queueing model is sufficiently general because any stationary arrival process and any probability distribution are approximated by the Markovian arrival process and phase type distribution with any desired accuracy, respectively (see, e.g., [1] and [2]). In this paper, we obtain the geometric tail asymptotics of the stationary distribution for our queueing model under a certain condition on the service time distribution.

We note that when the service times are exponentially distributed, that is, when there are no background states on the service times, it is possible to simplify the state description of the queueing model (see, e.g., Section 2 in [14] and Remark 3.1). On the other hand, our queueing model may have some background states on the service times because of the phase type distributions. Hence the state description of our queueing model (see (3.1)) becomes more complicated than the one of [14]. Because of this, it must be more difficult to verify the sufficient conditions for the geometric tail decay as used in [14] (see Sections 4 and 5 for the details). Specifically, to check the last condition of the sufficient conditions, we need an additional condition on the service time distribution (see (4.1)).

A similar queueing model with more than two servers is studied in [13], and the geometric tail asymptotics of the stationary distribution is derived. In [13], the last customer waiting in the longest queue is allowed to move to the shortest one if their difference exceeds a pre-determined threshold value, which is called jockeying. Because of the jockeying, the queueing model in [13] is formulated by a quasi-birth-and-death (QBD, for short) process with *finitely* many background states. The stationary distribution of the QBD process is known to have the matrix geometric form (see, e.g., [6] and [12]). Because of the finitely many background states, it is easy to get the geometric tail asymptotics by computing the maximal eigenvalue of the rate matrix in the matrix geometric form (see Theorem 4.1 of [13]).

Although our queueing model is also formulated by the QBD process, the number of the background states is *infinite* (see (3.1)) since the jockeying is not allowed. To study the tail asymptotics, we consider the convergence radius of the rate matrix because the size of the rate matrix is infinite. However, the convergence radius may not be the tail decay rate of the stationary distribution (see, e.g., [4], [9] and [10]). This is the reason why our queueing model is essentially difficult to analyze compared with the one of [13].

This paper is organized as follows. In Section 2, we introduce some techniques to obtain the geometric tail asymptotics of the stationary distribution. In Section 3, we formally describe our queueing model. In Section 4, we state our main result of this paper, and discuss difficulty when we apply the similar techniques as in [14]. As a result, we have to make an additional assumption on the service time (see (4.1)) to derive the geometric tail asymptotics. We conclude this paper with some remarks on future research in Section 5.

Throughout this paper, we shall use the following notations. For matrix A and vector $\boldsymbol{a}$, denote their (i,j)-th and i-th elements by $[A]_{ij}$ and $[\boldsymbol{a}]_i$, respectively. Let $\Delta_{\boldsymbol{a}}$ be the diagonal matrix whose (i,i)-th entry is the corresponding i-th entry of vector $\boldsymbol{a}$. Transpositions of matrix A and vector $\boldsymbol{a}$ are denoted by A^{t} and $\boldsymbol{a}^{\mathrm{t}}$, respectively. Let $\mathbf{0}$ be the null vector, and let $\mathbf{1}$ be the vector whose entries are all one, where their sizes will be determined by the context in which they appear. We denote the set of all integers by $\mathbb{Z}$, and the sets of all negative and positive integers by $\mathbb{Z}_-$ and $\mathbb{Z}_+$, respectively.

2. Sufficient Conditions for Geometric Tail Decay

We introduce the QBD process with infinitely many background states since our queueing model is described by this process. Furthermore, we state sufficient conditions for the

geometric tail decay of the stationary distribution of the QBD process. These kind of sufficient conditions were initially studied by [16], and were extended to a more general one by [11].

Let $(X(t), J(t))$ be a two dimensional continuous time Markov chain with state space $(\{0\} \times S_0) \cup (\mathbb{Z}_+ \times S)$, where S_0 and S are countable sets. The transition rate matrix of this Markov chain is assumed to have the following tridiagonal structure:

$$Q = \begin{pmatrix} Q_{00} & Q_{01} & & & \\ Q_{10} & Q_0 & Q_{+1} & & \\ & Q_{-1} & Q_0 & Q_{+1} & \\ & & \ddots & \ddots & \ddots \end{pmatrix},$$

where Q_{00}, Q_{01} and Q_{10} are the $|S_0| \times |S_0|$, $|S_0| \times |S|$ and $|S| \times |S_0|$ matrices, respectively, and Q_i is the $|S| \times |S|$ matrix for $i = 0, \pm 1$. The diagonal elements of Q are assumed to be bounded below. Then this Markov chain is referred to as the QBD process with infinitely many background states, where $X(t)$ and $J(t)$ are referred to as level and background processes, respectively. We assume that the QBD process has the stationary distribution $\boldsymbol{\pi} = (\boldsymbol{\pi}_n; n \in \{0\} \cup \mathbb{Z}_+)$, which is partitioned according to the value of the level. From [12], the stationary distribution has the matrix geometric form:

$$\boldsymbol{\pi}_n = \boldsymbol{\pi}_1 R^{n-1}, \quad n \geq 2, \tag{2.1}$$

where R is called the rate matrix (see, e.g., page 8 in [12]), and is given as the minimal nonnegative solution of the following matrix quadratic equation:

$$Q_{+1} + RQ_0 + R^2 Q_{-1} = O. \tag{2.2}$$

In what follows, we state the sufficient conditions for the geometric tail decay of the stationary distribution of the QBD process. To this end, we introduce the following matrix generating function:

$$Q^*(z) = z^{-1}Q_{-1} + Q_0 + zQ_{+1}, \quad z \neq 0.$$

As you will see in the last of this section, for the geometric tail decay, it is sufficient to find positive vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ and a constant $\alpha > 1$ satisfying the following four conditions:

$$\text{(C1) } \boldsymbol{x}Q^*(\alpha) = \mathbf{0}, \quad \text{(C2) } Q^*(\alpha)\boldsymbol{y} = \mathbf{0}, \quad \text{(C3) } \boldsymbol{x}\boldsymbol{y} < \infty, \quad \text{(C4) } \boldsymbol{\pi}_0 Q_{01}\boldsymbol{y} < \infty.$$

We note that

$$\begin{aligned} Q^*(\alpha) &= \alpha^{-1}Q_{-1} + Q_0 + \alpha(-RQ_0 - R^2Q_{-1}) \\ &= (I - \alpha R)Q_0 + (I - \alpha^2 R^2)\alpha^{-1}Q_{-1} \\ &= (I - \alpha R)(Q_0 + \alpha^{-1}Q_{-1} + RQ_{-1}), \end{aligned} \tag{2.3}$$

where the first equality follows from (2.2), and I is the identity matrix whose size is determined in which it appears. From Proposition 6.4.2 in [6], $Q_0 + \alpha^{-1}Q_{-1} + RQ_{-1}$ in (2.3) is rewritten by

$$Q_0 + \alpha^{-1}Q_{-1} + Q_{+1}G, \tag{2.4}$$

where G records the transition probability of the background process when the level process eventually moves down from the starting level. Since α is greater than one, (2.4) is a defective transition rate matrix. From (2.3) and (2.4), we obtain the following result.

Lemma 2.1. The following statements are equivalent.

(i) There exist positive vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ and a constant $\alpha > 1$ satisfying (C1) and (C2).

(ii) The convergence radius of R is given by α^{-1}, and R has the corresponding positive left and right invariant vectors $\boldsymbol{x}$ and $\boldsymbol{r} \equiv -(Q_0 + (\alpha^{-1}I + R)Q_{-1})\boldsymbol{y}$, respectively.

By the following lemma, conditions from (C1) to (C3) indicate the α-positivity of R (see [15] for the α-positivity).

Lemma 2.2. Conditions from (C1) to (C3) imply that $\boldsymbol{xr} < \infty$ and

$$\lim_{n\to\infty} \alpha^n R^n = \frac{\boldsymbol{r}\boldsymbol{x}}{\boldsymbol{x}\boldsymbol{r}}. \tag{2.5}$$

Proof. From Lemma 2.1 and conditions (C1) and (C2), we have

$$\alpha \boldsymbol{x} R = \boldsymbol{x}, \quad \alpha R \boldsymbol{r} = \boldsymbol{r}. \tag{2.6}$$

Let δ be a positive constant such that $\delta I + Q_0 + (\alpha^{-1}I + R)Q_{-1}$ becomes a nonnegative matrix, where the existence of such δ is assured since the diagonal elements of Q_0 is assumed to be bounded below. Then we have

$$\begin{aligned} \boldsymbol{x}\boldsymbol{r} &= \boldsymbol{x}(\delta I - \big(\delta I + Q_0 + (\alpha^{-1}I + R)Q_{-1}\big))\boldsymbol{y} \\ &\leq \delta \boldsymbol{x}\boldsymbol{y}, \end{aligned} \tag{2.7}$$

which is finite by (C3). Then R is shown to be α-positive from (2.6) and the finiteness of (2.7). Hence we obtain (2.5) by Theorem 6.5 in [15]. □

From (4.5) of Theorem 4.1 in [11], the last condition (C4) ensures the following interchange of the limit:

$$\lim_{n\to\infty} \alpha^n \boldsymbol{\pi}_1 R^{n-1} = \alpha \boldsymbol{\pi}_1 (\lim_{n\to\infty} \alpha^{n-1} R^{n-1}). \tag{2.8}$$

From Lemma 2.1, Lemma 2.2 and (2.8), we obtain the geometric tail asymptotics of the stationary distribution for the QBD process with infinitely many background states. The following proposition is a special case of Theorem 4.1 in [11] because the QBD process is an example of the GI/G/1 type Markov chain.

Proposition 2.1. If there exist positive vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ and a constant $\alpha > 1$ satisfying conditions from (C1) to (C4), then we have

$$\lim_{n\to\infty} \alpha^n \boldsymbol{\pi}_n = \frac{\alpha \boldsymbol{\pi}_1 \boldsymbol{r}}{\boldsymbol{x}\boldsymbol{r}} \boldsymbol{x}. \tag{2.9}$$

Remark 2.1. Conditions (C3) and (C4) are automatically satisfied when the number of the background states of the QBD process is finite. For example, the queueing model studied in [13] is formulated by this type of the QBD process because of the jockeying. Hence the queueing model in [13] is essentially easy to study compared with the one in this paper.

Remark 2.2. When the QBD process has infinitely many background states, it seems hard to directly verify (C4) since the condition includes the unknown vector $\boldsymbol{\pi}_0$, whose size is infinite. To this end, we find a rough upper bound of (C4), and show the finiteness of the bound in Section 4.3.

3. MArP/PH/2 Queue with Join the Shortest Queue Discipline

In the rest of this paper, we use same notations used in the preceding section if there is no ambiguity. We are concerned with the following queueing model. There are two servers named servers 1 and 2, where each server has a single waiting line with infinite capacity. The arrival process of customers is driven by a continuous time Markov chain $B_0(t)$ with state space $\mathcal{S}_0 = \{1, 2, \ldots, m_0\}$, where m_0 is a positive integer. We denote the transition rate matrix of this Markov chain by the $m_0 \times m_0$ matrix $C + D$, where C is an ML-matrix and D is an nonnegative and non-null matrix. A customer is assumed to arrive only when $B_0(t)$ changes according to D. This arrival process is referred to as the Markovian arrival process (MArP(C, D), for short). The arriving customer is assumed to join the shortest queue. If he finds no difference between the queue lengths, he joins either of the queues with equal probabilities. This joining rule of arriving customers is referred to as join the shortest queue discipline.

In each queue, customers are served according to first-come first-served discipline. The service time at each server is independently and identically distributed to a phase type distribution with representation PH$(\boldsymbol{\alpha}, T)$, where $\boldsymbol{\alpha}$ is the m dimensional probability vector satisfying $\boldsymbol{\alpha}\mathbf{1} = 1$ and T is a defective transition rate matrix of size m, where m is a positive integer. The phase type distribution is the distribution of absorbing time for a transient Markov chain with the transition rate matrix T, starting from transient states with the initial distribution $\boldsymbol{\alpha}$. When server i $(= 1, 2)$ is not idle at time t, we denote its service phase by $B_i(t) \in \mathcal{S}$ $(\equiv \{1, 2, \ldots, m\})$. Otherwise, that is, when server i is idle, we put $B_i(t) = 0$. This queueing model is referred to as the MArP/PH/2 with join the shortest queue discipline (see also Figure 1), which includes the queueing model studied in [14] as a special case.

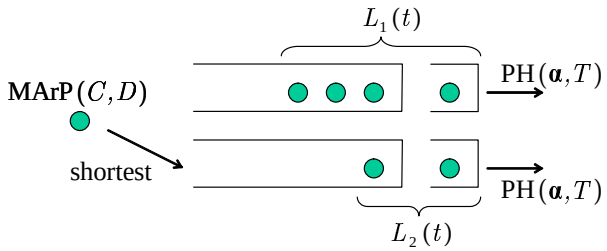

Figure 1: MArP/PH/2 queue with join the shortest queue discipline

We describe our queueing model by the QBD process with infinitely many background states. To this end, we denote the queue length including a customer being served at server i by $L_i(t)$ for $i = 1, 2$. By taking $\min\{L_1(t), L_2(t)\}$ and $(L_2(t) - L_1(t), B_0(t), B_1(t), B_2(t))$ as level and background processes, respectively, it is easy to see that

$$(\min\{L_1(t), L_2(t)\}, (L_2(t) - L_1(t), B_0(t), B_1(t), B_2(t))) \tag{3.1}$$

is the QBD process with infinitely many background states. The level partitioned state space of (3.1) is denoted by $\mathcal{U} = \cup_{n=0}^{\infty} \mathcal{U}_n$, where

$$\mathcal{U}_n = \{n\} \times \mathbb{Z} \times \mathcal{S}_0 \times \mathcal{S} \times \mathcal{S}, \quad n \geq 1,$$

and $\mathcal{U}_0$ is further partitioned according to the difference between queue lengths as follows:

$$\mathcal{U}_0 = \cup_{\ell=-\infty}^{\infty} \mathcal{U}_{0\ell}, \tag{3.2}$$

where $\mathcal{U}_{00} = \{0\} \times \{0\} \times \mathcal{S}_0 \times \{0\} \times \{0\}$, $\mathcal{U}_{0\ell} = \{0\} \times \{\ell\} \times \mathcal{S}_0 \times \{0\} \times \mathcal{S}$ for $\ell \geq 1$ and $\mathcal{U}_{0\ell} = \{0\} \times \{\ell\} \times \mathcal{S}_0 \times \mathcal{S} \times \{0\}$ for $\ell \leq -1$.

Remark 3.1. When the service time has an exponential distribution, the state description (3.1) can be simplified to

$$(\min\{L_1(t), L_2(t)\}, (|L_2(t) - L_1(t)|, B_0(t))),$$

where the absolute value is ensured by the memoryless property of the exponential distribution (see also [14]).

The level partitioned transition rate matrix of (3.1) is given by

$$Q = \begin{pmatrix} Q_{00} & Q_{01} & & & \\ Q_{10} & Q_0 & Q_{+1} & & \\ & Q_{-1} & Q_0 & Q_{+1} & \\ & & \ddots & \ddots & \ddots \end{pmatrix},$$

where each submatrix is given by

$$Q_{00} = \begin{pmatrix} \ddots & \ddots & & & & & \\ & C \oplus T & I_a \otimes \boldsymbol{t}\boldsymbol{\alpha} & & & & \\ & & C \oplus T & I_a \otimes \boldsymbol{t} & & & \\ & & 2^{-1}D \otimes \boldsymbol{\alpha} & C & 2^{-1}D \otimes \boldsymbol{\alpha} & & \\ & & & I_a \otimes \boldsymbol{t} & C \oplus T & & \\ & & & & I_a \otimes \boldsymbol{t}\boldsymbol{\alpha} & C \oplus T & \\ & & & & & \ddots & \ddots \end{pmatrix},$$

$$Q_{01} = \begin{pmatrix} \ddots & & & & & \\ & D \otimes I_s \otimes \boldsymbol{\alpha} & & & & \\ & & D \otimes I_s \otimes \boldsymbol{\alpha} & & & \\ & & O & & & \\ & & D \otimes \boldsymbol{\alpha} \otimes I_s & & & \\ & & & D \otimes \boldsymbol{\alpha} \otimes I_s & & \\ & & & & \ddots & \end{pmatrix},$$

$$Q_{10} = \begin{pmatrix} \ddots & & & & & & \\ & I_a \otimes I_s \otimes \boldsymbol{t} & & & & & \\ & & I_a \otimes I_s \otimes \boldsymbol{t} & O & I_a \otimes \boldsymbol{t} \otimes I_s & & \\ & & & & & I_a \otimes \boldsymbol{t} \otimes I_s & \\ & & & & & & \ddots \end{pmatrix},$$

$$Q_{-1} = \begin{pmatrix} \ddots & & & & & & \\ & I_a \otimes I_s \otimes \boldsymbol{t}\boldsymbol{\alpha} & & & & & \\ & & I_a \otimes I_s \otimes \boldsymbol{t}\boldsymbol{\alpha} & O & I_a \otimes \boldsymbol{t}\boldsymbol{\alpha} \otimes I_s & & \\ & & & & & I_a \otimes \boldsymbol{t}\boldsymbol{\alpha} \otimes I_s & \\ & & & & & & \ddots \end{pmatrix}, \quad (3.3)$$

$$Q_0 = \begin{pmatrix} \ddots & \ddots & & & & & \\ & C\oplus T\oplus T & I_a\otimes \boldsymbol{t}\boldsymbol{\alpha}\otimes I_s & & & & \\ & & C\oplus T\oplus T & I_a\otimes \boldsymbol{t}\boldsymbol{\alpha}\otimes I_s & & & \\ & & 2^{-1}D\otimes I_s\otimes I_s & C\oplus T\oplus T & 2^{-1}D\otimes I_s\otimes I_s & & \\ & & & I_a\otimes I_s\otimes \boldsymbol{t}\boldsymbol{\alpha} & C\oplus T\oplus T & & \\ & & & & I_a\otimes I_s\otimes \boldsymbol{t}\boldsymbol{\alpha} & C\oplus T\oplus T & \\ & & & & & \ddots & \ddots \end{pmatrix}, \quad (3.4)$$

$$Q_{+1} = \begin{pmatrix} \ddots & & & & & \\ & D\otimes I_s\otimes I_s & & & & \\ & & D\otimes I_s\otimes I_s & & & \\ & & O & & & \\ & & D\otimes I_s\otimes I_s & & & \\ & & & D\otimes I_s\otimes I_s & & \\ & & & & \ddots & \end{pmatrix}, \quad (3.5)$$

where I_a (resp. I_s) is the $m_0 \times m_0$ (resp. $m \times m$) identity matrix, and $\otimes$ (resp. $\oplus$) stands for the Kronecker product (resp. sum) (see [3] for the details of the Kronecker operators).

We assume that $C + D$ and $T + \boldsymbol{t}\boldsymbol{\alpha}$ are irreducible, where $\boldsymbol{t} \equiv -T\mathbf{1}$. Let $\boldsymbol{\kappa}_0$ and $\boldsymbol{\kappa}$ be the stationary distributions of $C + D$ and $T + \boldsymbol{t}\boldsymbol{\alpha}$, respectively. Throughout this paper, we assume the following stability condition:

$$\lambda < 2\mu, \quad (3.6)$$

where $\lambda = \boldsymbol{\kappa}_0 D\mathbf{1}$ and $\mu = \boldsymbol{\kappa}\boldsymbol{t}$. The stability is intuitively obvious since the level process of the QBD process has a negative drift under (3.6). This is formally verified through the truncation arguments in Lemma 4.6. We denote the level partitioned stationary distribution of (3.1) by $\boldsymbol{\pi} = (\boldsymbol{\pi}_n; n \geq 0)$. In the next section, we obtain the tail decay rate of $\boldsymbol{\pi}_n$ as n increases.

4. Geometric Tail Asymptotics for the Shortest Queue

In this section, we obtain the geometric tail asymptotics of the stationary distribution for the MArP/PH/2 queue with join the shortest queue discipline. To this end, we verify the sufficient conditions (see conditions from (C1) to (C4) in Section 2) for the geometric tail decay. As will be shown in Sections 4.1 and 4.2, similar techniques as in [14] are applicable to check conditions from (C1) to (C3). In Section 4.3, we verify the last condition (C4) under an additional assumption on the service time (see (4.1)). The difficulty when we try to remove the assumption is also discussed.

We state our main result of this paper, whose proof will be given through the following subsections. By Theorem 4.1 below, the tail probability of the stationary distribution for the shortest queue length geometrically decays with rate r^2 under (4.1), where $r \in (0, 1)$ is the tail decay rate for the corresponding queueing model with a single waiting line, and is determined by (4.3).

Theorem 4.1. Under the following assumption on the service time:

$$\boldsymbol{t} > \mathbf{0}, \quad (4.1)$$

we have

$$\lim_{n\to\infty} r^{-2n}\boldsymbol{\pi}_n = \frac{r^{-2}\boldsymbol{\pi}_1\boldsymbol{r}}{\boldsymbol{x}\boldsymbol{r}}\boldsymbol{x}, \quad (4.2)$$

where $\boldsymbol{r} \equiv -(Q_0 + (r^2 I + R)Q_{-1})\boldsymbol{y}$ and R is the minimal nonnegative solution of

$$Q_{+1} + RQ_0 + R^2 Q_{-1} = O.$$

Remark 4.1. The assumption (4.1) implies that the service completion must occur from any service phase with positive probability. When the service time has an exponential distribution with mean μ^{-1}, i.e., in the case of the MArP/M/2 with join the shortest queue discipline, (4.1) holds since we have $\boldsymbol{t} = \mu > 0$. Hence our result includes Theorem 5.1 of [14] as a special case. However, even for an Erlang distribution with rate μ and shape parameter 2, we have $\boldsymbol{t} = (0, \mu)^t$, which does not satisfy (4.1).

Conjecture 4.1. The assumption (4.1) can be removed possibly using the approaches in [9] and [10], but we have not yet proved this conjecture (see also Section 5).

4.1. Derivation of positive vector satisfying (C2)

To find the positive vector $\boldsymbol{y}$ satisfying (C2), we consider the corresponding queueing model with a single waiting line. This queueing model consists of two servers and a single waiting line with infinite capacity, where customers arrive according to the same Markovian arrival process MArP(C, D), and the service times at each server are independently and identically distributed to the same phase type distribution PH$(\boldsymbol{\alpha}, T)$. This queueing model is referred to as an MArP/PH/2 queue with a single waiting line. We denote the queue length including customers being served by $L(t)$, and the arrival phase by $D_0(t)$. The service phase of server i is denoted by $D_i(t)$ for $i = 1, 2$. By taking $L(t)$ and $(D_0(t), D_1(t), D_2(t))$ as level and background processes, respectively, $(L(t), (D_0(t), D_1(t), D_2(t)))$ is a QBD process with finitely many background states. The level partitioned transition rate matrix of this QBD process is denoted by

$$P = \begin{pmatrix} P_{00} & P_{01} & & & & \\ P_{10} & P_{11} & P_{12} & & & \\ & P_{21} & P_0 & P_{+1} & & \\ & & P_{-1} & P_0 & P_{+1} & \\ & & & \ddots & \ddots & \ddots \end{pmatrix},$$

where each submatrix is given as follows:

$$P_{00} = C,\ P_{01} = (2^{-1}D \otimes \boldsymbol{\alpha}, 2^{-1}D \otimes \boldsymbol{\alpha}),\ P_{10} = \begin{pmatrix} I_a \otimes \boldsymbol{t} \\ I_a \otimes \boldsymbol{t} \end{pmatrix},\ P_{11} = \begin{pmatrix} C \oplus T & O \\ O & C \oplus T \end{pmatrix},$$
$$P_{12} = \begin{pmatrix} D \otimes I_s \otimes \boldsymbol{\alpha} \\ D \otimes \boldsymbol{\alpha} \otimes I_s \end{pmatrix},\ P_{21} = (I_a \otimes I_s \otimes \boldsymbol{t}, I_a \otimes \boldsymbol{t} \otimes I_s),$$
$$P_{-1} = I_a \otimes (\boldsymbol{t}\boldsymbol{\alpha} \oplus \boldsymbol{t}\boldsymbol{\alpha}),\ P_0 = C \oplus T \oplus T,\ P_{+1} = D \otimes I_s \otimes I_s.$$

Since the QBD process under consideration here has a finite number of background states, the tail probability of the marginal distribution for the queue length geometrically decays with rate r under the stability condition (3.6). Then we get the following lemma which plays key role when we construct the positive vector $\boldsymbol{y}$ in (C2).

Lemma 4.1. There exist positive vectors $\boldsymbol{p}$ and $\boldsymbol{u}$ such that

$$\boldsymbol{p}P^*(r^{-1}) = \boldsymbol{0}, \quad P^*(r^{-1})\boldsymbol{u} = \boldsymbol{0}, \tag{4.3}$$

where $P^*(z) = z^{-1}P_{-1} + P_0 + zP_{+1}$ for $z \neq 0$. Furthermore, up to constant multiples, we have

$$\boldsymbol{p} = \boldsymbol{\beta}_0 \otimes \boldsymbol{\beta} \otimes \boldsymbol{\beta}, \quad \boldsymbol{u} = \boldsymbol{g}_0 \otimes \boldsymbol{g} \otimes \boldsymbol{g}, \tag{4.4}$$

where $\boldsymbol{\beta}_0$ and $\boldsymbol{g}_0$ are the Perron Frobeneus left and right eigenvectors of $C + r^{-1}D$, respectively, and $\boldsymbol{\beta}$ and $\boldsymbol{g}$ are the Perron Frobeneus left and right eigenvectors of $T + r\boldsymbol{t}\boldsymbol{\alpha}$, respectively.

Proof. The existence of the positive vectors satisfying (4.3) follows from Lemma 2.1. We denote the Perron Frobeneus eigenvalue of $C + r^{-1}D$ (resp. $T + r\boldsymbol{t}\boldsymbol{\alpha}$) by $s_0 > 0$ (resp. $s < 0$). Noting that $P^*(r^{-1}) = (C + r^{-1}D) \oplus (T + r\boldsymbol{t}\boldsymbol{\alpha}) \oplus (T + r\boldsymbol{t}\boldsymbol{\alpha})$, then we have

$$(\boldsymbol{\beta}_0 \otimes \boldsymbol{\beta} \otimes \boldsymbol{\beta})P^*(r^{-1}) = (s_0 + 2s)(\boldsymbol{\beta}_0 \otimes \boldsymbol{\beta} \otimes \boldsymbol{\beta}), \tag{4.5}$$

$$P^*(r^{-1})(\boldsymbol{g}_0 \otimes \boldsymbol{g} \otimes \boldsymbol{g}) = (s_0 + 2s)(\boldsymbol{g}_0 \otimes \boldsymbol{g} \otimes \boldsymbol{g}). \tag{4.6}$$

By postmultiplying $\boldsymbol{u}$ to (4.5), we have $(s_0 + 2s)(\boldsymbol{\beta}_0 \otimes \boldsymbol{\beta} \otimes \boldsymbol{\beta})\boldsymbol{u} = 0$, which implies that

$$s_0 + 2s = 0 \tag{4.7}$$

since $\boldsymbol{\beta}_0$, $\boldsymbol{\beta}$ and $\boldsymbol{u}$ are positive vectors. From (4.3), (4.5), (4.6) and (4.7), we obtain (4.4). □

From (3.3), (3.4) and (3.5), the matrix generating function $Q^*(r^{-2})$ has the following tridiagonal structure:

$$Q^*(r^{-2}) = \begin{pmatrix} \ddots & \ddots & \ddots & & & & & & \\ & \underline{Q}^*_{-1}(r) & Q^*_0 & \underline{Q}^*_{+1}(r) & & & & & \\ & & \underline{Q}^*_{-1}(r) & Q^*_0 & \underline{Q}^*_{+1}(r) & & & & \\ & & & Q^*_{-1}(r) & Q^*_0 & Q^*_{+1}(r) & & & \\ & & & & \overline{Q}^*_{-1}(r) & Q^*_0 & \overline{Q}^*_{+1}(r) & & \\ & & & & & \overline{Q}^*_{-1}(r) & Q^*_0 & \overline{Q}^*_{+1}(r) & \\ & & & & & & \ddots & \ddots & \ddots \end{pmatrix},$$

where

$$Q^*_0 = C \oplus T \oplus T, \quad Q^*_{+1}(r) = r^2(I_a \otimes \boldsymbol{t}\boldsymbol{\alpha} \otimes I_s) + 2^{-1}(D \otimes I_s \otimes I_s)$$
$$Q^*_{-1}(r) = r^2(I_a \otimes I_s \otimes \boldsymbol{t}\boldsymbol{\alpha}) + 2^{-1}(D \otimes I_s \otimes I_s), \quad \overline{Q}^*_{+1}(r) = r^2(I_a \otimes \boldsymbol{t}\boldsymbol{\alpha} \otimes I_s)$$
$$\overline{Q}^*_{-1}(r) = I_a \otimes I_s \otimes \boldsymbol{t}\boldsymbol{\alpha} + r^{-2}(D \otimes I_s \otimes I_s), \quad \underline{Q}^*_{+1}(r) = I_a \otimes \boldsymbol{t}\boldsymbol{\alpha} \otimes I_s + r^{-2}(D \otimes I_s \otimes I_s),$$
$$\underline{Q}^*_{-1}(r) = r^2(I_a \otimes I_s \otimes \boldsymbol{t}\boldsymbol{\alpha}).$$

Then the positive vector satisfying (C2) is given by the following geometric form.

Lemma 4.2. Let $\boldsymbol{y} = (\boldsymbol{y}_n; n \in \mathbb{Z})$, where $\boldsymbol{y}_n = r^{-|n|}\boldsymbol{u}$ for $n \in \mathbb{Z}$. Then $\boldsymbol{y}$ is the positive vector satisfying (C2).

Proof. Note that the second equation of (4.3) is rewritten by

$$\Big(D \otimes I_s \otimes I_s + r(C \oplus T \oplus T) + r^2(I_a \otimes (\boldsymbol{t}\boldsymbol{\alpha} \oplus \boldsymbol{t}\boldsymbol{\alpha}))\Big)\boldsymbol{u} = \boldsymbol{0}, \tag{4.8}$$

then we have

$$Q^*_{-1}(r)\boldsymbol{y}_{-1} + Q^*_0\boldsymbol{y}_0 + Q^*_{+1}(r)\boldsymbol{y}_1 = r^{-1}\{D \otimes I_s \otimes I_s + r(C \oplus T \oplus T) + r^2(I_a \otimes (\boldsymbol{t}\boldsymbol{\alpha} \oplus \boldsymbol{t}\boldsymbol{\alpha}))\}\boldsymbol{u},$$

which equals zero by (4.8). Similarly, both $\underline{Q}^*_{-1}(r)\boldsymbol{y}_{i-1} + Q^*_0\boldsymbol{y}_i + \underline{Q}^*_{+1}(r)\boldsymbol{y}_{i+1}$ $(i \in \mathbb{Z}_-)$ and $\overline{Q}^*_{-1}(r)\boldsymbol{y}_{i-1} + Q^*_0\boldsymbol{y}_i + \overline{Q}^*_{+1}(r)\boldsymbol{y}_{i+1}$ $(i \in \mathbb{Z}_+)$ are zeros by (4.8), which completes the proof of the lemma. □

4.2. Derivation of positive vector satisfying (C1) and (C3)

To prove the existence of the positive vector $\boldsymbol{x}$ satisfying (C1) and (C3), we consider a continuous time Markov chain with transition rate matrix $\Delta_{\boldsymbol{y}}^{-1}Q^*(r^{-2})\Delta_{\boldsymbol{y}}$, and show that this Markov chain is positive recurrent. We denote this Markov chain by $(Y(t), Z(t))$, where $Y(t)$ and $Z(t)$ take values in $\mathbb{Z}$ and $\mathcal{S}_0 \times \mathcal{S} \times \mathcal{S}$, respectively. From the transition structure of $\Delta_{\boldsymbol{y}}^{-1}Q^*(r^{-2})\Delta_{\boldsymbol{y}}$, it is easy to see that $Z(t)$ is independent of $Y(t)$, and its transition rate matrix is given by

$$r^{-1}\Delta_{\boldsymbol{u}}^{-1}\Big(r^2(I_a \otimes (\boldsymbol{t}\boldsymbol{\alpha} \oplus \boldsymbol{t}\boldsymbol{\alpha})) + r(C \oplus T \oplus T) + D \otimes I_s \otimes I_s\Big)\Delta_{\boldsymbol{u}}. \tag{4.9}$$

We denote the stationary distribution of (4.9) by $\boldsymbol{\zeta}$. Similarly to Theorem 3.1.1 in [12], $(Y(t), Z(t))$ is shown to be positive recurrent by the following lemma because $Z(t)$ takes its values in finite set $\mathcal{S} \times \mathcal{S}$.

Lemma 4.3. The Markov chain $(Y(t), Z(t))$ has drifts to the origin, that is,

$$\boldsymbol{\zeta}(r\Delta_{\boldsymbol{u}}^{-1}\overline{Q}_{-1}^*(r)\Delta_{\boldsymbol{u}})\mathbf{1} > \boldsymbol{\zeta}(r^{-1}\Delta_{\boldsymbol{u}}^{-1}\overline{Q}_{+1}^*(r)\Delta_{\boldsymbol{u}})\mathbf{1} \tag{4.10}$$

and

$$\boldsymbol{\zeta}(r\Delta_{\boldsymbol{u}}^{-1}\underline{Q}_{+1}^*(r)\Delta_{\boldsymbol{u}})\mathbf{1} > \boldsymbol{\zeta}(r^{-1}\Delta_{\boldsymbol{u}}^{-1}\underline{Q}_{-1}^*(r)\Delta_{\boldsymbol{u}})\mathbf{1}. \tag{4.11}$$

Proof. By the first equation of (4.3), the stationary distribution $\boldsymbol{\zeta}$ is given by

$$\boldsymbol{\zeta} = \frac{\boldsymbol{p}\Delta_{\boldsymbol{u}}}{\boldsymbol{p}\boldsymbol{u}}.$$

Then we have

$$\boldsymbol{\zeta}(r\Delta_{\boldsymbol{u}}^{-1}\overline{Q}_{-1}^*(r)\Delta_{\boldsymbol{u}})\mathbf{1} - \boldsymbol{\zeta}(r^{-1}\Delta_{\boldsymbol{u}}^{-1}\overline{Q}_{+1}^*(r)\Delta_{\boldsymbol{u}})\mathbf{1} = \frac{r^{-1}}{\boldsymbol{p}\boldsymbol{u}}(\boldsymbol{\beta}_0 D\boldsymbol{g}_0)(\boldsymbol{\beta}\boldsymbol{g})^2 \tag{4.12}$$

from (4.4). Since the right-hand side of (4.12) is positive, we obtain (4.10). The proof for (4.11) is similar. □

We denote the stationary distribution of $(Y(t), Z(t))$ by $\boldsymbol{\xi}$, and partition it according to the value of $Y(t)$ such that $\boldsymbol{\xi} = (\boldsymbol{\xi}_n; n \in \mathbb{Z})$. Then we have the following lemma.

Lemma 4.4. Let $\boldsymbol{x} = (\boldsymbol{x}_n; n \in \mathbb{Z})$, where $\boldsymbol{x}_n = r^{|n|}\boldsymbol{\xi}_n\Delta_{\boldsymbol{u}}^{-1}$ for $n \in \mathbb{Z}$. Then $\boldsymbol{x}$ satisfies (C1) and (C3).

Proof. Since $\boldsymbol{\xi}$ is the stationary distribution for $(Y(t), Z(t))$, we have

$$\boldsymbol{\xi}\Delta_{\boldsymbol{y}}^{-1}Q^*(r^{-2})\Delta_{\boldsymbol{y}} = \mathbf{0}, \quad \boldsymbol{\xi}\mathbf{1} = 1,$$

which imply that (C1) is satisfied by $\boldsymbol{x} = (r^{|n|}\boldsymbol{\xi}_n\Delta_{\boldsymbol{u}}^{-1}; n \in \mathbb{Z})$. Condition (C3) is readily satisfied since $\boldsymbol{x}\boldsymbol{y} = \boldsymbol{\xi}\mathbf{1} = 1$. □

4.3. Verification of (C4)

It remains to check the last condition (C4). By Lemma 4.2, we have

$$\boldsymbol{\pi}_0 Q_{01}\boldsymbol{y} = \sum_{\ell\geq 0} r^{-\ell}\boldsymbol{\pi}_{0,\ell+1}(D\otimes\boldsymbol{\alpha}\otimes I_s)\boldsymbol{u} + \sum_{\ell\geq 0} r^{-\ell}\boldsymbol{\pi}_{0,-(\ell+1)}(D\otimes I_s\otimes\boldsymbol{\alpha})\boldsymbol{u}, \tag{4.13}$$

where $\boldsymbol{\pi}_0 = (\boldsymbol{\pi}_{0\ell}; \ell\in\mathbb{Z})$ which is partitioned according to (3.2).

For large $|\ell|$, $\boldsymbol{\pi}_{0,\ell+1}$ and $\boldsymbol{\pi}_{0,-(\ell+1)}$ consist of the stationary probabilities of the background process when the queue lengths are unbalanced. These probabilities must be small since the unbalanced situation is a rare event because of the joining rule of customers. Thus the finiteness of (4.13) intuitively seems to be obtained. In the rest of this section, we formally prove the finiteness of (4.13) by introducing a rough upper bound.

Conjecture 4.2. In the case of Poisson arrival and two exponential servers, the tail decay rate of the difference between the queue lengths is obtained by using a stationary equation (see, e.g., [10]). Hence we conjecture that the approach in [10] may be useful to show the finiteness of (4.13), but we have not yet proved this conjecture.

The rough upper bound for (4.13) is given by the following lemma, which is similar to (5.2) in [14] except for a few modifications.

Lemma 4.5. Under the assumption (4.1), there is a positive constant θ such that

$$\begin{aligned}\boldsymbol{\pi}_0 Q_{01}\boldsymbol{y} &\leq \boldsymbol{\pi}_{0,1}(D\otimes\boldsymbol{\alpha}\otimes I_s)\boldsymbol{u} + \boldsymbol{\pi}_{0,-1}(D\otimes I_s\otimes\boldsymbol{\alpha})\boldsymbol{u}\\ &\quad +2\theta\sum_{n\geq 1} r^{-n}\boldsymbol{\pi}_{n0}(D\otimes I_s\otimes I_s)\mathbf{1}.\end{aligned} \tag{4.14}$$

Proof. For $n\geq 1$, let $\mathcal{M}_n$ be a subset of the state space $\mathcal{U}$ for (3.1) such that it consists of the states in which the longer queue is not greater than n (see Figure 2). We consider the flow balance equation between $\mathcal{M}_n$ and $\mathcal{M}_n^c$ for $n\geq 1$. Note that the set of states through which the process could leave $\mathcal{M}_n$ for $\mathcal{M}_n^c$ is $\{(n,0,i,j,k); i\in\mathcal{S}_0, j,k\in\mathcal{S}\}$ due to transition rate $D\otimes I_s\otimes I_s$. On the other hand, the sets of states through which the process could leave $\mathcal{M}_n^c$ for $\mathcal{M}_n$ are $\{(0,n+1,i,0,j); i\in\mathcal{S}_0, j\in\mathcal{S}\}$, $\{(0,-(n+1),i,j,0); i\in\mathcal{S}_0, j\in\mathcal{S}\}$, $\cup_{\ell=1}^{n}\{(\ell,n+1-\ell,i,j,k); i\in\mathcal{S}_0, j,k\in\mathcal{S}\}$ and $\cup_{\ell=-n}^{-1}\{(\ell,-(n+1)-\ell,i,j,k); i\in\mathcal{S}_0, j,k\in\mathcal{S}\}$ due to transition rates $I_a\otimes\boldsymbol{t}\boldsymbol{\alpha}$, $I_a\otimes\boldsymbol{t}\boldsymbol{\alpha}$, $I_a\otimes I_s\otimes\boldsymbol{t}\boldsymbol{\alpha}$ and $I_a\otimes\boldsymbol{t}\boldsymbol{\alpha}\otimes I_s$, respectively. Therefore we obtain the following flow balance equation

$$\begin{aligned}\boldsymbol{\pi}_{n0}(D\otimes I_s\otimes I_s)\mathbf{1} &= \boldsymbol{\pi}_{0,n+1}(I_a\otimes\boldsymbol{t}\boldsymbol{\alpha})\mathbf{1} + \sum_{\ell=1}^{n}\boldsymbol{\pi}_{\ell,n+1-\ell}(I_a\otimes I_s\otimes\boldsymbol{t}\boldsymbol{\alpha})\mathbf{1}\\ &\quad +\boldsymbol{\pi}_{0,-(n+1)}(I_a\otimes\boldsymbol{t}\boldsymbol{\alpha})\mathbf{1} + \sum_{\ell=1}^{n}\boldsymbol{\pi}_{\ell,-(n+1-\ell)}(I_a\otimes\boldsymbol{t}\boldsymbol{\alpha}\otimes I_s)\mathbf{1},\end{aligned}$$

which implies that

$$\boldsymbol{\pi}_{n0}(D\otimes I_s\otimes I_s)\mathbf{1}\geq\boldsymbol{\pi}_{0,n+1}(\mathbf{1}\otimes\boldsymbol{t}),\quad \boldsymbol{\pi}_{n0}(D\otimes I_s\otimes I_s)\mathbf{1}\geq\boldsymbol{\pi}_{0,-(n+1)}(\mathbf{1}\otimes\boldsymbol{t}) \tag{4.15}$$

for $n\geq 1$. By (4.1), there is a constant $\theta>0$ such that

$$\theta(\mathbf{1}\otimes\boldsymbol{t})\geq(D\otimes\boldsymbol{\alpha}\otimes I_s)\boldsymbol{u},\quad \theta(\mathbf{1}\otimes\boldsymbol{t})\geq(D\otimes I_s\otimes\boldsymbol{\alpha})\boldsymbol{u}. \tag{4.16}$$

From (4.15) and (4.16), we have

$$\theta\boldsymbol{\pi}_{n0}(D\otimes I_s\otimes I_s)\mathbf{1} \geq \boldsymbol{\pi}_{0,n+1}(D\otimes\boldsymbol{\alpha}\otimes I_s)\boldsymbol{u}, \tag{4.17}$$

$$\theta\boldsymbol{\pi}_{n0}(D\otimes I_s\otimes I_s)\mathbf{1} \geq \boldsymbol{\pi}_{0,-(n+1)}(D\otimes I_s\otimes\boldsymbol{\alpha})\boldsymbol{u} \tag{4.18}$$

for $n\geq 1$. Then we obtain (4.14) from (4.13), (4.17) and (4.18). □

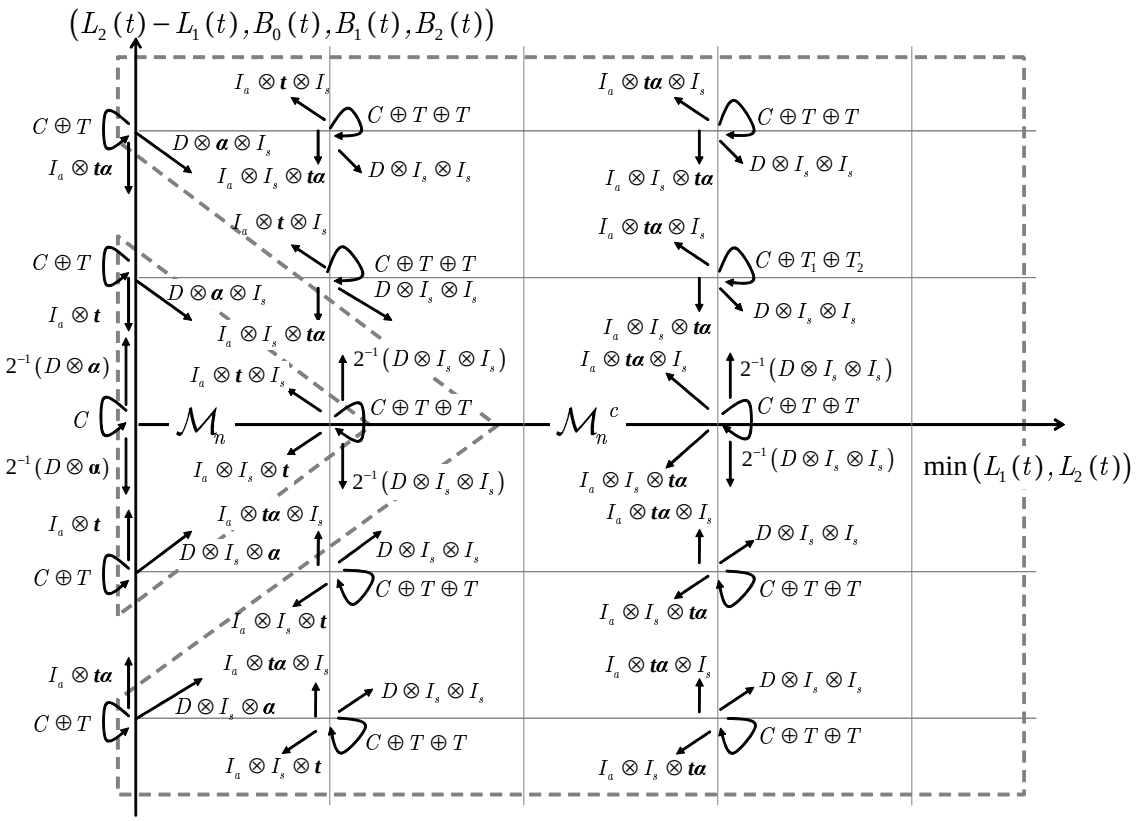

Figure 2: MArP/PH/2 with shortest queue discipline

By the following lemma, $\boldsymbol{\pi}_{n0}$ is shown to decay faster than r^n as n increases. Then condition (C4) is verified by Lemma 4.5, which completes the proof of Theorem 4.1.

Lemma 4.6. For any small $\epsilon > 0$, we have

$$\limsup_{n\to\infty} r^{(-2+\epsilon)n}\boldsymbol{\pi}_{n0} = \mathbf{0}.$$

This lemma is proved through the following three steps which are similar to [14] except for some technical modifications.

Step 1 For fixed $M \geq 3$, we modify the QBD process (3.1) such that its state space $\mathcal{U}$ is truncated by removing the state transitions from (n, M, i, j, k) to $(n-1, M+1, i, j, k)$ for $n \geq 1$, $i \in \mathcal{S}_0$ and $j, k \in \mathcal{S}$. This modification is equivalent to that the service of the shorter queue is stopped when the difference between the queue lengths attains M in the original model. For this truncated process, we assume that state (n, ℓ, i, j, k) is a member of new level $n + \ell$. This truncated model is denoted by

$$(\max\{L_1^{(M)}(t), L_2^{(M)}(t)\}, (L_2^{(M)}(t) - L_1^{(M)}(t), B_0^{(M)}(t), B_1^{(M)}(t), B_2^{(M)}(t))), \tag{4.19}$$

which is a QBD process with finitely many background states. Let $Q_i^{(M)}$ be the transition rate matrix of the background process when the level increases by i $(= 0, \pm 1)$ provided that the level is greater than $M + 1$. Then these transition rate matrices are given by the

following $((2M+1)m_0m^2) \times ((2M+1)m_0m^2)$ matrices

$$Q_{-1}^{(M)} = \begin{pmatrix} O & I_a \otimes \boldsymbol{t}\boldsymbol{\alpha} \otimes I_s & & & & & \\ & \ddots & & & & & \\ & & I_a \otimes \boldsymbol{t}\boldsymbol{\alpha} \otimes I_s & & & & \\ & & & O & & & \\ & & & I_a \otimes I_s \otimes \boldsymbol{t}\boldsymbol{\alpha} & & & \\ & & & & \ddots & & \\ & & & & & I_a \otimes I_s \otimes \boldsymbol{t}\boldsymbol{\alpha} & O \end{pmatrix},$$

$$Q_0^{(M)} = \begin{pmatrix} \underline{A}_0 & D \otimes I_s \otimes I_s & & & & & & & & & \\ I_a \otimes I_s \otimes \boldsymbol{t}\boldsymbol{\alpha} & A_0 & D \otimes I_s \otimes I_s & & & & & & & & \\ & \ddots & \ddots & \ddots & & & & & & & \\ & & I_a \otimes I_s \otimes \boldsymbol{t}\boldsymbol{\alpha} & A_0 & D \otimes I_s \otimes I_s & & & & & & \\ & & & I_a \otimes I_s \otimes \boldsymbol{t}\boldsymbol{\alpha} & A_0 & I_a \otimes \boldsymbol{t}\boldsymbol{\alpha} \otimes I_s & & & & & \\ & & & & D \otimes I_s \otimes I_s & A_0 & I_a \otimes \boldsymbol{t}\boldsymbol{\alpha} \otimes I_s & & & \\ & & & & & \ddots & \ddots & \ddots & & \\ & & & & & & D \otimes I_s \otimes I_s & A_0 & I_a \otimes \boldsymbol{t}\boldsymbol{\alpha} \otimes I_s & \\ & & & & & & & D \otimes I_s \otimes I_s & \overline{A}_0 \end{pmatrix},$$

where $\underline{A}_0 = C \oplus T \oplus (T + \boldsymbol{t}\boldsymbol{\alpha})$, $A_0 = C \oplus T \oplus T$, $\overline{A}_0 = C \oplus (T + \boldsymbol{t}\boldsymbol{\alpha}) \oplus T$, and

$$Q_{+1}^{(M)} = \begin{pmatrix} O & & & & & & \\ & \ddots & & & & & \\ & & O & & & & \\ & & 2^{-1}(D \otimes I_s \otimes I_s) & O & 2^{-1}(D \otimes I_s \otimes I_s) & & \\ & & & & O & & \\ & & & & & \ddots & \\ & & & & & & O \end{pmatrix}.$$

Step 2 For sufficiently large M, we have

$$\tilde{\boldsymbol{p}}^{(M)} Q_{+1}^{(M)} \mathbf{1} - \tilde{\boldsymbol{p}}^{(M)} Q_{-1}^{(M)} \mathbf{1} < 0, \tag{4.20}$$

where $\tilde{\boldsymbol{p}}^{(M)}$ is the stationary distribution for $Q_{-1}^{(M)} + Q_0^{(M)} + Q_{+1}^{(M)}$ (see Appendix A for the proof of (4.20)). This implies that the truncated model is positive recurrent for sufficiently large M by Theorem 3.1.1 in [12]. Since (4.19) is the QBD process with finitely many background states, the tail decay rate of the stationary distribution for $\max(L_1^{(M)}(t), L_2^{(M)}(t))$ decays geometrically with a rate $r_M \in (0,1)$ (see also Remark 2.1). From Lemma 2.1, the decay rate r_M is determined by the following equations:

$$\boldsymbol{p}^{(M)} K^{(M)}(r_M^{-1}) = \mathbf{0}, \qquad K^{(M)}(r_M^{-1}) \boldsymbol{q}^{(M)} = \mathbf{0}, \tag{4.21}$$

where $\boldsymbol{p}^{(M)}$ and $\boldsymbol{q}^{(M)}$ are positive vectors and $K^{(M)}(z) \equiv z^{-1} Q_{-1}^{(M)} + Q_0^{(M)} + z Q_{+1}^{(M)}$ for $z \neq 0$.
Step 3 Since the difference between $\max(L_1^{(M)}(t), L_2^{(M)}(t))$ and $\min(L_1^{(M)}(t), L_2^{(M)}(t))$ is bounded by M, r_M also is the tail decay rate of the shortest queue for the truncated model. As the truncation level M increases, the system becomes more stable. Hence r_M monotonically decreases as $M \to \infty$, and the limit is denoted by $r_\infty \equiv \lim_{M\to\infty} r_M$. Let $\underline{\boldsymbol{p}} = (\underline{\boldsymbol{p}}_\ell; \ell \in \mathbb{Z})$ and $\underline{\boldsymbol{q}} = (\underline{\boldsymbol{q}}_\ell; \ell \in \mathbb{Z})$, where

$$\underline{\boldsymbol{p}}_\ell = \liminf_{M\to\infty} \boldsymbol{p}_\ell^{(M)}, \qquad \underline{\boldsymbol{q}}_\ell = \liminf_{M\to\infty} \boldsymbol{q}_\ell^{(M)}$$

for $\ell \in \mathbb{Z}$. Similarly to [14], by taking $\liminf_{M\to\infty}$ of (4.21), it is shown that these vectors are positive and satisfy

$$\underline{\boldsymbol{p}} K^{(\infty)}(r_\infty^{-1}) = \mathbf{0}, \quad K^{(\infty)}(r_\infty^{-1})\underline{\boldsymbol{q}} = \mathbf{0}, \quad \underline{\boldsymbol{p}}\ \underline{\boldsymbol{q}} < \infty, \tag{4.22}$$

where $K^{(\infty)}(z) \equiv z^{-1}Q_{-1}^{(\infty)} + Q_0^{(\infty)} + zQ_{+1}^{(\infty)}$ for $z \neq 0$ and $Q_i^{(\infty)}$ $(i = 0, \pm 1)$ is obtained from $Q_i^{(M)}$ by letting $M \to \infty$. Furthermore, we can show the existence of the positive vectors $\boldsymbol{p}^{(\infty)}$ and $\boldsymbol{q}^{(\infty)}$ such that

$$\boldsymbol{p}^{(\infty)} K^{(\infty)}(r^{-2}) = \mathbf{0}, \quad K^{(\infty)}(r^{-2})\boldsymbol{q}^{(\infty)} = \mathbf{0}, \quad \boldsymbol{p}^{(\infty)}\boldsymbol{q}^{(\infty)} < \infty \tag{4.23}$$

(see Appendix B for the proofs of (4.22) and (4.23)). From (4.22), (4.23) and Remark 3.2 in [14], r_M converges to r^2 as M increases, which completes the proof of Lemma 4.6.

5. Conclusion and Future Work

We studied the geometric tail asymptotics of the stationary distribution for the MArP/PH/2 with join the shortest queue discipline. To this end, we formulated our queueing model by the QBD process with infinitely many background states, and showed that the sufficient conditions for the geometric tail decay were verified under the additional condition on the service time distribution.

It may be interesting to consider whether the present approach can be applied to the case of many heterogeneous servers. Similarly to [13], the positive vectors in (C1) and (C2) will be obtained for more than two servers. However, the last two conditions (C3) and (C4) are not easy to verify even for the case of two servers by the following reasons.

Firstly, condition (C3) is not trivial to verify even for the case of two heterogeneous servers. Suppose that there exist two servers, and that the service time distribution of server i has a representation $\mathrm{PH}(\boldsymbol{\alpha}_i, T_i)$ for $i = 1, 2$. Similarly to Lemma 4.2, the existence of the positive vector $\boldsymbol{y}$ satisfying (C2) is readily shown. Then there exists the positive vector $\boldsymbol{x}$ satisfying (C1) by the Perron Frobeneus theorem. To see that these positive vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ satisfy (C3), we must show that (4.12) is positive, which is required to show that the Markov chain in Lemma 4.3 is positive recurrent. Let $\boldsymbol{\beta}_i$ (resp. $\boldsymbol{g}_i$) be the Perron Frobeneus left (resp. right) eigenvector of $T_i + r\boldsymbol{t}_i\boldsymbol{\alpha}_i$, where $\boldsymbol{t}_i = -T_i\mathbf{1}$, then the right-hand side of (4.12) is rewritten by

$$\frac{r(\boldsymbol{\beta}_0\boldsymbol{g}_0)\left\{(\boldsymbol{\beta}_1\boldsymbol{g}_1)(\boldsymbol{\beta}_2\boldsymbol{t}_2)(\boldsymbol{\alpha}_2\boldsymbol{g}_2) - (\boldsymbol{\beta}_1\boldsymbol{t}_1)(\boldsymbol{\beta}_2\boldsymbol{g}_2)(\boldsymbol{\alpha}_1\boldsymbol{g}_1)\right\} + r^{-1}(\boldsymbol{\beta}_0 D\boldsymbol{g}_0)(\boldsymbol{\beta}_1\boldsymbol{g}_1)(\boldsymbol{\beta}_2\boldsymbol{g}_2)}{(\boldsymbol{\beta}_0\boldsymbol{g}_0)(\boldsymbol{\beta}_1\boldsymbol{g}_1)(\boldsymbol{\beta}_2\boldsymbol{g}_2)}. \tag{5.1}$$

It seems very difficult to show the positivity of (5.1).

Secondly, the verification of the last condition (C4) is difficult because of the unknown vector $\boldsymbol{\pi}_0$, i.e., the stationary probability vector at level 0 of the QBD process. Hence we showed the finiteness of (C4) by using the rough upper bound as in [14] (see (4.14)), which required the additional condition (4.1) on the service time.

As noted in Conjecture 4.1, the finiteness of (C4) may be directly verified by using the similar techniques as in [9] and [10], in which a two dimensional random walk is studied. To this end, we need to extend the results in [9] and [10] to a multi-dimensional random walk with some background states because of many servers and the background states on both the arrival and service processes. We leave these problems for future work.

Acknowledgment
The author is greatly thankful to Professor Masakiyo Miyazawa at Tokyo University of Science for invaluable discussions. The author also would like to thank the referees for their careful review and the valuable comments, which improved this paper.

A. Proof of (4.20)

To prove (4.20), we consider a two dimensional continuous time Markov chain $(\tilde{Y}(t), \tilde{Z}(t)) \in \mathbb{Z} \times (\mathcal{S}_0 \times \mathcal{S} \times \mathcal{S})$ with transition rate matrix $Q_{-1}^{(\infty)} + Q_0^{(\infty)} + Q_{+1}^{(\infty)}$, where $Q_i^{(\infty)}$ $(i = 0, \pm 1)$ is obtained from $Q_i^{(M)}$ by letting $M \to \infty$. We note that $Q_{-1}^{(\infty)} + Q_0^{(\infty)} + Q_{+1}^{(\infty)}$ has the following tridiagonal structure:

$$\begin{pmatrix}
\ddots & \ddots & \ddots & & & & \\
 & I \otimes I \otimes \boldsymbol{t\alpha} & C \oplus T \oplus T & D \otimes I \otimes I + I \otimes \boldsymbol{t\alpha} \otimes I & & & \\
 & & 2^{-1}(D \otimes I \otimes I) + I \otimes I \otimes \boldsymbol{t\alpha} & C \oplus T \oplus T & 2^{-1}(D \otimes I \otimes I) + I \otimes \boldsymbol{t\alpha} \otimes I & & \\
 & & & D \otimes I \otimes I + I \otimes I \otimes \boldsymbol{t\alpha} & C \oplus T \oplus T & I \otimes \boldsymbol{t\alpha} \otimes I & \\
 & & & & \ddots & \ddots & \ddots
\end{pmatrix},$$

where the suffixes of the identity matrices are omitted. Then it is easy to see that the stationary distribution of $\tilde{Z}(t)$ is given by $\boldsymbol{\kappa}_0 \otimes \boldsymbol{\kappa} \otimes \boldsymbol{\kappa}$, and $\tilde{Y}(t)$ has positive drifts to the origin, i.e.,

$$\begin{aligned}
&(\boldsymbol{\kappa}_0 \otimes \boldsymbol{\kappa} \otimes \boldsymbol{\kappa})(D \otimes I_s \otimes I_s + I_a \otimes I_s \otimes \boldsymbol{t\alpha})\mathbf{1} - (\boldsymbol{\kappa}_0 \otimes \boldsymbol{\kappa} \otimes \boldsymbol{\kappa})(I_a \otimes \boldsymbol{t\alpha} \otimes I_s)\mathbf{1} = \boldsymbol{\kappa}_0 D\mathbf{1} > 0, \\
&(\boldsymbol{\kappa}_0 \otimes \boldsymbol{\kappa} \otimes \boldsymbol{\kappa})(D \otimes I_s \otimes I_s + I_a \otimes \boldsymbol{t\alpha} \otimes I_s)\mathbf{1} - (\boldsymbol{\kappa}_0 \otimes \boldsymbol{\kappa} \otimes \boldsymbol{\kappa})(I_a \otimes I_s \otimes \boldsymbol{t\alpha})\mathbf{1} = \boldsymbol{\kappa}_0 D\mathbf{1} > 0.
\end{aligned}$$

Hence $Q_{-1}^{(\infty)} + Q_0^{(\infty)} + Q_{+1}^{(\infty)}$ has the stationary distribution $\tilde{\boldsymbol{p}}^{(\infty)}$, which is partitioned according to the value of $\tilde{Y}(t)$ as $\tilde{\boldsymbol{p}}^{(\infty)} = (\tilde{\boldsymbol{p}}_\ell^{(\infty)}; \ell \in \mathbb{Z})$. Then we obtain the following result, which is similar to Lemma 5.4 in [14] except for some technical modifications.

Lemma A.1. Under the stability condition (3.6), we have

$$\tilde{\boldsymbol{p}}^{(\infty)} Q_{+1}^{(\infty)} \mathbf{1} - \tilde{\boldsymbol{p}}^{(\infty)} Q_{-1}^{(\infty)} \mathbf{1} = \frac{1}{2}\lambda - \mu < 0.$$

Proof. By the tridiagonal structure of $Q_{-1}^{(\infty)} + Q_0^{(\infty)} + Q_{+1}^{(\infty)}$, the stationary distribution $(\tilde{\boldsymbol{p}}_\ell^{(\infty)}; \ell \in \mathbb{Z})$ has the following forms:

$$\tilde{\boldsymbol{p}}_{-n}^{(\infty)} = \tilde{\boldsymbol{p}}_{-1}^{(\infty)} \underline{R}^{n-1}, \quad n \geq 1, \tag{A.1}$$

$$\tilde{\boldsymbol{p}}_{n}^{(\infty)} = \tilde{\boldsymbol{p}}_{1}^{(\infty)} \overline{R}^{n-1}, \quad n \geq 1, \tag{A.2}$$

where $\tilde{\boldsymbol{p}}_\ell^{(\infty)}$ $(\ell = 0, \pm 1, \pm 2)$ satisfies

$$\begin{aligned}
&\tilde{\boldsymbol{p}}_{-2}^{(\infty)}(D \otimes I_s \otimes I_s + I_a \otimes \boldsymbol{t\alpha} \otimes I_s) + \tilde{\boldsymbol{p}}_{-1}^{(\infty)}(C \oplus T \oplus T) \\
&\quad + \tilde{\boldsymbol{p}}_0^{(\infty)}(2^{-1}(D \otimes I_s \otimes I_s) + I_a \otimes I_s \otimes \boldsymbol{t\alpha}) = \mathbf{0}, \\
&\tilde{\boldsymbol{p}}_{-1}^{(\infty)}(D \otimes I_s \otimes I_s + I_a \otimes \boldsymbol{t\alpha} \otimes I_s) + \tilde{\boldsymbol{p}}_0^{(\infty)}(C \oplus T \oplus T) \\
&\quad + \tilde{\boldsymbol{p}}_1^{(\infty)}(D \otimes I_s \otimes I_s + I_a \otimes I_s \otimes \boldsymbol{t\alpha}) = \mathbf{0}, \\
&\tilde{\boldsymbol{p}}_0^{(\infty)}(2^{-1}(D \otimes I_s \otimes I_s) + I_a \otimes \boldsymbol{t\alpha} \otimes I_s) + \tilde{\boldsymbol{p}}_1^{(\infty)}(C \oplus T \oplus T) \\
&\quad + \tilde{\boldsymbol{p}}_2^{(\infty)}(D \otimes I_s \otimes I_s + I_a \otimes I_s \otimes \boldsymbol{t\alpha}) = \mathbf{0},
\end{aligned} \tag{A.3}$$

and $\underline{R}$ and $\overline{R}$ are the minimal nonnegative solutions of

$$I_a \otimes I_s \otimes \boldsymbol{t}\boldsymbol{\alpha} + \underline{R}(C \oplus T \oplus T) + \underline{R}^2(D \otimes I_s \otimes I_s + I_a \otimes \boldsymbol{t}\boldsymbol{\alpha} \otimes I_s) = O, \quad (A.4)$$

$$I_a \otimes \boldsymbol{t}\boldsymbol{\alpha} \otimes I_s + \overline{R}(C \oplus T \oplus T) + \overline{R}^2(D \otimes I_s \otimes I_s + I_a \otimes I_s \otimes \boldsymbol{t}\boldsymbol{\alpha}) = O, \quad (A.5)$$

respectively. Note that

$$\boldsymbol{\kappa}_0 \otimes \boldsymbol{\kappa} \otimes \boldsymbol{\kappa} = \tilde{\boldsymbol{p}}_{-1}^{(\infty)}(I - \underline{R})^{-1} + \tilde{\boldsymbol{p}}_0^{(\infty)} + \tilde{\boldsymbol{p}}_1^{(\infty)}(I - \overline{R})^{-1}, \quad (A.6)$$

then we have

$$\begin{aligned}
&\tilde{\boldsymbol{p}}^{(\infty)}Q_{+1}^{(\infty)}\mathbf{1} - \tilde{\boldsymbol{p}}^{(\infty)}Q_{-1}^{(\infty)}\mathbf{1} \\
&= \tilde{\boldsymbol{p}}_1^{(\infty)}(I - \overline{R})^{-1}(I_a \otimes I_s \otimes \boldsymbol{t}\boldsymbol{\alpha})\mathbf{1} + \tilde{\boldsymbol{p}}_{-1}^{(\infty)}(I - \underline{R})^{-1}(I_a \otimes \boldsymbol{t}\boldsymbol{\alpha} \otimes I_s)\mathbf{1} - \tilde{\boldsymbol{p}}_0^{(\infty)}(D \otimes I_s \otimes I_s)\mathbf{1} \\
&= 2\boldsymbol{\kappa}\boldsymbol{t} + \tilde{\boldsymbol{p}}_1^{(\infty)}(I - \overline{R})^{-1}(I_a \otimes T \otimes I_s)\mathbf{1} + \tilde{\boldsymbol{p}}_{-1}^{(\infty)}(I - \underline{R})^{-1}(I_a \otimes I_s \otimes T)\mathbf{1} \\
&\quad + \tilde{\boldsymbol{p}}_0^{(\infty)}(C \oplus T \oplus T)\mathbf{1} \\
&= 2\boldsymbol{\kappa}\boldsymbol{t} + \tilde{\boldsymbol{p}}_{-1}^{(\infty)}\{(I - \underline{R})^{-1}(I_a \otimes I_s \otimes T) - (D \otimes I_s \otimes I_s + I_a \otimes \boldsymbol{t}\boldsymbol{\alpha} \otimes I_s)\}\mathbf{1} \\
&\quad + \tilde{\boldsymbol{p}}_1^{(\infty)}\{(I - \overline{R})^{-1}(I_a \otimes T \otimes I_s) - (D \otimes I_s \otimes I_s + I_a \otimes I_s \otimes \boldsymbol{t}\boldsymbol{\alpha})\}\mathbf{1}, \quad (A.7)
\end{aligned}$$

where the second equality follows from (A.6), and the last one is obtained by postmultiplying $\mathbf{1}$ to (A.3). Because of the invertibilities of $I - \underline{R}$ and $I - \overline{R}$, we have

$$(I_a \otimes I_s \otimes T)\mathbf{1} = \underline{R}(C \otimes I_s \otimes I_s + I_a \otimes T \otimes I_s)\mathbf{1}, \quad (A.8)$$

$$(I_a \otimes T \otimes I_s)\mathbf{1} = \overline{R}(C \otimes I_s \otimes I_s + I_a \otimes I_s \otimes T)\mathbf{1} \quad (A.9)$$

by postmultiplying $\mathbf{1}$ to (A.4) and (A.5), respectively. From (A.7), (A.8), (A.9) and $\mu = \boldsymbol{\kappa}\boldsymbol{t}$, we obtain

$$\begin{aligned}
\tilde{\boldsymbol{p}}^{(\infty)}Q_{+1}^{(\infty)}\mathbf{1} - \tilde{\boldsymbol{p}}^{(\infty)}Q_{-1}^{(\infty)}\mathbf{1} &= 2\mu - \tilde{\boldsymbol{p}}_{-1}^{(\infty)}(I - \underline{R})^{-1}(D\mathbf{1} \otimes \mathbf{1} \otimes \mathbf{1} + \mathbf{1} \otimes \boldsymbol{t} \otimes \mathbf{1}) \\
&\quad - \tilde{\boldsymbol{p}}_1^{(\infty)}(I - \overline{R})^{-1}(D\mathbf{1} \otimes \mathbf{1} \otimes \mathbf{1} + \mathbf{1} \otimes \mathbf{1} \otimes \boldsymbol{t}). \quad (A.10)
\end{aligned}$$

On the other hand, from (A.1), (A.2) and (A.7), we have

$$\tilde{\boldsymbol{p}}^{(\infty)}Q_{+1}^{(\infty)}\mathbf{1} = \tilde{\boldsymbol{p}}_{-1}^{(\infty)}(I - \underline{R})^{-1}(\mathbf{1} \otimes \boldsymbol{t} \otimes \mathbf{1}) + \tilde{\boldsymbol{p}}_1^{(\infty)}(I - \overline{R})^{-1}(\mathbf{1} \otimes \mathbf{1} \otimes \boldsymbol{t}). \quad (A.11)$$

From (A.6) and $\lambda = \boldsymbol{\kappa}_0 D\mathbf{1}$, we have

$$\tilde{\boldsymbol{p}}^{(\infty)}Q_{-1}^{(\infty)}\mathbf{1} = \lambda - \tilde{\boldsymbol{p}}_{-1}^{(\infty)}(I - \underline{R})^{-1}(D\mathbf{1} \otimes \mathbf{1} \otimes \mathbf{1}) - \tilde{\boldsymbol{p}}_1^{(\infty)}(I - \overline{R})^{-1}(D\mathbf{1} \otimes \mathbf{1} \otimes \mathbf{1}). \quad (A.12)$$

Combining (A.11) and (A.12) yields

$$\begin{aligned}
\tilde{\boldsymbol{p}}^{(\infty)}Q_{-1}^{(\infty)}\mathbf{1} - \tilde{\boldsymbol{p}}^{(\infty)}Q_{+1}^{(\infty)}\mathbf{1} &= \lambda - \tilde{\boldsymbol{p}}_{-1}^{(\infty)}(I - \underline{R})^{-1}(D\mathbf{1} \otimes \mathbf{1} \otimes \mathbf{1} + \mathbf{1} \otimes \boldsymbol{t} \otimes \mathbf{1}) \\
&\quad - \tilde{\boldsymbol{p}}_1^{(\infty)}(I - \overline{R})^{-1}(D\mathbf{1} \otimes \mathbf{1} \otimes \mathbf{1} + \mathbf{1} \otimes \mathbf{1} \otimes \boldsymbol{t}). \quad (A.13)
\end{aligned}$$

From (A.10) and (A.13), we finally obtain

$$\tilde{\boldsymbol{p}}^{(\infty)}Q_{-1}^{(\infty)}\mathbf{1} - \tilde{\boldsymbol{p}}^{(\infty)}Q_{+1}^{(\infty)}\mathbf{1} = \frac{1}{2}(2\mu - \lambda),$$

which is positive by (3.6). This completes the proof of the lemma. □

We are now ready to prove the stability condition (4.20) for the truncated model. Partition $\tilde{\boldsymbol{p}}^{(M)}$ into blocks as $\tilde{\boldsymbol{p}}^{(M)} = (\tilde{\boldsymbol{p}}_n^{(M)}; -M \le n \le M)$, then we have

$$\begin{aligned}\tilde{\boldsymbol{p}}^{(M)}Q_{+1}^{(M)}\mathbf{1} - \tilde{\boldsymbol{p}}^{(M)}Q_{-1}^{(M)}\mathbf{1} &= \sum_{n=1}^{M}\tilde{\boldsymbol{p}}_n^{(M)}(I_a\otimes I_s\otimes \boldsymbol{t}\boldsymbol{\alpha})\mathbf{1} + \sum_{n=-M}^{-1}\tilde{\boldsymbol{p}}_n^{(M)}(I_a\otimes \boldsymbol{t}\boldsymbol{\alpha}\otimes I_s)\mathbf{1} \\ &\quad -\tilde{\boldsymbol{p}}_0^{(M)}(D\otimes I_s\otimes I_s)\mathbf{1}. \qquad \text{(A.14)}\end{aligned}$$

Applying Theorem 3.4 in [8] to $Q_{-1}^{(M)} + Q_0^{(M)} + Q_{+1}^{(M)}$ and $Q_{-1}^{(\infty)} + Q_0^{(\infty)} + Q_{+1}^{(\infty)}$, we have

$$\lim_{M\to\infty}\sum_{n=-M}^{-1}\tilde{\boldsymbol{p}}_n^{(M)} = \tilde{\boldsymbol{p}}_{-1}^{(\infty)}(I-\underline{R})^{-1},\ \lim_{M\to\infty}\sum_{n=1}^{M}\tilde{\boldsymbol{p}}_n^{(M)} = \tilde{\boldsymbol{p}}_1^{(\infty)}(I-\overline{R})^{-1},\ \lim_{M\to\infty}\overline{\boldsymbol{p}}_0^{(M)} = \overline{\boldsymbol{p}}_0^{(\infty)}.$$

Hence $\tilde{\boldsymbol{p}}^{(M)}Q_{+1}^{(M)}\mathbf{1} - \tilde{\boldsymbol{p}}^{(M)}Q_{-1}^{(M)}\mathbf{1}$ converges to $\tilde{\boldsymbol{p}}^{(\infty)}Q_{+1}^{(\infty)}\mathbf{1} - \tilde{\boldsymbol{p}}^{(\infty)}Q_{-1}^{(\infty)}\mathbf{1}$ from the first equality in (A.7) and (A.14). Then (4.20) holds for sufficiently large M by Lemma A.1.

B. Proof of (4.22) and (4.23)

Except for minor modifications, (4.22) and (4.23) are obtained by applying similar techniques as in [14]. We partition $K^{(M)}(z)$ into blocks such that

$$K^{(M)}(z) = \begin{pmatrix}
\underline{K}_{-M,-M} & \underline{K}_{+1}(z) & & & & & & & & \\
\underline{K}_{-1} & K_0 & \underline{K}_{+1}(z) & & & & & & & \\
 & \ddots & \ddots & \ddots & & & & & & \\
 & & \underline{K}_{-1} & K_0 & \underline{K}_{+1}(z) & & & & & \\
 & & & K_{-1}(z) & K_0 & K_{+1}(z) & & & & \\
 & & & & \overline{K}_{-1}(z) & K_0 & \overline{K}_{+1} & & & \\
 & & & & & \ddots & \ddots & \ddots & & \\
 & & & & & & \overline{K}_{-1}(z) & K_0 & \overline{K}_{+1} & \\
 & & & & & & & \overline{K}_{-1}(z) & \overline{K}_{MM}
\end{pmatrix},$$

where

$$\begin{aligned}
&K_0 = C\oplus T\oplus T, \quad K_{+1}(z) = I_a\otimes \boldsymbol{t}\boldsymbol{\alpha}\otimes I_s + 2^{-1}z(D\otimes I_s\otimes I_s),\\
&K_{-1}(z) = I_a\otimes I_s\otimes \boldsymbol{t}\boldsymbol{\alpha} + 2^{-1}z(D\otimes I_s\otimes I_s), \quad \overline{K}_{+1} = I_a\otimes \boldsymbol{t}\boldsymbol{\alpha}\otimes I_s,\\
&\overline{K}_{-1}(z) = z^{-1}(I_a\otimes I_s\otimes \boldsymbol{t}\boldsymbol{\alpha}) + D\otimes I_s\otimes I_s, \quad \overline{K}_{MM} = C\oplus(T+\boldsymbol{t}\boldsymbol{\alpha})\oplus T,\\
&\underline{K}_{+1}(z) = z^{-1}(I_a\otimes \boldsymbol{t}\boldsymbol{\alpha}\otimes I_s) + D\otimes I_s\otimes I_s, \quad \underline{K}_{-1} = I_a\otimes I_s\otimes \boldsymbol{t}\boldsymbol{\alpha},\\
&\underline{K}_{-M,-M} = C\oplus T\oplus(T+\boldsymbol{t}\boldsymbol{\alpha}).
\end{aligned}$$

For $1 \le k,\ell \le M-1$ and $\boldsymbol{i},\boldsymbol{j} \in \mathcal{S}_0\times\mathcal{S}\times\mathcal{S}$, let $[\overline{N}_{k\ell}^{(M)}(z)]_{\boldsymbol{i}\boldsymbol{j}}$ be the mean sojourn time of the Markov additive process generated by $\{\overline{K}_{-1}(z), K_0, \overline{K}_{+1}\}$ at level ℓ with background state $\boldsymbol{j}$ before the additive component hits level 0 or M given that it starts from level k with background state $\boldsymbol{i}$. Similarly, for $-(M-1) \le k,\ell \le -1$, let $[\underline{N}_{k\ell}^{(M)}(z)]_{\boldsymbol{i}\boldsymbol{j}}$ be the mean sojourn time of the Markov additive process generated by $\{\underline{K}_{-1}, K_0, \underline{K}_{+1}(z)\}$ at level ℓ with background state $\boldsymbol{j}$ before the additive component hits level 0 or $-M$ given that it starts from level k with background state $\boldsymbol{i}$. When $M\to\infty$, we denote $\overline{N}_{k\ell}^{(M)}(z)$ and $\underline{N}_{k\ell}^{(M)}(z)$ by $\overline{N}_{k\ell}^{(\infty)}(z)$ and $\underline{N}_{k\ell}^{(\infty)}(z)$, respectively.

By the similar argument to the proof of Lemma 5.5 in [14], we can show that the vectors $\underline{\boldsymbol{p}} = (\underline{\boldsymbol{p}}_\ell; \ell \in \mathbb{Z})$ and $\underline{\boldsymbol{q}} = (\underline{\boldsymbol{q}}_\ell; \ell \in \mathbb{Z})$ are positive, and have the following forms:

$$\begin{aligned}
\underline{\boldsymbol{p}}_\ell &= \underline{\boldsymbol{p}}_0 K_{-1}(r_\infty^{-1}) N_{-1,\ell}^{(\infty)}(r_\infty^{-1}), \quad \ell \in \mathbb{Z}_-, \\
\underline{\boldsymbol{p}}_\ell &= \underline{\boldsymbol{p}}_0 K_{+1}(r_\infty^{-1}) N_{1\ell}^{(\infty)}(r_\infty^{-1}), \quad \ell \in \mathbb{Z}_+, \\
\underline{\boldsymbol{q}}_\ell &= N_{\ell,-1}^{(\infty)}(r_\infty^{-1}) \underline{K}_{+1}(r_\infty^{-1}) \underline{\boldsymbol{q}}_0, \quad \ell \in \mathbb{Z}_-, \\
\underline{\boldsymbol{q}}_\ell &= N_{\ell 1}^{(\infty)}(r_\infty^{-1}) \overline{K}_{-1}(r_\infty^{-1}) \underline{\boldsymbol{q}}_0, \quad \ell \in \mathbb{Z}_+,
\end{aligned}$$

where $\underline{\boldsymbol{p}}_0$ and $\underline{\boldsymbol{q}}_0$ are determined by

$$\begin{aligned}
\underline{\boldsymbol{p}}_0 \Big(K_0 + K_{+1}(r_\infty^{-1}) \overline{N}_{11}^{(\infty)}(r_\infty^{-1}) \overline{K}_{-1}(r_\infty^{-1}) + K_{-1}(r_\infty^{-1}) \underline{N}_{-1,-1}^{(\infty)}(r_\infty^{-1}) \underline{K}_{+1}(r_\infty^{-1}) \Big) &= \mathbf{0}, \\
\Big(K_0 + K_{+1}(r_\infty^{-1}) \overline{N}_{11}^{(\infty)}(r_\infty^{-1}) \overline{K}_{-1}(r_\infty^{-1}) + K_{-1}(r_\infty^{-1}) \underline{N}_{-1,-1}^{(\infty)}(r_\infty^{-1}) \underline{K}_{+1}(r_\infty^{-1}) \Big) \underline{\boldsymbol{q}}_0 &= \mathbf{0}.
\end{aligned}$$

Since $\underline{K}_{-1} + K_0 + \underline{K}_{+1}(r_\infty^{-1})$ and $\overline{K}_{-1}(r_\infty^{-1}) + K_0 + \overline{K}_{+1}$ are defective matrices, $N_{-1,\ell}^{(\infty)}(r_\infty^{-1})$, $N_{1\ell}^{(\infty)}(r_\infty^{-1})$, $N_{\ell,-1}^{(\infty)}(r_\infty^{-1})$ and $N_{\ell 1}^{(\infty)}(r_\infty^{-1})$ geometrically decay entry wise as $|\ell|$ increases. Hence we have $\underline{\boldsymbol{p}}\ \underline{\boldsymbol{q}} < \infty$, which completes the proof of (4.22).

We next show the existence of the positive vectors $\boldsymbol{p}^{(\infty)}$ and $\boldsymbol{q}^{(\infty)}$ satisfying (4.23). Let $\boldsymbol{q}^{(\infty)} = (\boldsymbol{q}_\ell^{(\infty)}; \ell \in \mathbb{Z})$, where $\boldsymbol{q}_\ell^{(\infty)} \equiv r^{|\ell|}\boldsymbol{u}$ for positive vector $\boldsymbol{u}$ in Lemma 4.1. Then it is easy to check that $K^{(\infty)}(r^{-2})\boldsymbol{q}^{(\infty)} = \mathbf{0}$ by (4.8). Similarly to the proof of Lemma 4.4, let $\tilde{K} = \Delta_{\boldsymbol{q}^{(\infty)}}^{-1} K^{(\infty)}(r^{-2}) \Delta_{\boldsymbol{q}^{(\infty)}}$, i.e.,

$$\tilde{K} = \begin{pmatrix}
\ddots & \ddots & \ddots & & & & \\
 & \underline{K}_{-1} & K_0 & \underline{K}_{+1} & & & \\
 & & K_{-1} & K_0 & K_{+1} & & \\
 & & & \overline{K}_{-1} & K_0 & \overline{K}_{+1} & \\
 & & & & \ddots & \ddots & \ddots
\end{pmatrix},$$

where

$$\begin{aligned}
&K_0 = \Delta_{\boldsymbol{u}}^{-1}(C \oplus T \oplus T)\Delta_{\boldsymbol{u}}, \quad K_{+1} = \Delta_{\boldsymbol{u}}^{-1}(r(I_a \otimes \boldsymbol{t}\boldsymbol{\alpha} \otimes I_s) + 2^{-1}r(D \otimes I_s \otimes I_s))\Delta_{\boldsymbol{u}}, \\
&K_{-1} = \Delta_{\boldsymbol{u}}^{-1}(r(I_a \otimes I_s \otimes \boldsymbol{t}\boldsymbol{\alpha}) + 2^{-1}r^{-1}(D \otimes I_s \otimes I_s))\Delta_{\boldsymbol{u}}, \quad \overline{K}_{+1} = \Delta_{\boldsymbol{u}}^{-1}(r(I_a \otimes \boldsymbol{t}\boldsymbol{\alpha} \otimes I_s))\Delta_{\boldsymbol{u}}, \\
&\overline{K}_{-1} = \Delta_{\boldsymbol{u}}^{-1}(r(I_a \otimes I_s \otimes \boldsymbol{t}\boldsymbol{\alpha}) + r^{-1}(D \otimes I_s \otimes I_s))\Delta_{\boldsymbol{u}}, \\
&\underline{K}_{+1} = \Delta_{\boldsymbol{u}}^{-1}(r(I_a \otimes \boldsymbol{t}\boldsymbol{\alpha} \otimes I_s) + r^{-1}(D \otimes I_s \otimes I_s))\Delta_{\boldsymbol{u}}, \quad \underline{K}_{-1} = \Delta_{\boldsymbol{u}}^{-1}(r(I_a \otimes I_s \otimes \boldsymbol{t}\boldsymbol{\alpha}))\Delta_{\boldsymbol{u}}.
\end{aligned}$$

It is easy to see that the Markov additive processes generated by $\{\overline{K}_{-1}, K_0, \overline{K}_{+1}\}$ and $\{\underline{K}_{-1}, K_0, \underline{K}_{+1}\}$, respectively, have drifts to the origin. Hence $\tilde{K}$ has the stationary distribution $\tilde{\boldsymbol{\xi}}$. Let $\boldsymbol{p}^{(\infty)} = \tilde{\boldsymbol{\xi}}\Delta_{\boldsymbol{q}^{(\infty)}}^{-1}$, then we have

$$\boldsymbol{p}^{(\infty)}\boldsymbol{q}^{(\infty)} = \tilde{\boldsymbol{\xi}}\mathbf{1} = 1 < \infty,$$

which completes the proof of (4.23).

References

[1] S. Asmussen: *Applied Probability and Queues*, second edition (Springer, New York, 2003).

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

[3] R. Bellman: *Introduction to Matrix Analysis*, second edition (SIAM, Philadelphia, 1997).

[4] R.D. Foley and D.R. McDonald: Join the shortest queue: stability and exact asymptotics. *Annals of Applied Probability*, **11** (2001), 569–607.

[5] J.F.C. Kingman: Two similar queues in parallel. *Annals of Mathematical Statistics*, **32** (1961), 1314–1323.

[6] G. Latouche and V. Ramaswami: *Introduction to Matrix Analytic Methods in Stochastic Modeling* (American Statistical Association and the Society for Industrial and Applied Mathematics, Philadelphia, 1999).

[7] H. Li, M. Miyazawa, and Y.Q. Zhao: Geometric decay in a QBD process with countable background states with applications to shortest queues. *Stochastic Models*, **23** (2007), 413-438.

[8] H. Li and Y.Q. Zhao: Stochastic block-monotonicity in the approximation of the stationary distribution of infinite Markov chains. *Stochastic Models*, **16** (2000), 313–333.

[9] M. Miyazawa: Two sided DQBD process and solutions to the tail decay rate problem and their applications to the generalized join shortest queue. In W. Yue, Y. Takahashi and H. Takagi (eds.): *Advances in Queueing Theory and Network Applications* (Springer, Cambridge, MA, 2008), 3–33.

[10] M. Miyazawa: Tail decay rates in double QBD processes and related reflected random walks. *Mathematics of Operations Research*, **34** (2009), 547–575.

[11] M. Miyazawa and Y.Q. Zhao: The stationary tail asymptotics in the $GI/G/1$ type queue with countably many background states. *Advances in Applied Probability*, **36** (2004), 1231–1251.

[12] M.F. Neuts: *Matrix-Geometric Solutions in Stochastic Models* (Johns Hopkins University Press, Baltimore, 1983).

[13] Y. Sakuma: Asymptotic behavior for MArP/PH/c queue with shortest queue discipline and jockeying. *Operations Research Letters*, **38** (2010), 7-10.

[14] Y. Sakuma, M. Miyazawa, and Y.Q. Zhao: Decay rate for a PH/M/2 queue with shortest queue discipline. *Queueing Systems*, **53** (2006), 189-202.

[15] E. Seneta: *Nonnegative Matrices and Markov Chains* (Springer, New York, 1981).

[16] Y. Takahashi, K. Fujimoto, and N. Makimoto: Geometric decay of the steady-state probabilities in a quasi-birth-and-death process with a countable number of phases. *Stochastic Models*, **17** (2001), 1–24.

Yutaka Sakuma
Department of Information Sciences
Tokyo University of Science
Noda-City, Chiba 278-8510, Japan
E-mail: sakuma-y@rs.noda.tus.ac.jp