RAIRO-Oper. Res. 51 (2017) 945–963
DOI: 10.1051/ro/2017012

RAIRO Operations Research
www.rairo-ro.org

ANALYSIS OF A TWO-CLASS CONTINUOUS-TIME QUEUEING MODEL WITH TWO TANDEM DEDICATED SERVERS

Tao jiang[1] and Liwei liu[2]

Abstract. Mélange *et al.* (2016) investigated a continuous-time queueing system with two types of customers each having their own dedicated server, where the two dedicated servers are in parallel and have different service rates, meanwhile, the system adopts a global First-Come-First-Served (gFCFS) service discipline, *i.e.*, all new arrivals queue together in a common FCFS queue, regardless of their types. In the present paper, we aim to give a further study on this queueing model, in which the two dedicated servers are accommodated in series. By using matrix analytic method and spectral expansion method, steady state probabilities are derived to make the straightforward computation of performance measures and the sojourn time of an arbitrary customer. Finally, some numerical examples are provided to show the effect of several system parameters on performance measures.

Mathematics Subject Classification. 68M20, 60K20, 90B22.

Received August 29, 2016. Accepted March 2, 2017.

1. Introduction

In most traditional queueing models, a service facility provides exactly one type of service and that all customers requiring this type of service are accommodated in one common queue. For multi-class customers in a queueing system, multiple different service facilities are needed to provide service for each type of customers, and individual queues are formed before these service facilities. In these queueing models, customers are only blocked by other customers of same type. However, in real life, some queueing systems also adopt a gFCFS service discipline, *i.e.*, customers requiring different types of service are accommodated in a common queue and are served in their order of arrival, regardless of the class they belong to. In these models, customers of one given type are not only hindered by customers of same type, but also hindered by customers of other types.

For this kind of queueing models, Bruneel *et al.* [1] studied a simple discrete-time queueing model with two types of customers each having their own dedicated server under a gFCFS discipline, in which the service times of all customers are deterministically equal to 1 slot each. They also applied the queueing model into practice, such as security checkpoint in international airports or train stations, switching nodes of telecommunication networks and traffic junctions in the context of road networks, *etc.* We refer to Bruneel *et al.* [1] for more details of those applications. Then, Bruneel *et al.* [2] continued to investigate the effect of gFCFS and relative

Keywords. Continuous-time, dedicated servers, tandem, steady state, sojourn time.

[1] College of Economics and Management, Shandong University of Science and Technology, Qingdao, 266590, China.
jtao0728@163.com

[2] School of Science, Nanjing University of Science and Technology, Nanjing, 210094, China.

Article published by EDP Sciences

load distribution in two-class queues with dedicated servers. In [2], they gave the stability condition of the system and derived the system size distribution at random slot boundaries. Recently, Mélange *et al.* [3] went on considering a continuous-time queueing model with class clustering (customers of the same type having the tendency to arrive "back-to-back") and gFCFS policy, where customers of different types have the same service rate μ, and gave the system size distribution and sojourn time distribution of an arbitrary customer. Next, Mélange *et al.* [4] analysed a continuous-time queueing system with two classes of customers each having their own dedicated server, where the two servers have different service rates, and analysed the blocked impact in this model, meanwhile, they gave a comparison between two systems: one with block effect and one without block effect. Lately, Bruneel *et al.* [5] extended the model in Bruneel *et al.* [1], where service times are deterministically equal to $s \geq 1$ time slots and arriving customers enter into the system according to a general independent arrival process.

From these papers, we find that the two dedicated servers in these queueing models are in parallel, however, in some practical scenarios, such as toll collection systems, we may encounter this situation: the servers are accommodated in series. Some excellent papers on tollbooth tandem queues can be seen in He and Chao [6], Chao *et al.* [7] and Do [8], etc. Actually, two-class queueing model with two tandem dedicated servers also has important applications in real life, especially in China. Indeed, at highway exit in some cities of China, we may experience the two dedicated servers are accommodated in series. For example, at a highway exit, if the vehicles want to enter into downtown, they need to go through tollbooths. Usually, the highway exit consists of a waiting region and two tollbooth in series, one is for the local vehicles, the other one is for non-local vehicles. Vehicles receive service from their dedicated tollbooths, and each vehicle receives service from one and only one tollbooth. From Figure 1a, local vehicles (customer 1) receive service from tollbooth 1 (server 1), and non-local vehicles (customer 2) have to go through tollbooth 1, and receive service from tollbooth 2 (server 2). After served by their dedicated tollbooths, they continue to drive along their own exit path into downtown. In this queueing model, customers are not only blocked by the gFCFS service discipline, but also by the two tandem servers. It is because that non-local vehicles have to go through tollbooth 1, *i.e.*, if tollbooth 1 is busy and tollbooth 2 is idle, even if a non-local vehicle situates at the head of the waiting room, it may be blocked to get service from tollbooth 2 and must wait until the departure of a local vehicle at tollbooth 1.

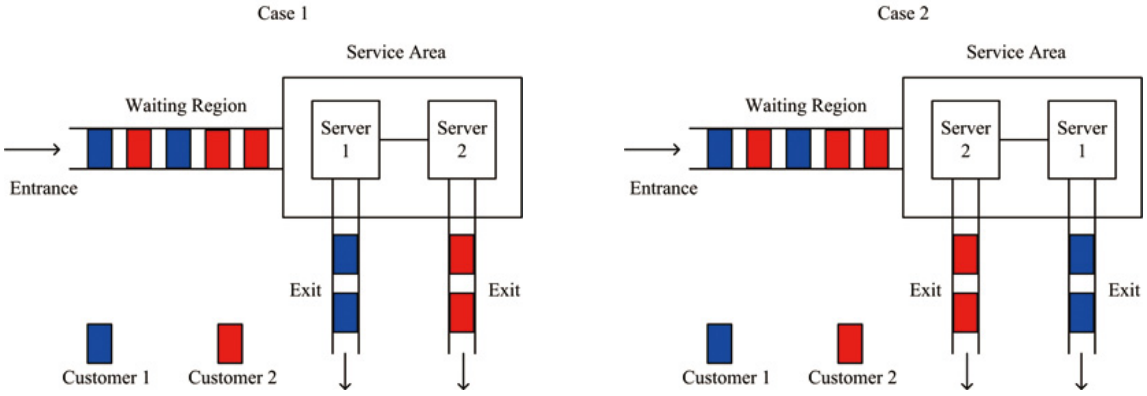
Motivated by the applications of the servers are accommodated in series, it is crucial to analyse the queueing model, so, in this paper, we extend the queueing model in Mélange *et al.* [4]. Different from Mélange *et al.* [4], we assume that the two dedicated are accommodated in series. We further investigate two cases (See Figs. 1a and 1b) by matrix analytic method and spectral expansion method, and give a comparison between the two cases in terms of some parameters. We hope that our results can provide guidance suggestions on how to keep the waiting region clear and reduce traffic congestion at the highway exit.

The paper is organized as follows: in Section 2, we give the model description. In Section 3, we obtain the sufficient and necessary stability condition. Section 4 is devoted to giving the steady state probabilities by matrix analytic method and spectral expansion method, respectively. Section 5 gives various performance measures and sojourn time distribution of an arbitrary customer. Numerical examples are presented in Section 6. Section 7 is the conclusion.

2. Model description

We investigate a continuous-time queue that consists of two types customers and two tandem dedicated servers. According to the order of the two servers, we consider two cases of the system (See Figs. 1a and 1b). The queueing model is described in detail below:

(1) Customers arrive the system according to a Poisson process with arrival rate λ.
(2) The types of consecutive customers are independent, *i.e.*, an arriving customer is of type 1 with probability p and of type 2 with probability $q = 1 - p$. That is, the arrival rate of the two types customers are denoted by λ_1 and λ_2, where $\lambda_1 = \lambda p$ and $\lambda_2 = \lambda q$.



(a) Case 1 (b) Case 2

FIGURE 1. The structure of the queueing system.

(3) Service times follow an exponential distribution with parameter μ_1 at server 1 and parameter μ_2 at server 2. The servers are dedicated to a given type of customers, *i.e.*, server 1 only serves type 1 customers and server 2 serves type 2 customers.
(4) There is a waiting room of infinite size in front of server 1 (Case 1) or server 2 (Case 2) and there is no waiting room between the two servers.

According to the structure of the queueing system, there are two blocking effects in our system. One is the gFCFS service discipline, which is same with Mélange *et al.* [4]. The other one is the two tandem servers. Since we assume that the two servers are in series, in Case 1, type 2 customers need to go through server 1 to receive service at server 2. Hence, if server 1 is busy and server 2 is idle, even if a type 2 customer situates at the head of the waiting room, he may be blocked to get service from server 2 and must wait until the departure of type 1 customer at server 1. Oppositely, in Case 2, type 1 customers need to go through server 2 to receive service at server 1.

In the following part, we focus on the analysis of Case 1, and Case 2 can be analysed with the same method. So, let $N(t)$ denote the number of customers in the system at time t, $I(t)$ denote the customer type in the first position of the line at time t, and $J(t)$ denote the customer type in the second position of the line at time t. The whole system can be described by a continuous-time Markov chain where the state of the system is characterized by $\{(N(t), I(t), J(t)), t \geq 0\}$, with state space

$$\Omega = \{\{0\} \cup \{(1,1)\} \cup \{(1,2)\} \cup \{(n,i,j), n \geq 2, i,j = 1,2\}\}.$$

$\{0\}$ is the state that there are no customers in the system. $\{(1,1)\}$ is the state that there is only a customer in the system, and the first position is type 1 customer. $\{(1,2)\}$ denotes the state that there is a customer in the system, and the first position is type 2 customer. In order to avoid confusion, we give an explicit explanation on the states as follows: For $n \geq 2$,

(1) $(n,1,1)$ denotes that there are n customers in the system, the two customers at the front of the line are both of type 1, that is, the oldest one (in the first position of the line) is served at server 1, the second oldest one (in the second position of the line or at the head of the waiting room) is waiting at the front of the waiting room.
(2) $(n,1,2)$ denotes that the types of two customers at the front of the line are different, *i.e.*, the oldest one (in the first position of the line) is type 1, which is in service at server 1, and the second oldest one (in the

second position of the line or at the head of the waiting room) is type 2, which is blocked to receive service from server 2.

(3) $(n, 2, 1)$ denotes that both of the servers are busy, *i.e.*, the oldest customer (in the first position of the line) is type 2, which is in service at server 2, and the second oldest customer (in the second position of the line) is type 1, which is served at server 1.

(4) $(n, 2, 2)$ denotes that there are n customers in the system, the two customers at the front of the line are both of type 2, that is, one is served at server 2, the other one is waiting at the head of the waiting room.

3. Stability analysis

In this section, we use the mean drift result of Neuts [9] to obtain the stability condition. By referring to the continuous-time Markov process, we can obtain the state-transition-rate matrix as follows

$$\boldsymbol{Q} = \begin{pmatrix} \overline{B}_0 & \overline{\boldsymbol{B}}_{\boldsymbol{1}} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots \\ \overline{\boldsymbol{B}}_{\boldsymbol{2}} & \overline{\boldsymbol{A}}_{\boldsymbol{0}} & \overline{\boldsymbol{A}}_{\boldsymbol{1}} & \mathbf{0} & \mathbf{0} & \dots \\ \mathbf{0} & \overline{\boldsymbol{A}}_{\boldsymbol{2}} & \boldsymbol{A}_{\boldsymbol{0}} & \boldsymbol{A}_{\boldsymbol{1}} & \mathbf{0} & \dots \\ \mathbf{0} & \mathbf{0} & \boldsymbol{A}_{\boldsymbol{2}} & \boldsymbol{A}_{\boldsymbol{0}} & \boldsymbol{A}_{\boldsymbol{1}} & \dots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \boldsymbol{A}_{\boldsymbol{2}} & \boldsymbol{A}_{\boldsymbol{0}} & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}, \tag{3.1}$$

where

$$\overline{B}_0 = -\lambda, \overline{\boldsymbol{B}}_{\boldsymbol{1}} = (\lambda p, \lambda q), \overline{\boldsymbol{B}}_{\boldsymbol{2}} = (\mu_1, \mu_2)^T,$$

$$\overline{\boldsymbol{A}}_{\boldsymbol{0}} = \begin{pmatrix} -(\lambda+\mu_1) & 0 \\ 0 & -(\lambda+\mu_2) \end{pmatrix}, \overline{\boldsymbol{A}}_{\boldsymbol{1}} = \begin{pmatrix} \lambda p & \lambda q & 0 & 0 \\ 0 & 0 & \lambda p & \lambda q \end{pmatrix}, \overline{\boldsymbol{A}}_{\boldsymbol{2}} = \begin{pmatrix} \mu_1 & 0 \\ 0 & \mu_1 \\ \mu_2 & \mu_1 \\ 0 & \mu_2 \end{pmatrix},$$

$$\boldsymbol{A}_{\boldsymbol{0}} = \begin{pmatrix} -(\lambda+\mu_1) & 0 & 0 & 0 \\ 0 & -(\lambda+\mu_1) & 0 & 0 \\ 0 & 0 & -(\lambda+\mu_1+\mu_2) & 0 \\ 0 & 0 & 0 & -(\lambda+\mu_2) \end{pmatrix},$$

$$\boldsymbol{A}_{\boldsymbol{1}} = \begin{pmatrix} \lambda & 0 & 0 & 0 \\ 0 & \lambda & 0 & 0 \\ 0 & 0 & \lambda & 0 \\ 0 & 0 & 0 & \lambda \end{pmatrix}, \boldsymbol{A}_{\boldsymbol{2}} = \begin{pmatrix} p\mu_1 & q\mu_1 & 0 & 0 \\ 0 & 0 & p\mu_1 & q\mu_1 \\ p\mu_2 & q\mu_2 & p\mu_1 & q\mu_1 \\ 0 & 0 & p\mu_2 & q\mu_2 \end{pmatrix}.$$

Once state-transition-rate matrix is obtained, we can investigate the sufficient and necessary stability condition of our model in the following theorem.

Theorem 3.1. *The system under consideration is stable if and only if*

$$\rho = \frac{\lambda(p{\mu_2}^2 + q\mu_1\mu_2 + q^2{\mu_1}^2)}{\mu_1{\mu_2}^2 + q{\mu_1}^2\mu_2} < 1. \tag{3.2}$$

Proof. Based on the mean drift result in Neuts [9], the system would be stable and the stationary probability exists if and only if

$$\boldsymbol{x}\boldsymbol{A_1}\boldsymbol{e} < \boldsymbol{x}\boldsymbol{A_2}\boldsymbol{e},$$

where $\boldsymbol{e}$ is a column vector with four dimensions and all its elements are equal to one, $\boldsymbol{x} = (\tilde{x}_0, \tilde{x}_1, \tilde{x}_2, \tilde{x}_3)$ is the invariant probability vector of $\boldsymbol{A} = \boldsymbol{A_0} + \boldsymbol{A_1} + \boldsymbol{A_2}$, which satisfies $\boldsymbol{x}\boldsymbol{A} = \boldsymbol{0}$ and $\boldsymbol{x}\boldsymbol{e} = 1$. Notice that the generator $\boldsymbol{A}$

$$\boldsymbol{A} = \boldsymbol{A_0} + \boldsymbol{A_1} + \boldsymbol{A_2} = \begin{pmatrix} -q\mu_1 & q\mu_1 & 0 & 0 \\ 0 & -\mu_1 & p\mu_1 & q\mu_1 \\ p\mu_2 & q\mu_2 & -(q\mu_1 + \mu_2) & q\mu_1 \\ 0 & 0 & p\mu_2 & -p\mu_2 \end{pmatrix}$$

is irreducible, then, an immediate result is that

$$\tilde{x}_0 = \frac{p^2{\mu_2}^2}{p{\mu_2}^2 + q\mu_1\mu_2 + q^2{\mu_1}^2}, \tilde{x}_1 = \frac{pq{\mu_2}^2}{p{\mu_2}^2 + q\mu_1\mu_2 + q^2{\mu_1}^2},$$

$$\tilde{x}_2 = \frac{pq\mu_1\mu_2}{p{\mu_2}^2 + q\mu_1\mu_2 + q^2{\mu_1}^2}, \tilde{x}_3 = \frac{q^2\mu_1(\mu_1 + \mu_2)}{p{\mu_2}^2 + q\mu_1\mu_2 + q^2{\mu_1}^2}.$$

So from $\boldsymbol{x}\boldsymbol{A_1}\boldsymbol{e} < \boldsymbol{x}\boldsymbol{A_2}\boldsymbol{e}$, we can derive the sufficient and necessary stability condition

$$\lambda < \frac{\mu_1{\mu_2}^2 + q{\mu_1}^2\mu_2}{p{\mu_2}^2 + q\mu_1\mu_2 + q^2{\mu_1}^2},$$

which is equivalent to

$$\rho = \frac{\lambda(p{\mu_2}^2 + q\mu_1\mu_2 + q^2{\mu_1}^2)}{\mu_1{\mu_2}^2 + q{\mu_1}^2\mu_2} < 1. \qquad \square$$

Actually, using similar method in Mélange *et al.* [4], we can also obtain the stability condition. The more details can be seen in Appendix.

If $\rho < 1$, we continue to give the steady state probabilities for this queueing system. First, define the steady state probabilities by

$$\pi_0 = \lim_{t\to\infty} P(N(t) = 0), \pi_{1,k} = \lim_{t\to\infty} P(N(t) = 1, I(t) = k), k = 1, 2,$$

$$\pi_{n,i,j} = \lim_{t\to\infty} P(N(t) = n, I(t) = i, J(t) = j), n \geq 2, i, j = 1, 2,$$

$$\boldsymbol{\pi_1} = (\pi_{1,1}, \pi_{1,2}), \boldsymbol{\pi_k} = (\pi_{k,1,1}, \pi_{k,1,2}, \pi_{k,2,1}, \pi_{k,2,2}), k \geq 2, \boldsymbol{\pi} = (\pi_0, \boldsymbol{\pi_1}, \boldsymbol{\pi_2}, \ldots).$$

Then, the balance equations can be written as follows.

$$\lambda\pi_0 = \mu_1\pi_{1,1} + \mu_2\pi_{1,2}, \tag{3.3}$$

$$(\lambda + \mu_1)\pi_{1,1} = \mu_1\pi_{2,1,1} + \mu_2\pi_{2,2,1} + \lambda_1\pi_0, \tag{3.4}$$

$$(\lambda + \mu_2)\pi_{1,2} = \mu_1(\pi_{2,1,2} + \pi_{2,2,1}) + \mu_2\pi_{2,2,2} + \lambda_2\pi_0, \tag{3.5}$$

$$(\lambda + \mu_1)\pi_{2,1,1} = \lambda_1\pi_{1,1} + p\mu_1\pi_{3,1,1} + p\mu_2\pi_{3,2,1}, \tag{3.6}$$

$$(\lambda + \mu_1)\pi_{2,1,2} = \lambda_2\pi_{1,1} + q\mu_1\pi_{3,1,1} + q\mu_2\pi_{3,2,1}, \tag{3.7}$$

$$(\lambda+\mu_1+\mu_2)\pi_{2,2,1}=\lambda_1\pi_{1,2}+p\mu_1\pi_{3,1,2}+p\mu_1\pi_{3,2,1}+p\mu_2\pi_{3,2,2}, \tag{3.8}$$

$$(\lambda+\mu_2)\pi_{2,2,2}=\lambda_2\pi_{1,2}+q\mu_1\pi_{3,1,2}+q\mu_1\pi_{3,2,1}+q\mu_2\pi_{3,2,2}, \tag{3.9}$$

$$(\lambda+\mu_1)\pi_{n,1,1}=\lambda\pi_{n-1,1,1}+p\mu_1\pi_{n+1,1,1}+p\mu_2\pi_{n+1,2,1}, n\geq 3, \tag{3.10}$$

$$(\lambda+\mu_1)\pi_{n,1,2}=\lambda\pi_{n-1,1,2}+q\mu_1\pi_{n+1,1,1}+q\mu_2\pi_{n+1,2,1}, n\geq 3, \tag{3.11}$$

$$\begin{aligned}(\lambda+\mu_1+\mu_2)\pi_{n,2,1}=&\lambda\pi_{n-1,2,1}+p\mu_1\pi_{n+1,1,2}\\&+p\mu_1\pi_{n+1,2,1}+p\mu_2\pi_{n+1,2,2}, n\geq 3,\end{aligned} \tag{3.12}$$

$$\begin{aligned}(\lambda+\mu_2)\pi_{n,2,2}=&\lambda\pi_{n-1,2,2}+q\mu_1\pi_{n+1,1,2}\\&+q\mu_1\pi_{n+1,2,1}+q\mu_2\pi_{n+1,2,2}, n\geq 3.\end{aligned} \tag{3.13}$$

The normalization equation is

$$\pi_0+\pi_{1,1}+\pi_{1,2}+\sum_{n=2}^{\infty}(\pi_{n,1,1}+\pi_{n,1,2}+\pi_{n,2,1}+\pi_{n,2,2})=1. \tag{3.14}$$

4. Steady state probabilities

In this section, we first use matrix analytic method to find the steady state probabilities for the quasi birth-death (QBD) process. In order to analyse the system effectively by matrix analytic method, we need to derive the rate matrix $\boldsymbol{R}$, which is the minimal non-negative solution of

$$\boldsymbol{R}^2\boldsymbol{A_2}+\boldsymbol{R}\boldsymbol{A_0}+\boldsymbol{A_1}=\boldsymbol{0}. \tag{4.1}$$

In fact, due to the special structure of matrix $\boldsymbol{A_2}$, it is difficult to obtain the explicit expression of rate matrix $\boldsymbol{R}$. However, some existing algorithms such as the monotonic iterative algorithms (Neuts [9]), the matrix continued fraction approach presented in (Phung–Duc *et al.* [10]) and more efficient algorithms (Bini *et al.* [11]) can be used to compute the rate matrix $\boldsymbol{R}$.

Here, we use the iterative algorithms in Latouche and Ramaswami [12] to obtain $\boldsymbol{R}$. Based on the description of Theorem 8.7.2 in Latouche and Ramaswami [12], if the QBD is recurrent, then the sequences $\{\boldsymbol{U}(k), k\geq 1\}$ and $\{\boldsymbol{G}(k), k\geq 1\}$, where

$$\begin{aligned}\boldsymbol{U}(1)&=\boldsymbol{A_0}+\boldsymbol{A_1},\\ \boldsymbol{G}(k)&=(-\boldsymbol{U(k)})^{-1}\boldsymbol{A_2},\\ \boldsymbol{U}(k+1)&=\boldsymbol{A_0}+\boldsymbol{A_1}\boldsymbol{G}(k),\end{aligned}$$

for $k\geq 1$, are such that $\boldsymbol{U}(k)$ is substochastic and $\boldsymbol{G}(k)$ is stochastic for all k. Further, the two sequences are monotonically increasing and converge to $\boldsymbol{U}$ and $\boldsymbol{G}$, respectively. The iterative procedure stops until the condition

$$\|\boldsymbol{G}(k+1)-\boldsymbol{G}(k)\|_\infty\leq\varepsilon$$

satisfied. Then, the rate matrix $\boldsymbol{R}$ can be obtained by

$$\boldsymbol{R}=\boldsymbol{A_1}(-\boldsymbol{A_0}-\boldsymbol{A_1}\boldsymbol{G})^{-1}.$$

Once $\boldsymbol{R}$ is obtained, according to the matrix analytic method, we have

$$\boldsymbol{\pi_k}=\boldsymbol{\pi_2}\boldsymbol{R}^{k-2},\qquad k\geq 3.$$

Finally, the boundary vectors $\pi_0, \boldsymbol{\pi_1}, \boldsymbol{\pi_2}$ can be obtained by

$$\mathbf{0} = (\pi_0, \boldsymbol{\pi_1}, \boldsymbol{\pi_2})\boldsymbol{B}[\boldsymbol{R}], \tag{4.2}$$

$$1 = \pi_0 + \boldsymbol{\pi_1}\boldsymbol{e_2} + \sum_{i=2}^{\infty} \boldsymbol{\pi_i}\boldsymbol{e}, \tag{4.3}$$

where

$$\boldsymbol{B}[\boldsymbol{R}] = \begin{pmatrix} \overline{B}_0 & \overline{\boldsymbol{B}}_{\mathbf{1}} & \mathbf{0} \\ \overline{\boldsymbol{B}}_{\mathbf{2}} & \overline{\boldsymbol{A}}_{\mathbf{0}} & \overline{\boldsymbol{A}}_{\mathbf{1}} \\ \mathbf{0} & \overline{\boldsymbol{A}}_{\mathbf{2}} & \boldsymbol{A_0} + \boldsymbol{R}\boldsymbol{A_2} \end{pmatrix}, \qquad \boldsymbol{e_2} = (1,1)', \boldsymbol{e} = (1,1,1,1)'.$$

By using the censoring technique and referring to Li [13], we construct the UL-type RG-factorization for $\boldsymbol{B}[\boldsymbol{R}]$ to derive the expressions of $\pi_0, \boldsymbol{\pi_1}, \boldsymbol{\pi_2}$. First, we write the U-measure as

$$\boldsymbol{U_2} = \boldsymbol{R}\boldsymbol{A_2} + \boldsymbol{A_0}, \boldsymbol{U_1} = \overline{\boldsymbol{A}}_{\mathbf{0}} + \overline{\boldsymbol{A}}_{\mathbf{1}}(-\boldsymbol{U_2})^{-1}\overline{\boldsymbol{A}}_{\mathbf{2}}, U_0 = \overline{B}_0 + \overline{\boldsymbol{B}}_{\mathbf{1}}(-\boldsymbol{U_1})^{-1}\overline{\boldsymbol{B}}_{\mathbf{2}},$$

where $\boldsymbol{U_k}$ is the infinite generator obtained by Censoring technique for $0 \le k \le 2$, and from [13], the Markov chain $\boldsymbol{U_k}$ is transient, and thus the matrix $\boldsymbol{U_k}$ is invertible for $1 \le k \le 2$. While the Markov chain U_0 is positive recurrent if and only if the Markov chain $\boldsymbol{B}[\boldsymbol{R}]$ is positive recurrent. Based on the U-measure, we can define the UL-type R-measure and G-measure as follows:

$$\boldsymbol{R_1} = \overline{\boldsymbol{A}}_{\mathbf{1}}(-\boldsymbol{U_2})^{-1}, \boldsymbol{R_0} = \overline{\boldsymbol{B}}_{\mathbf{1}}(-\boldsymbol{U_1})^{-1},$$

$$\boldsymbol{G_1} = (-\boldsymbol{U_1})^{-1}\overline{\boldsymbol{B}}_{\mathbf{2}}, \boldsymbol{G_2} = (-\boldsymbol{U_2})^{-1}\overline{\boldsymbol{A}}_{\mathbf{2}}.$$

Note that the matrix $\boldsymbol{R_0}, \boldsymbol{R_1}$ and $\boldsymbol{G_1}, \boldsymbol{G_2}$ satisfy

$$\overline{\boldsymbol{B}}_{\mathbf{1}} + \boldsymbol{R_0}\overline{\boldsymbol{A}}_{\mathbf{0}} + \boldsymbol{R_0}\boldsymbol{R_1}\overline{\boldsymbol{A}}_{\mathbf{2}} = \mathbf{0},$$

and

$$\overline{\boldsymbol{A}}_{\mathbf{1}}\boldsymbol{G_2}\boldsymbol{G_1} + \overline{\boldsymbol{A}}_{\mathbf{0}}\boldsymbol{G_1} + \overline{\boldsymbol{B}}_{\mathbf{2}} = \mathbf{0},$$

with the boundary conditions

$$\boldsymbol{R_1} = \overline{\boldsymbol{A}}_{\mathbf{1}}(-\boldsymbol{U_2})^{-1}, \boldsymbol{G_2} = (-\boldsymbol{U_2})^{-1}\overline{\boldsymbol{A}}_{\mathbf{2}}.$$

Then, we have

$$\boldsymbol{R_0} = -\overline{\boldsymbol{B}}_{\mathbf{1}}(\overline{\boldsymbol{A}}_{\mathbf{0}} + \boldsymbol{R_1}\overline{\boldsymbol{A}}_{\mathbf{2}})^{-1},$$

and

$$\boldsymbol{G_1} = -(\overline{\boldsymbol{A}}_{\mathbf{1}}\boldsymbol{G_2} + \overline{\boldsymbol{A}}_{\mathbf{0}})^{-1}\overline{\boldsymbol{B}}_{\mathbf{2}}.$$

The UL-type RG-factorization of $\boldsymbol{B}[\boldsymbol{R}]$ is given by

$$\boldsymbol{B}[\boldsymbol{R}] = (\boldsymbol{I} - \boldsymbol{R_U})\boldsymbol{U_D}(\boldsymbol{I} - \boldsymbol{G_L}),$$

where

$$\boldsymbol{R_U} = \begin{pmatrix} 0 & \boldsymbol{R_0} & 0 \\ 0 & 0 & \boldsymbol{R_1} \\ 0 & 0 & 0 \end{pmatrix}, \quad \boldsymbol{G_L} = \begin{pmatrix} 0 & 0 & 0 \\ \boldsymbol{G_1} & 0 & 0 \\ 0 & \boldsymbol{G_2} & 0 \end{pmatrix}, \quad \boldsymbol{U_D} = \mathrm{diag}(U_0, \boldsymbol{U_1}, \boldsymbol{U_2}),$$

with $U_0 = 0$ and $\boldsymbol{U_1}, \boldsymbol{U_2}$ are invertible. Finally, the steady state probabilities can be obtained by the following theorem.

Theorem 4.1. *If $\rho < 1$, the stationary state probabilities are given as follows*:

$$\pi_0 = K, \boldsymbol{\pi_k} = K \prod_{i=0}^{k-1} \boldsymbol{R_i}, 1 \le k \le 2, \boldsymbol{\pi_k} = \boldsymbol{\pi_2} \boldsymbol{R}^{k-2}, k \ge 3, \tag{4.4}$$

where $K = (1 + \boldsymbol{R_0 e_2} + \boldsymbol{R_0 R_1} (\boldsymbol{I} - \boldsymbol{R})^{-1} \boldsymbol{e})^{-1}$.

Proof. Substituting the UL-type RG-factorization of $\boldsymbol{B}[\boldsymbol{R}]$ into $(\pi_0, \boldsymbol{\pi_1}, \boldsymbol{\pi_2}) \boldsymbol{B}[\boldsymbol{R}] = \boldsymbol{0}$, we have

$$(\pi_0, \boldsymbol{\pi_1}, \boldsymbol{\pi_2})(\boldsymbol{I} - \boldsymbol{R_U}) \boldsymbol{U_D} (\boldsymbol{I} - \boldsymbol{G_L}) = \boldsymbol{0}.$$

From the definition of $\boldsymbol{G_L}$, we know that $\boldsymbol{I} - \boldsymbol{G_L}$ is invertible and

$$(\pi_0, \boldsymbol{\pi_1}, \boldsymbol{\pi_2})(\boldsymbol{I} - \boldsymbol{R_U}) \boldsymbol{U_D} = \boldsymbol{0}.$$

Let

$$(Y_0, \boldsymbol{Y_1}, \boldsymbol{Y_2}) = (\pi_0, \boldsymbol{\pi_1}, \boldsymbol{\pi_2})(\boldsymbol{I} - \boldsymbol{R_U}),$$

using the properties of the U-measure, we have

$$\boldsymbol{Y_k U_k} = \boldsymbol{0}, 0 \le k \le 2,$$

and

$$\begin{cases} Y_0 = \pi_0, \\ \boldsymbol{Y_k} = \boldsymbol{\pi_k} - \boldsymbol{\pi_{k-1}} \boldsymbol{R_{k-1}}, 1 \le k \le 2. \end{cases}$$

Since $U_0 = 0$ and the matrix $\boldsymbol{U_k}$ is invertible for $1 \le k \le 2$, then, we have $Y_0 \ne 0$ and $\boldsymbol{Y_k} = \boldsymbol{0}$, *i.e.*, $\boldsymbol{\pi_k} = \boldsymbol{\pi_{k-1}} \boldsymbol{R_{k-1}}, 1 \le k \le 2$. Let $Y_0 = K$, we have

$$\pi_0 = K, \boldsymbol{\pi_k} = K \prod_{i=0}^{k-1} \boldsymbol{R_i}, 1 \le k \le 2, \boldsymbol{\pi_k} = \boldsymbol{\pi_2} \boldsymbol{R}^{k-2}, k \ge 3.$$

Using the normalization condition $\pi_0 + \boldsymbol{\pi_1 e_2} + \sum_{i=2}^{\infty} \boldsymbol{\pi_i e} = 1$, we have

$$K = (1 + \boldsymbol{R_0 e_2} + \boldsymbol{R_0 R_1} (\boldsymbol{I} - \boldsymbol{R})^{-1} \boldsymbol{e})^{-1}.$$

Then, the steady state probabilities can be obtained. □

The expected number of customers in the system can be obtained by

$$E[L] = \boldsymbol{\pi_1 e_2} + \sum_{k=2}^{\infty} k \boldsymbol{\pi_k e} = K \boldsymbol{R_0 e_2} + K \boldsymbol{R_0 R_1} [(\boldsymbol{I} - \boldsymbol{R})^{-1} + (\boldsymbol{I} - \boldsymbol{R})^{-2}] \boldsymbol{e}. \tag{4.5}$$

Next, an alternate method called spectral expansion method is applied to find the steady state probabilities for this queueing system. From $\boldsymbol{\pi Q} = \boldsymbol{0}$, we have

$$\boldsymbol{\pi_{i-1} A_1} + \boldsymbol{\pi_i A_0} + \boldsymbol{\pi_{i+1} A_2} = \boldsymbol{0}, i \ge 3. \tag{4.6}$$

This is a homogeneous vector difference equation of order 2, with constant coefficients. Associated with it is the characteristic matrix polynomial, $\boldsymbol{Q}(x) = \boldsymbol{A_1} + \boldsymbol{A_0} x + \boldsymbol{A_2} x^2$. Following Mitrani and Chakka [14], Chakka [15]

and Do [8], if the system is stable, the number of eigenvalues in unit cycle of the characteristic polynomial $\boldsymbol{Q}(x)$ is four, so we could write the steady state probabilities as follows:

$$\boldsymbol{\pi_i} = \sum_{k=1}^{4} a_k x_k^{i-2} \boldsymbol{\psi_k}, i \geq 2, \tag{4.7}$$

where x_k are the eigenvalues inside the unit circle, $\boldsymbol{\psi_k}$ are corresponding left eigenvectors, and coefficients a_k can be determined from the balance equations and the normalization equation. The eigenvalue-eigenvector pairs $(x_k, \boldsymbol{\psi_k})$ of $\boldsymbol{Q}(x)$ satisfy

$$\boldsymbol{\psi_k Q}(x_k) = 0, \det(\boldsymbol{Q}(x_k)) = 0, k = 1, 2, 3, 4.$$

In the present paper, $\boldsymbol{Q}(x)$ has the following the structure:

$$\boldsymbol{Q}(x) = \boldsymbol{A_1} + \boldsymbol{A_0}x + \boldsymbol{A_2}x^2 = \begin{pmatrix} f_1(x) & q\mu_1 x^2 & 0 & 0 \\ 0 & f_2(x) & p\mu_1 x^2 & q\mu_1 x^2 \\ p\mu_2 x^2 & q\mu_2 x^2 & f_3(x) & q\mu_1 x^2 \\ 0 & 0 & p\mu_2 x^2 & f_4(x) \end{pmatrix},$$

where

$$f_1(x) = \lambda - (\lambda + \mu_1)x + p\mu_1 x^2, \quad f_2(x) = \lambda - (\lambda + \mu_1)x,$$
$$f_3(x) = \lambda - (\lambda + \mu_1 + \mu_2)x + p\mu_1 x^2, \quad f_4(x) = \lambda - (\lambda + \mu_2)x + q\mu_2 x^2.$$

From the expression $\boldsymbol{Q}(x)$, we conclude that $\boldsymbol{Q}(x)$ has seven eigenvalues. Referring to Proposition 2 in [14], if the system is stable, the number of eigenvalues of $\boldsymbol{Q}(x)$ strictly inside the unit disk is equal to the degree of $\boldsymbol{Q}(x)$, *i.e.*, $\boldsymbol{Q}(x)$ should have four eigenvalues that are inside the unit circle, denoted by x_1, x_2, x_3, x_4. Then, by $\boldsymbol{\psi_k Q}(x_k) = 0$, we can derive the eigenvectors corresponding to x_1, x_2, x_3, x_4:

$$\boldsymbol{\psi_k} = (1, \psi_{k,2}, \psi_{k,3}, \psi_{k,4}), \tag{4.8}$$

where

$$\psi_{k,2} = \frac{q}{p}, \quad \psi_{k,3} = -\frac{f_1(x_k)}{p\mu_2 {x_k}^2}, \quad \psi_{k,4} = \frac{q[\lambda - (\lambda + \mu_1 + \mu_2)x]}{p[\lambda - (\lambda + \mu_2)x]}\psi_{k,3}, \quad k = 1, 2, 3, 4.$$

Once the eigenvalues and corresponding left eigenvectors are obtained, we can compute the coefficients a_i by the balance equations and the normalization condition. With the aid of the balance equations (3.3)-(3.5), the probability vectors $\pi_0, \boldsymbol{\pi_1}$ can be obtained in terms of $\boldsymbol{\pi_2}$:

$$\pi_0 = (\mu_1\pi_{1,1} + \mu_2\pi_{1,2})/\lambda, \tag{4.9}$$

$$\pi_{1,1} = \frac{(\lambda + p\mu_2)A + p\mu_2 B}{\lambda^2 + \lambda(p\mu_2 + q\mu_1)}, \tag{4.10}$$

$$\pi_{1,2} = \frac{(\lambda + q\mu_1)B + q\mu_1 A}{\lambda^2 + \lambda(p\mu_2 + q\mu_1)}, \tag{4.11}$$

where

$$A = \mu_1\pi_{2,1,1} + \mu_2\pi_{2,2,1}, \quad B = \mu_1\pi_{2,1,2} + \mu_1\pi_{2,2,1} + \mu_2\pi_{2,2,2}.$$

Substituting (4.7) into (3.6)−(3.9), and (4.7), (4.9)−(4.11) and (4.8) into (3.14) (it is not difficult to find that there are four independent linear equations), we can derive the four unknown coefficients a_1, a_2, a_3, a_4. After the coefficients are obtained, all the steady state probabilities can be derived.

By using this method, the expected number of customers in the system can be obtained by

$$E[L] = \boldsymbol{\pi_1 e_2} + \sum_{i=2}^{\infty} i\boldsymbol{\pi_i e} = \boldsymbol{\pi_1 e_2} + \sum_{k=1}^{4} a_k \left(\frac{1}{1 - x_k} + \frac{1}{(1 - x_k)^2}\right) \boldsymbol{\psi_k e}.$$

5. Some performance measures and sojourn time

From the obtained steady state probabilities, in this section, we first give some performance measures.
The probability that both servers are idle (the system is empty):

$$P_e = \pi_0.$$

The probability that server 1 is busy and server 2 is idle:

$$P_{1,b} = \pi_{1,1} + \sum_{n=2}^{\infty}(\pi_{n,1,1} + \pi_{n,1,2}).$$

The probability that server 2 is busy and server 1 is idle:

$$P_{2,b} = \pi_{1,2} + \sum_{n=2}^{\infty}\pi_{n,2,2}.$$

The probability that both servers are busy:

$$P_b = \sum_{n=2}^{\infty}\pi_{n,2,1}.$$

Next, considering a tagged customer, we derive the Laplace-Stieltjes transforms (LST) of the stationary sojourn time distribution of an arbitrary customer, where the sojourn time is the period from the epoch at which he enters the system to the epoch at which he leaves the system. Let W and $W^*(s)$ respectively denote the sojourn time of a customer and its corresponding LST. For $n \geq 2$, define $W_{n,i,j}$ and $W^*_{n,i,j}(s)$ as the conditional remaining sojourn time of a customer given that the customer sees the state $(n, i, j), i, j = 1, 2$ upon arrival and its corresponding LST. By conditioning on the next future event (first-step analysis) and using the strong Markov property, after some computations, we have

$$W^*_{n,1,1}(s) = \frac{\mu_1}{\mu_1 + s}(pW^*_{n-1,1,1}(s) + qW^*_{n-1,1,2}(s)), n \geq 3, \tag{5.1}$$

$$W^*_{n,1,2}(s) = \frac{\mu_1}{\mu_1 + s}(pW^*_{n-1,2,1}(s) + qW^*_{n-1,2,2}(s)), n \geq 3, \tag{5.2}$$

$$W^*_{n,2,1}(s) = \frac{\mu_1(pW^*_{n-1,2,1}(s) + qW^*_{n-1,2,2}(s))}{\mu_1 + \mu_2 + s} + \frac{\mu_2(pW^*_{n-1,1,1}(s) + qW^*_{n-1,1,2}(s))}{\mu_1 + \mu_2 + s}, n \geq 3, \tag{5.3}$$

$$W^*_{n,2,2}(s) = \frac{\mu_2}{\mu_2 + s}(pW^*_{n-1,2,1}(s) + qW^*_{n-1,2,2}(s)), n \geq 3, \tag{5.4}$$

with the boundary conditions

$$W^*_0(s) = \frac{\mu_1 p}{\mu_1 + s} + \frac{\mu_2 q}{\mu_2 + s},$$

$$W^*_{1,1}(s) = p\left(\frac{\mu_1}{\mu_1 + s}\right)^2 + q\frac{\mu_1\mu_2}{(\mu_1 + s)(\mu_2 + s)},$$

$$W^*_{1,2}(s) = p\frac{\mu_1}{\mu_1+s} + q\left(\frac{\mu_2}{\mu_2+s}\right)^2,$$

$$W^*_{2,1,1}(s) = p\left(\frac{\mu_1}{\mu_1+s}\right)^3 + q\left(\frac{\mu_1}{\mu_1+s}\right)^2\frac{\mu_2}{\mu_2+s},$$

$$W^*_{2,1,2}(s) = p\left(\frac{\mu_1}{\mu_1+s}\right)^2 + q\frac{\mu_1}{\mu_1+s}\left(\frac{\mu_2}{\mu_2+s}\right)^2,$$

$$W^*_{2,2,1}(s) = p\left(\frac{\mu_1}{\mu_1+\mu_2+s}\frac{\mu_1}{\mu_1+s} + \frac{\mu_2}{\mu_1+\mu_2+s}\left(\frac{\mu_1}{\mu_1+s}\right)^2\right)$$
$$+ q\left(\frac{\mu_2}{\mu_1+\mu_2+s}\frac{\mu_1\mu_2}{(\mu_1+s)(\mu_2+s)} + \frac{\mu_1}{\mu_1+\mu_2+s}\left(\frac{\mu_2}{\mu_2+s}\right)^2\right),$$

$$W^*_{2,2,2}(s) = p\frac{\mu_1\mu_2}{(\mu_1+s)(\mu_2+s)} + q\left(\frac{\mu_2}{\mu_2+s}\right)^3,$$

where $W_0(s)$, $W_{1,1}$ and $W_{1,2}$ denote that the conditional remaining sojourn time of a customer given that the customer sees the state $\{0\}$, $\{(1,1)\}$ and$\{(1,2)\}$ upon arrival. Define

$$H_j(s) = a_j x_j^{-2}(\psi_{j,1}M^j_{1,1}(s) + \psi_{j,2}M^j_{1,2}(s) + \psi_{j,3}M^j_{2,1}(s) + \psi_{j,4}M^j_{2,2}(s)), j = 1,2,3,4,$$

where

$$M^j_{1,1}(s) = \sum_{n=2}^{\infty} W^*_{n,1,1}(s)x_j^n, M^j_{1,2}(s) = \sum_{n=2}^{\infty} W^*_{n,1,2}(s)x_j^n,$$

$$M^j_{2,1}(s) = \sum_{n=2}^{\infty} W^*_{n,2,1}(s)x_j^n, M^j_{2,2}(s) = \sum_{n=2}^{\infty} W^*_{n,2,2}(s)x_j^n.$$

Multiplying the equations by x_j^n and summing over n from 3 to ∞, we have

$$(\mu_1+s)[M^j_{1,1}(s) - W^*_{2,1,1}(s)x_j^2] = \mu_1[px_jM^j_{1,1}(s) + qx_jM^j_{1,2}(s)],$$

$$(\mu_1+s)[M^j_{1,2}(s) - W^*_{2,1,2}(s)x_j^2] = \mu_1[px_jM^j_{2,1}(s) + qx_jM^j_{2,2}(s)],$$

$$(\mu_1+\mu_2+s)[M^j_{2,1}(s) - W^*_{2,2,1}(s)x_j^2] = \mu_1[px_jM^j_{2,1}(s) + qx_jM^j_{2,2}(s)]$$
$$+ \mu_2[px_jM^j_{1,1}(s) + qx_jM^j_{1,2}(s)],$$

$$(\mu_2+s)[M^j_{2,2}(s) - W^*_{2,2,2}(s)x_j^2] = \mu_2[px_jM^j_{2,1}(s) + qx_jM^j_{2,2}(s)].$$

Multiplying above equations with $a_j x_j^{-2}\psi_{j,1}$, $a_j x_j^{-2}\psi_{j,2}$, $a_j x_j^{-2}\psi_{j,3}$ and $a_j x_j^{-2}\psi_{j,4}$, summing over them yields

$$\begin{aligned}
&sH_j(s) - a_j[(\mu_1+s)\psi_{j,1}W^*_{2,1,1}(s) + (\mu_1+s)\psi_{j,2}W^*_{2,1,2}(s)\\
&+ (\mu_1+\mu_2+s)\psi_{j,3}W^*_{2,2,1}(s) + (\mu_2+s)\psi_{j,4}W^*_{2,2,2}(s)]\\
&= a_j x_j^{-2}M^j_{1,1}(s)[-\mu_1\psi_{j,1} + \mu_1 p x_j\psi_{j,1} + \mu_2 p x_j\psi_{j,3}]\\
&+ a_j x_j^{-2}M^j_{1,2}(s)[-\mu_1\psi_{j,2} + \mu_1 q x_j\psi_{j,1} + \mu_2 q x_j\psi_{j,3}]\\
&+ a_j x_j^{-2}M^j_{2,1}(s)[-(\mu_1+\mu_2)\psi_{j,3} + \mu_1 p x_j\psi_{j,3} + \mu_1 p x_j\psi_{j,2} + \mu_2 p x_j\psi_{j,4}]\\
&+ a_j x_j^{-2}M^j_{2,2}(s)[-\mu_2\psi_{j,4} + \mu_2 q x_j\psi_{j,4} + \mu_1 q x_j\psi_{j,2} + \mu_1 q x_j\psi_{j,3}]. \qquad (5.5)
\end{aligned}$$

According to the balance equations for $n \geq 3$, equation (5.5) can be simplified as

$$\begin{aligned}
&sH_j(s) - a_j[(\mu_1+s)\psi_{j,1}W^*_{2,1,1}(s) + (\mu_1+s)\psi_{j,2}W^*_{2,1,2}(s)\\
&+ (\mu_1+\mu_2+s)\psi_{j,3}W^*_{2,2,1}(s) + (\mu_2+s)\psi_{j,4}W^*_{2,2,2}(s)]\\
&= a_j x_j^{-3}M^j_{1,1}(s)[\lambda(x_j-1)]\psi_{j,1} + a_j x_j^{-3}M^j_{1,2}(s)[\lambda(x_j-1)]\psi_{j,2}\\
&+ a_j x_j^{-3}M^j_{2,1}(s)[\lambda(x_j-1)]\psi_{j,3} + a_j x_j^{-3}M^j_{2,2}(s)[\lambda(x_j-1)]\psi_{j,4},
\end{aligned}$$

then, $H_j(s)$ can be obtained by

$$\begin{aligned}
H_j(s) =& \frac{a_j}{s-\lambda(1-x_j^{-1})}[(\mu_1+s)\psi_{j,1}W^*_{2,1,1}(s) + (\mu_1+s)\psi_{j,2}W^*_{2,1,2}(s)\\
&+ (\mu_1+\mu_2+s)\psi_{j,3}W^*_{2,2,1}(s) + (\mu_2+s)\psi_{j,4}W^*_{2,2,2}(s)].
\end{aligned}$$

Finally, the LST of an arbitrary customer's sojourn time can be obtained by

$$\begin{aligned}
W^*(s) =& \pi_0 W^*_0(s) + \pi_{1,1}W^*_{1,1}(s) + \pi_{1,2}W^*_{1,2}(s)\\
&+ \sum_{n=2}^{\infty}[\pi_{n,1,1}W^*_{n,1,1}(s) + \pi_{n,1,2}W^*_{n,1,2}(s) + \pi_{n,2,1}W^*_{n,2,1}(s) + \pi_{n,2,2}W^*_{n,2,2}(s)]\\
=& \pi_0 W^*_0(s) + \pi_{1,1}W^*_{1,1}(s) + \pi_{1,2}W^*_{1,2}(s) + \sum_{j=1}^{4} H_j(s).
\end{aligned}$$

6. Numerical examples

In this section, we provide a set of numerical examples to show the effect of system parameters on π_0 and $E[L]$, and give comparisons between Case 1 and Case 2. First, we give the state-transition-rate matrix of Case 2.

$$\boldsymbol{Q}^* = \begin{pmatrix}
\overline{B}_0^* & \overline{\boldsymbol{B}}_{\boldsymbol{1}}^* & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \cdots\\
\overline{\boldsymbol{B}}_{\boldsymbol{2}}^* & \overline{\boldsymbol{A}}_{\boldsymbol{0}}^* & \overline{\boldsymbol{A}}_{\boldsymbol{1}}^* & \boldsymbol{0} & \boldsymbol{0} & \cdots\\
\boldsymbol{0} & \overline{\boldsymbol{A}}_{\boldsymbol{2}}^* & \boldsymbol{A}_{\boldsymbol{0}}^* & \boldsymbol{A}_{\boldsymbol{1}}^* & \boldsymbol{0} & \cdots\\
\boldsymbol{0} & \boldsymbol{0} & \boldsymbol{A}_{\boldsymbol{2}}^* & \boldsymbol{A}_{\boldsymbol{0}}^* & \boldsymbol{A}_{\boldsymbol{1}}^* & \cdots\\
\boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{A}_{\boldsymbol{2}}^* & \boldsymbol{A}_{\boldsymbol{0}}^* & \cdots\\
\vdots & \vdots & \vdots & \vdots & \vdots & \ddots
\end{pmatrix},$$

TABLE 1. A comparison of two method in computing the steady state probabilities.

$\boldsymbol{\pi_k}$	RG-factorization	Spectral Expansion Method
π_0	0.445417	0.445367
$\pi_{1,1}$	0.140959	0.140945
$\pi_{1,2}$	0.132493	0.132476
$\pi_{2,1,1}$	0.043321	0.043322
$\pi_{2,1,2}$	0.028880	0.028883
$\pi_{2,2,1}$	0.029717	0.029709
$\pi_{2,2,2}$	0.033679	0.033671
$\pi_{3,1,1}$	0.021965	0.021997
$\pi_{3,1,2}$	0.014643	0.014665
$\pi_{3,2,1}$	0.012770	0.012749
$\pi_{3,2,2}$	0.020020	0.019994
$\vdots$	$\vdots$	$\vdots$

where

$$\overline{B}_0^* = \overline{B}_0 = -\lambda, \overline{\boldsymbol{B}}_{\boldsymbol{1}}^* = \overline{\boldsymbol{B}}_{\boldsymbol{1}} = (\lambda p, \lambda q), \overline{\boldsymbol{B}}_{\boldsymbol{2}}^* = \overline{\boldsymbol{B}}_{\boldsymbol{2}} = (\mu_1, \mu_2)^T,$$

$$\overline{\boldsymbol{A}}_{\boldsymbol{0}}^* = \begin{pmatrix} -(\lambda+\mu_1) & 0 \\ 0 & -(\lambda+\mu_2) \end{pmatrix}, \overline{\boldsymbol{A}}_{\boldsymbol{1}}^* = \begin{pmatrix} \lambda p & \lambda q & 0 & 0 \\ 0 & 0 & \lambda p & \lambda q \end{pmatrix}, \overline{\boldsymbol{A}}_{\boldsymbol{2}}^* = \begin{pmatrix} \mu_1 & 0 \\ \mu_2 & \mu_1 \\ \mu_2 & 0 \\ 0 & \mu_2 \end{pmatrix},$$

$$\boldsymbol{A}_{\boldsymbol{0}}^* = \begin{pmatrix} -(\lambda+\mu_1) & 0 & 0 & 0 \\ 0 & -(\lambda+\mu_1+\mu_2) & 0 & 0 \\ 0 & 0 & -(\lambda+\mu_2) & 0 \\ 0 & 0 & 0 & -(\lambda+\mu_2) \end{pmatrix},$$

$$\boldsymbol{A}_{\boldsymbol{1}}^* = \boldsymbol{A}_{\boldsymbol{1}} = \begin{pmatrix} \lambda & 0 & 0 & 0 \\ 0 & \lambda & 0 & 0 \\ 0 & 0 & \lambda & 0 \\ 0 & 0 & 0 & \lambda \end{pmatrix}, \boldsymbol{A}_{\boldsymbol{2}}^* = \begin{pmatrix} p\mu_1 & q\mu_1 & 0 & 0 \\ p\mu_2 & q\mu_2 & p\mu_1 & q\mu_1 \\ p\mu_2 & q\mu_2 & 0 & 0 \\ 0 & 0 & p\mu_2 & q\mu_2 \end{pmatrix}.$$

Corollary 6.1. *In this case, the sufficient and necessary stability condition is*

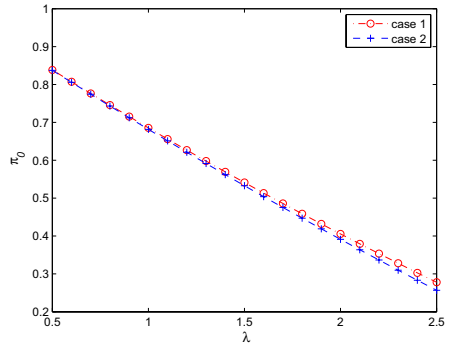
$$\lambda < \frac{\mu_1^2\mu_2 + p\mu_1\mu_2^2}{p\mu_1\mu_2 + p^2\mu_2^2 + q\mu_1^2}.$$

Proof. The proof of this corollary is similar to the proof of Theorem 3.1, and we do not explain it here any more. □

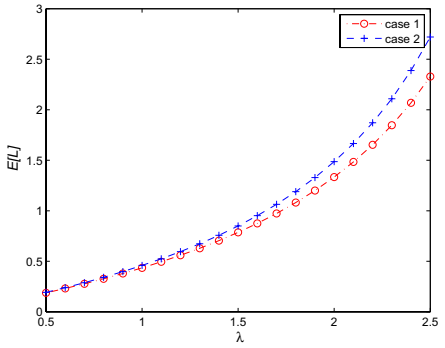
Similar to the analysis in Case 1, if the stability condition is satisfied, we can obtain the steady state probabilities and sojourn time in Case 2.

Now, we first provide a table (see Tab. 1) to show the stationary probabilities π_0 and $\boldsymbol{\pi_k}, k \geq 1$ obtained by using UL-type RG-factorization and spectral expansion method, where $\lambda = 2$, $\mu_1 = 3.5$, $\mu_2 = 3$ and $p = 0.6$. By using the two method to calculate the steady state probabilities $\boldsymbol{\pi_k}$, we find that spectral expansion method offers considerable advantages in efficiency, and it has a faster speed than RG-factorization in computing the stationary probabilities.

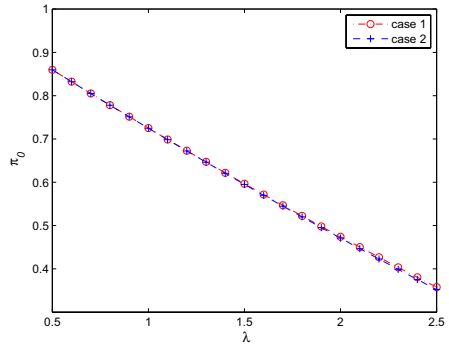
Next, based on the theoretical framework given by the above analysis, we present some figures below to study the impact of the parameters on π_0 and $E[L]$ under stability condition. Without loss of generality, we



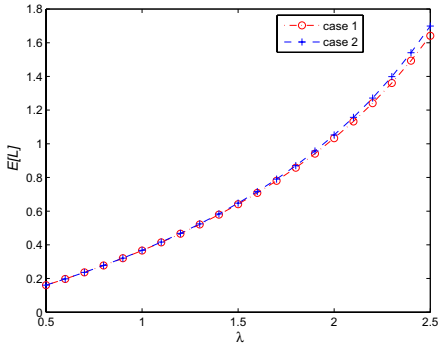
(a) π_0 *vs.* $\lambda(\mu_1 = 3.5, \mu_2 = 2.5, p = 0.6)$



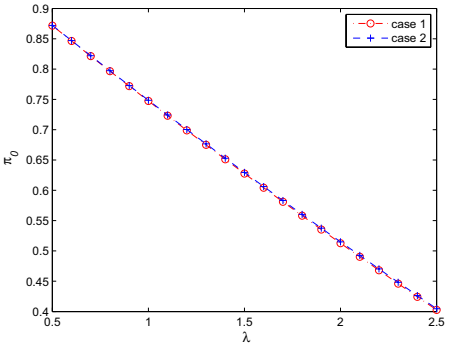
(b) E[L] *vs.* $\lambda(\mu_1 = 3.5, \mu_2 = 2.5, p = 0.6)$



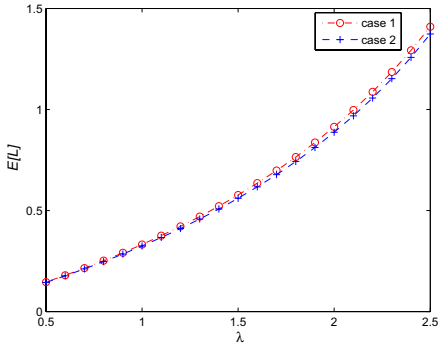
(c) π_0 *vs.* $\lambda(\mu_1 = 3.5, \mu_2 = 3.5, p = 0.6)$



(d) E[L] *vs.* $\lambda(\mu_1 = 3.5, \mu_2 = 3.5, p = 0.6)$



(e) π_0 *vs.* $\lambda(\mu_1 = 3.5, \mu_2 = 4.5, p = 0.6)$

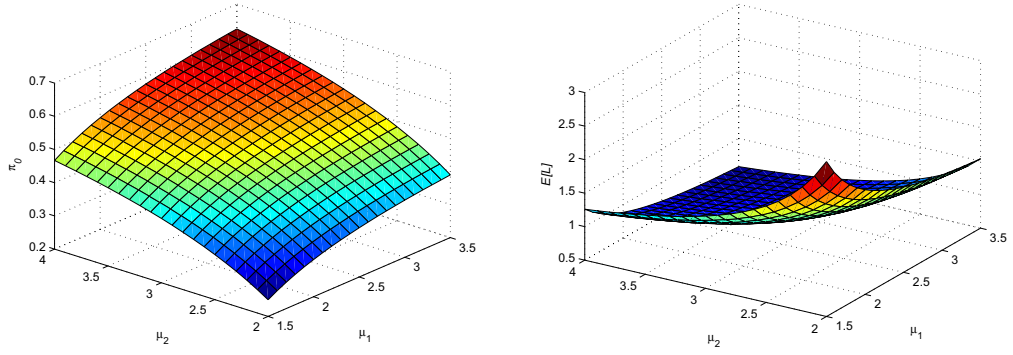


(f) E[L] *vs.* $\lambda(\mu_1 = 3.5, \mu_2 = 4.5, p = 0.6)$

FIGURE 2. π_0 and $E[L]$ *versus* λ for different cases.

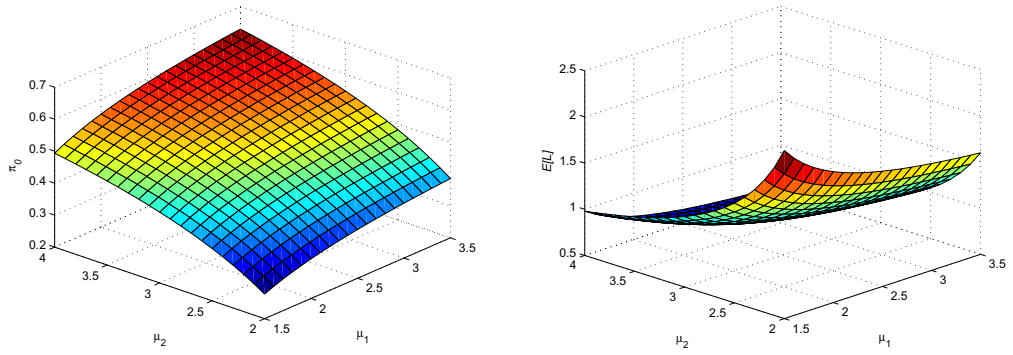
first assume $\mu_1 = 3.5, p = 0.6$, $\mu_2 = 2.5,\ 3.5,\ 4.5$, and plot the trend of the change for π_0 and $E[L]$ as arrival rate λ increases from 0.5 to 1.5.

Clearly, from Figure 2, we find that π_0 decreases with the increase of λ and $E[L]$ increases with the increase of λ, which are identical to the intuitive expectations. Actually, higher arrival rate λ leads to more customers staying in the system. From Figures 2a and 2d, we also find that if λ is fixed, Case 1 has a bigger value π_0 and a smaller value $E[L]$ than Case 2, *i.e.*, under this assumption, Case 1 has a greater ability to reduce congestion.



(a) π_0 *vs.* μ_1 and $\mu_2(\lambda = 1.5, p = 0.3)$ (b) E[L] *vs.* μ_1 and $\mu_2(\lambda = 1.5, p = 0.3)$

FIGURE 3. π_0 and $E[L]$ *versus* μ_1 and μ_2 for Case 1.



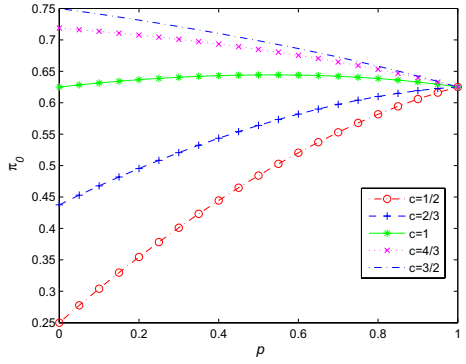
(a) π_0 *vs.* μ_1 and $\mu_2(\lambda = 1.5, p = 0.3)$ (b) E[L] *vs.* μ_1 and $\mu_2(\lambda = 1.5, p = 0.3)$

FIGURE 4. π_0 and $E[L]$ *versus* μ_1 and μ_2 for Case 2.

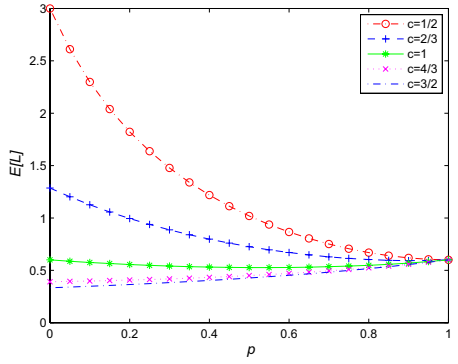
From Figures 2e and 2f, we find that if λ is fixed, Case 2 has a bigger value π_0 and a smaller value $E[L]$ than Case 1.

In Figures 3 and 4, we assume $\lambda = 1.5, p = 0.3$, and investigate the values π_0 and $E[L]$ regarding the combinations of the values μ_1 and μ_2 under the two cases, respectively. As expected, from Figure 3, for Case 1, π_0 increases as μ_1 increases and increases as μ_2 increases. Conversely $E[L]$ decreases with the increase of μ_1 and decreases with the increase of μ_2. From Figure 4, for Case 2, we find that π_0 and $E[L]$ have the same variation trend as that for Case 1.

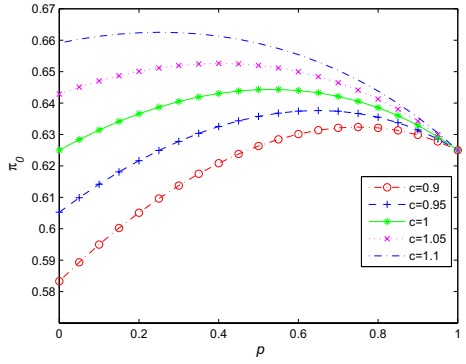
In Figure 5, we pay attention to the curves of π_0 and $E[L]$ with the change of p with $\lambda = 1.5$, $\mu_1 = 4$ and $\mu_2 = c\mu_1$. From Figure 5a, we find that, For $c = 3/2, 4/3$, *i.e.*, $\mu_2 > \mu_1$, π_0 decreases with the increase of p. For $c = 1$, *i.e.*, $\mu_1 = \mu_2$, π_0 increases with the increase of p from 0 to 0.5, and decreases as p increases from 0.5 to 1. For $c = 2/3, 1/2$, *i.e.*, $\mu_1 > \mu_2$, π_0 increases with the increase of p. It is also obvious that, if p fixed, the larger c is, *i.e.*, the bigger μ_2 is, the larger π_0 becomes. We also find that, as p approaches to 1, π_0 tends to a fix value no matter the values c. It is reasonable that when p reaches to 1, the queue reduces to a classic $M/M/1$ queue with arrival rate λ and service rate μ_1, and μ_2 has no impact on π_0. From Figure 5b, we find that $E[L]$ has the opposite variation trend to π_0.



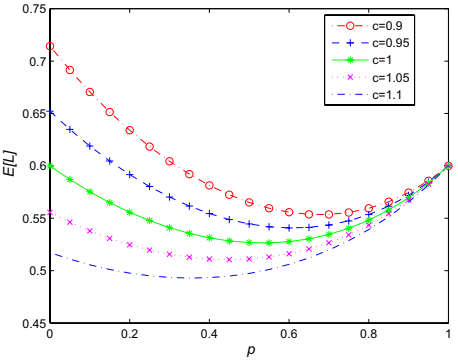
(a) π_0 *vs.* p $(\lambda = 1.5, \mu_1 = 4, \mu_2 = c\mu_1)$



(b) E[L] *vs.* $p(\lambda = 1.5, \mu_1 = 4, \mu_2 = c\mu_1)$



(c) π_0 *vs.* p $(\lambda = 1.5, \mu_1 = 4, \mu_2 = c\mu_1)$

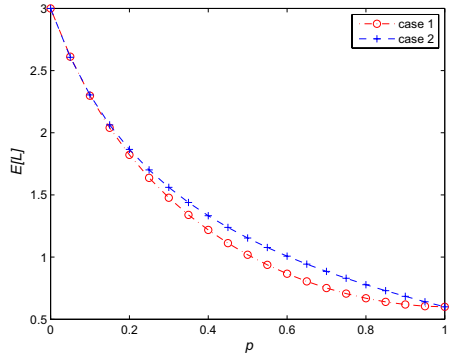


(d) E[L] *vs.* p $(\lambda = 1.5, \mu_1 = 4, \mu_2 = c\mu_1)$

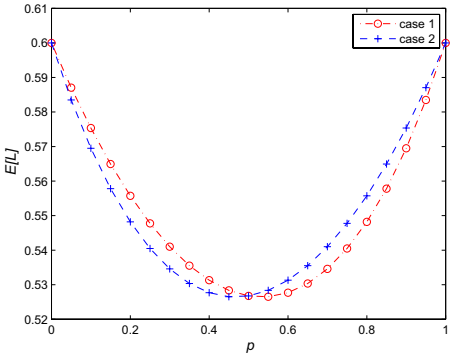
FIGURE 5. π_0 and $E[L]$ *versus* p for Case 1 of different values c.

In order to further explore the impact of value c on $E[L]$ and π_0, we assume $c = 0.9,\ 0.95,\ 1,\ 1.05,\ 1.1$ and investigate the variation trend of π_0 and $E[L]$ with the change of p. As expected, from Figures 5c and 5d, we find that π_0 has a maximum for p that is not in the extreme (1 or 0) and $E[L]$ has a minimum for p that is not in the extreme (0 or 1). Hence, we can make a guess that if $c \in [1-\delta, 1+\delta]$, $E[L]$ has a minimum for p that lies in the interval $(0, 1)$, if $c < 1-\delta$, $E[L]$ has a minimum for p that is in the extreme 1, and if $c > 1+\delta$, $E[L]$ has a minimum for p that is in the extreme 0, where δ is a given relatively small constant. In my opinion, this may be because, when $c \in [1-\delta, 1+\delta]$, *i.e.*, μ_2 is approximately equal to μ_1, as p increases, server 1 can share the workload and serve type 1 customers, which leads to the reduction of the number of customers. As p continues to increase, more and more customers belong to type 1, whether server 2 is also active or not, is not very relevant, that is, at this moment, server 2 plays a relatively minor role in reducing the number of customers. So, if $c \in [1-\delta, 1+\delta]$, $E[L]$ has a minimum for p that lies in the interval $(0, 1)$. When $c > 1+\delta$, *i.e.*, μ_2 is larger than μ_1, when $p = 0$, server 2 plays a dominant role in reducing the number of customers, as p increases, the role of server 2 is weaken and the role of server 1 is enhanced, which leads to more customers stays in system, so $E[L]$ has a minimum for p at 0. Similarly, when $c < 1-\delta$, that is sever 1 has a faster service rate than server 2, so, as p increases, the number of type 1 customers increases and server 1 plays a dominant role in reducing the number of customer, which leads to more customers leaving the system, so $E[L]$ has a minimum for p at 1.

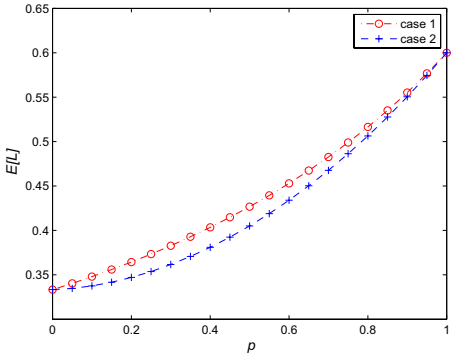
Then, we assume $\lambda = 1.5$, $\mu_1 = 4$ and $\mu_2 = c\mu_1$ and present a comparison between the two cases. In Figure 6, we plot the trend of the change for $E[L]$ as p from 0 to 1. From Figure 6a, if $c = 1/2$, we find that $E[L]$ decreases



(a) E[L] *vs.* p ($\lambda = 1.5, \mu_1 = 4, \mu_2 = \frac{1}{2}\mu_1$)



(b) E[L] *vs.* p ($\lambda = 1.5, \mu_1 = 4, \mu_2 = \mu_1$)

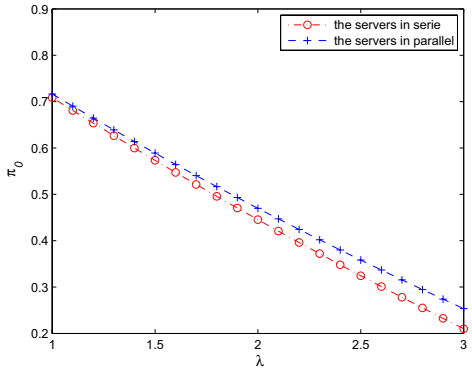


(c) E[L] *vs.* p ($\lambda = 1.5, \mu_1 = 4, \mu_2 = \frac{3}{2}\mu_1$)

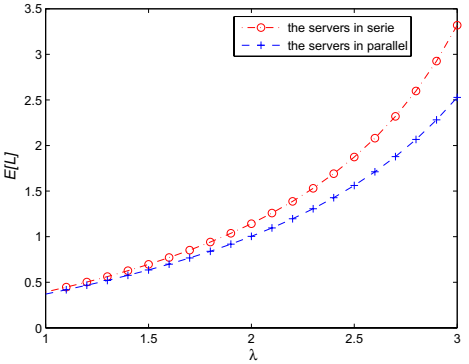
FIGURE 6. $E[L]$ *versus* p of different cases for different values c.

with the increase of p. From Figure 6b, if $c = 1$, *i.e.*, $\mu_1 = \mu_2$, we find that $E[L]$ first decreases with the increase of p and then increases with the increase of p. From Figure 6c, if $c = 3/2$, an increase in p results in the increase of $E[L]$. We also find that, from Figure 6a, if p is fixed, Case 2 has a bigger value than Case 1, *i.e.*, Case 1 has a greater ability to reduce congestion. From Figure 6b, we find that when p increases from 0 to 0.5 and for fixed p in this interval, Case 1 has a bigger value than Case 2, when p increases from 0.5 to 1 and for fixed p in $[0.5, 1]$, Case 2 has a bigger value than Case 1. From Figure 6c, if p is fixed, Case 1 has a bigger value than Case 2. It is worth noting that, as p approaches to 0 or 1, for both cases, the values $E[L]$ are equal. It is because that, as p reaches to 0, the queues of both cases reduce to a classic $M/M/1$ queue with arrival rate λ and service rate μ_2. Similarly, as p reaches to 1, the queues of both cases reduce to a classic $M/M/1$ queue with arrival rate λ and service rate μ_1.

Finally, we provide a comparison between the queueing model in consideration and the queueing model in [4] and show what the exact influence of the servers in series is. Assume that $\mu_1 = 3.5$, $\mu_2 = 3$, and $p = 0.6$. From Figures 7a and 7b, as λ increases, the values $E[L]$ of the queueing model that the servers in parallel and the queueing model that the servers in series have the same variation trend. We also find that, for a fixed λ, the value $E[L]$ of the queueing model in consideration is larger than the queueing model in [4], moreover, the larger λ is, the greater difference between the two models becomes, that is, the block effect caused by the two tandem servers can make a main influence on the number of customers in the system especially when λ is large.



(a) π_0 *vs.* λ for the servers in parallel and in series



(b) E[L] *vs.* λ for the servers in parallel and in series

FIGURE 7. π_0 and $E[L]$ *versus* λ for the servers in parallel and in series.

The numerical examples indicates that if we want to reduce congestion and improve the service ability, we need to choose a suitable case on the basis of the customer types. We hope the results can be applied to more practical queueing systems.

7. Conclusion

In this paper, we investigated a two-class continuous-time queueing model with two tandem dedicated servers. We established the theoretical foundations for applications and obtained the explicit computation expressions for the performance measures. By using mean drift result, we first gave the stability condition for the system. Then, in terms of matrix analytic method and spectral expansion method, we respectively obtained the steady state probabilities. Further, we provided the elaborate analysis of the stationary sojourn time of an arbitrary customer. Finally, we presented some numerical examples to show the impact of parameters on the performance measures and gave a comparison between the two cases. We expect that the results can be applied to more practical queueing systems.

Appendix A.

First, we introduce the average amount of work (of type 1 and type 2) that enters the system per unit time:

$$\rho^* = \rho_1 + \rho_2 = \frac{\lambda p}{\mu_1} + \frac{\lambda q}{\mu_2}.$$

So, the stability condition can be expressed as

$$\rho^* < t_1 + t_2 + 2t,$$

where t_1 denotes the fraction of time when only server 1 is busy, t_2 denotes the fraction of time when only server 2 is busy, t is the fraction of time both the two servers are busy. Assuming the system is continuously provided with new arrivals and there are always at least two customers in the system. The system is stable if $\rho^* < t_1 + t_2 + 2t$ is satisfied. Since [4] has given a detail explanation on this expression, we don't explain it any more. In order to derive the fractions of time t_0, t_1 and t_2, we note that the busy servers form a simple

four-state Markov chain with the state space {(1,1), (1,2), (2,1), (2,2)}. Then we have

$$t_{1,1} = \frac{\mu_2^2 p^2}{\mu_2^2 p + \mu_1\mu_2 q + \mu_1^2 q^2}, \qquad t_{1,2} = \frac{\mu_2^2 pq}{\mu_2^2 p + \mu_1\mu_2 q + \mu_1^2 q^2},$$

$$t_{2,1} = \frac{\mu_1\mu_2 pq}{\mu_2^2 p + \mu_1\mu_2 q + \mu_1^2 q^2}, \qquad t_{2,2} = \frac{\mu_1 q^2(\mu_1 + \mu_2)}{\mu_2^2 p + \mu_1\mu_2 q + \mu_1^2 q^2},$$

where $t_1 = t_{1,1} + t_{1,2}$, $t_{2,1} = t$, $t_{2,2} = t_2$. Then after some algebraic computations, the expression $\rho^* < t_1 + t_2 + 2t$ translates into

$$\lambda < \frac{\mu_1{\mu_2}^2 + q{\mu_1}^2\mu_2}{p{\mu_2}^2 + q\mu_1\mu_2 + q^2{\mu_1}^2},$$

which is the stability condition of the system, which is accordance with Theorem 3.1.

Acknowledgements. The authors would like to thank the editor and the referees for the helpful suggestions and comments to improve the quality of this paper.

References

[1] H. Bruneel, W. Mélange, B. Steyaert, D. Claeys and J. Walraevens, A two-class discrete-time queueing model with two dedicated servers and global FCFS service discipline. *Eur. J. Oper. Res.* **223** (2012) 123–132.

[2] H. Bruneel, W. Mélange, B. Steyaert, D. Claeys and J. Walraevens, Effect of global FCFS and relative load distribution in two-class queues with dedicated servers. *4OR Q. J. Oper. Res.* **11**(4) (2013) 375–391.

[3] W. Mélange, H. Bruneel, B. Steyaert, D. Claeys and J. Walraevens, A continuous-time queueing model with class clustering and global FCFS service discipline. *J. Ind. Manag. Optimiz.* **10** (2014) 193–206.

[4] W. Mélange, J. Walraevens, D. Claeys, B. Steyaert and H. Bruneel, The impact of a global FCFS service discipline in a two-class queue with dedicated servers. *Comput. Oper. Res.* **71** (2016) 23–33.

[5] H. Bruneel, W. Mélange, D. Claeys and J. Walraevens, A two-class global FCFS discrete-time queueing model with arbitrary-length constant service times. *TOP* **25** (2017) 164–178.

[6] Q.M. He and X. Chao, A tollbooth tandem queue with heterogeneous servers. *Eur. J. Oper. Res.* **236** (2014) 177–189.

[7] X. Chao, Q.M. He and S. Ross, Tollbooth tandem queues with infinite homogeneous servers. *J. Appl. Prob.* **52** (2015) 941–961.

[8] T.V. Do, A closed-form solution for a toll booth tandem queue with two heterogeneous servers and exponential service times. *Eur. J. Oper. Res.* **247** (2015) 672–675.

[9] M.F. Neuts, Matrix-Geometric Solutions in Stochastic Models: Algorithmic Approach. Johns Hopkins University Press, Baltimore (1981).

[10] T. Phung−Duc, H. Masuyama, S. Kasahara and Y. Takahashi, A simple algorithm for the rate matrices of level-dependent QBD processes, in *Proc. of the 5th International Conference on Queueing Theory and Network Applications,* Beijing (2010) 46–52.

[11] D.A. Bini, B. Meini, S. Steffe and B. Van Houdt, Structured Markov chain solver: The algorithms, in *Proc. of the SMCTOOLS workshop,* Pisa (2006).

[12] G. Latouche and V. Ramaswami, Introduction to matrix analytic methods in stochastic modeling. SIAM, Philadelphia (1999).

[13] Q.L. Li, Constructive Computation in Stochastic Models with Applications: the RG-Factorizations. Springer, Berlin and Tsinghua University Press, Beijing (2010).

[14] I. Mitrani and R. Chakka, Spectral expansion solution for a class of Markov models: application and comparison with the matrix-geometric method. *Perform. Eval.* **23** (1995) 241–260.

[15] R. Chakka, *Performance and reliability modelling of computing systems using spectral expansion.* Ph.D. Thesis, University of Newcastle upon Tyne, Newcastle upon Tyne (1995).