Computers and Operations Research 114 (2020) 104809

Contents lists available at ScienceDirect

Computers and Operations Research

journal homepage: www.elsevier.com/locate/cor

A polling system with ‘Join the shortest - serve the longest’ policy

Efrat Perel[a], Nir Perel[a,*], Uri Yechiali[b]

[a] School of Industrial Engineering and Management, Afeka College of Engineering, Tel-Aviv, Israel
[b] Department of Statistics and Operations Research, School of Mathematical Sciences, Tel-Aviv University, Tel-Aviv, Israel

ARTICLE INFO

Article history:
Received 25 February 2019
Revised 23 August 2019
Accepted 26 September 2019
Available online 28 September 2019

2010 MSC:
60K25
90B22

Keywords:
Polling systems
Join the shortest
Serve the longest
Probability generating functions
Matrix geometric

ABSTRACT

This paper studies a Markovian single-server non-symmetric two-queue polling system, operating simultaneously under a combination of two well-known queueing regimes: (*i*) ‘Join the Shortest Queue’ and (*ii*) ‘Serve the Longest Queue’. The system is defined as a two-dimensional continuous-time Markov chain, and analyzed via both probability generating functions approach and matrix geometric method. Although both queues are unbounded, by applying a non-conventional representation and without resorting to involved boundary-value problem analysis, we derive the joint steady-state probability distribution of the system's states, and consequently calculate its performance measures and derive its stability condition. Numerical results are presented, as well as a comparison with a corresponding $M/G/1$ queue.

1. Introduction

Two queueing models that have been extensively studied in the literature are (*i*) the so called ‘Join the Shortest Queue’ (JSQ), involving a single arrival stream of customers and multiple queues, in which a new arrival joins the shortest queue; and (*ii*) the so called ‘Serve the Longest Queue’ (SLQ), in which a single server attends several queues, and always chooses the next customer to be served from the longest queue. Each model aims at balancing the queue lengths: Under the JSQ regime, arriving customers are the decision makers, while in the SLQ policy the server is the controller.

This paper combines the above two operating policies into a unified model. Specifically, we consider a polling system comprised of two non-identical Markovian queues, denoted by Q_1 and Q_2, attended by a single server that alternates between them. An arriving customer always joins the shortest queue, unless the queue lengths are equal, in which case the customer joins Q_i $(i = 1, 2)$ w.p. p_i $(p_1 + p_2 = 1)$. The server always serves the longest queue while exercising a queue-size depending preemptive priority policy, i.e. the server never resides in a shorter queue, giving priority to the longest queue. This preemptive-type policy implies that at a moment when the number of customers in an un-served queue exceeds the number of customers in the served queue (either at service completion or when arrival occurs), the server immediately switches to the longer queue. This is in contrast to the classical multi-class priority model where the priority level of each class is pre-determined and does not change. In the present model, the priority levels (of the queues) are size-dependent and dynamically change according to queue lengths. As a possible illustration of our JSQ-SLQ model, one may consider a medical clinic with several nurse rooms and a single physician. A newly arriving patient is directed to a nurse room having the shortest queue and is treated there (e.g. her blood is sampled). When the results are obtained, the single physician attends the patient, always choosing a patient from the longest queue. Naturally, the service rates of the nurses are not identical.

Single-server polling systems have been widely studied in the queueing literature, see e.g. Takagi (1986), Yechiali (1993), Boon et al. (2011), and the extensive references therein. The main service disciplines applied by the server are the Exhaustive, Gated, Globally-Gated and k-limited. In most cases, the server visits the queues in a cyclic (Round-Robin) order, incurring non-zero switch-over times when switching between queues. Server's dynamic switching rules were also investigated, see e.g. Browne and Yechiali (1989). Single-server two-queue polling models with switching decisions depending on the queue sizes, but with zero switch-over times were studied by Perel and Yechiali (2017), while

[*] Corresponding author.
E-mail addresses: efratp@afeka.ac.il (E. Perel), nirp@afeka.ac.il (N. Perel), uriy@tauex.tau.ac.il (U. Yechiali).

https://doi.org/10.1016/j.cor.2019.104809

a corresponding model with non-zero switching times was investigated by Jolles et al. (2018).

'Serving the longer queue' (SLQ) regime was introduced by Cohen (1987), who studied a system with two queues, where a single server serves the longer queue under non-preemptive priority policy. Each queue has its own generally distributed service time and its own Poisson rate of arrival. Flatto (1989) considered a Markovian system with two identical queues and a single server that serves the longer queue under preemptive priority policy. Zheng and Zipkin (1990) considered the SLQ policy in the context of inventory control. Houtum et al. (1997) studied a fully symmetric non-preemptive Markovian system with a single server and N queues, where upon service completion, the server picks the next job from the longest queue. Knessl and Yao (2013) provided asymptotic properties of heavy traffic limits for a Markovian non-symmetric two-queue model under the SLQ policy. Baharian and Tezcan (2011) considered the SLQ mechanism and studied the stability of a system with parallel queues and different classes, both of customers and servers; Ravid et al. (2013) analyzed a Markovian repair system with a single repairman and two queues having non identical arrival rates. They obtained expressions for queue lengths and sojourn times and pointed out a direct relation between their model and a corresponding SLQ model; Maguluri et al. (2014) investigated the SLQ mechanism in the context of scheduling in wireless networks, while Pedarsani and Walrand (2016) studied the stability of SLQ scheduling in open multi-class queueing networks.

While the SLQ policy concentrates on the server's selection policy of jobs to be processed, the JSQ mechanism deals with customers' decisions upon arrival. Winston (1977) considered a fully symmetric Markovian system with multiple servers, each having its own queue, where Poisson arriving customers join the shortest queue. It is shown that the JSQ policy is optimal in the sense that it maximizes the discounted number of customers to complete their service in any time t. Adan et al. (1991a) studied a system with two parallel queues, each having its specific exponential service time and a single stream of Poisson arrivals, where customers follow the JSQ policy. It is shown that the joint equilibrium distribution of the queue lengths can be represented by an infinite sum of product-form solutions. In a following paper, Adan et al. (1991b) studied the same system, while allowing jockeying between the queues whenever the difference between the queue lengths exceeds some threshold T. Adan et al. (2013) further analyzed a system with two single server queues, where customers inter-arrival times follow an Erlang distribution, an arriving customer joins the shortest queue, and service times are exponentially distributed. Furthermore, Adan et al. (2016) considered a Markovian polling system with two symmetric queues with a single server operating according the exhaustive switching regime, i.e., the server stays at the current queue if the system is completely empty after a service completion. The latter authors derived the equilibrium distribution of the joint queue lengths by using the compensation approach and by defining and solving a boundary value problem. Additional studies on JSQ policy can be found in Halfin (1985), Menich (1987), Hordijk and Koole (1990), Menich and Serfozo (1991), Cohen (1998), Turner (2000), Foley and McDonald (2001), Yao and Knessl (2005, 2006), Gupta et al. (2007), Blanc (2009) and Dester et al. (2017). However, in the above mentioned papers, each of the queues has its own server, while in the current study we analyze a single-server polling system.

In most studies mentioned above, two dimensional Markovian queueing systems were investigated when one of the dimensions was bounded. The common analysis methods are (i) via probability generating functions (PGFs) (see e.g. Perel and Yechiali, 2008 and Perel and Yechiali, 2014), and (ii) via matrix geometric analysis (see e.g. books by Neuts, 1981 and Latouche and Ramaswami, 1999). However, although both methods rely on the same system's parameters, the complete relationship between the two methods has not been revealed yet. Recent papers that use both methods and explore relationships between them are Perel and Yechiali (2013), Paz and Yechiali (2014), Perel and Yechiali (2017) and Phung-Duc (2017). This paper, in addition to solving the joint JSQ and SLQ models, further investigates the above relationships. We note that in cases of special structure of the matrices A_0, A_1 and A_2 appearing in the infinitesimal generator matrix Q, used in the matrix geometric analysis (see Section 4), it is possible to derive a direct calculation of the entries of the rate matrix R (see Latouche and Ramaswami, 1999, Van Leeuwaarden and Winands, 2006, Van Houdt and van Leeuwaarden, 2011, Van Leeuwaarden et al., 2009, Hanukov and Yechiali, 2019), but none of the above cases is applicable in the current model.

When both queues are unbounded, one can apply a boundary value problem analysis, see e.g. Flatto (1989), Avrachenkov et al. (2014), Adan et al. (2016), or a truncation method as carried out in Bright and Taylor (1995). In contrast, in this paper, where we combine both customers' JSQ policy and server's SLQ regime into a unified system, we allow both dimensions of the non-symmetric two-queue polling system to be unbounded. We are able to derive the equilibrium joint probability distribution function of the queue lengths by using an un-conventional approach when forming relevant PGFs and when applying the matrix geometric method, thus avoiding an intricate boundary value problem analysis.

The paper continues as follows. The model is described and formulated in Section 2. In Section 3, steady-state equations, as well as probability generating functions are derived, and performance measures are calculated. In Section 4 the matrix geometric method is employed and the system's stability condition is derived, while Section 5 shortly presents a special case. Numerical results are presented in Section 6, as well as a comparison with a corresponding regular $M/G/1$ queue where service time is exponentially distributed with rate μ_i with probability p_i. Section 7 concludes the paper.

2. The model

We consider a polling system with a single server and two non-identical queues, denoted by Q_1 and Q_2. Customers arrive at the system according to a Poisson process with rate λ. Each arriving customer exercises the 'Join the Shortest Queue' (JSQ) policy, whereas if the lengths of the queues are equal, the customer joins Q_i w.p. p_i, where $p_1 + p_2 = 1$. Service duration of an arbitrary customer in Q_i is exponentially distributed with mean $1/\mu_i$, $i = 1, 2$. On the other hand, the server alternates between the two queues according to the 'Serve the Longest Queue' (SLQ) switching policy. That is, the server always attends the longest queue. As soon as the number of customers in the non-attendant queue rises above the number of customers in the attendant queue, the server stops serving the served customer and immediately switches to the other queue. The service of the interrupted customer will resume anew when its turn comes. Note that, if the server has completed service in Q_i $(i = 1, 2)$ and both queues are equal in size, the server does not switch. In this case, if a new customer arrives before the server completes one more service, then the new customer will join Q_1 or Q_2 with probability p_1 and p_2, respectively, independent of the servers' position. Consequently, if the latter customer joins Q_j (while the server is in Q_i, $i \neq j$), the server immediately switches to Q_j. The server will return to Q_i as soon as the number of customers in Q_i exceeds the number of customers in Q_j, and so on.

At time $t > 0$, let $L_i(t)$ denote the number of customers present in Q_i, $i = 1, 2$, and, assuming stability, let $L_i = \lim_{t \to \infty} L_i(t)$. Also, define $D(t) = L_1(t) - L_2(t)$ and $D = \lim_{t \to \infty} D(t)$. The dual regime

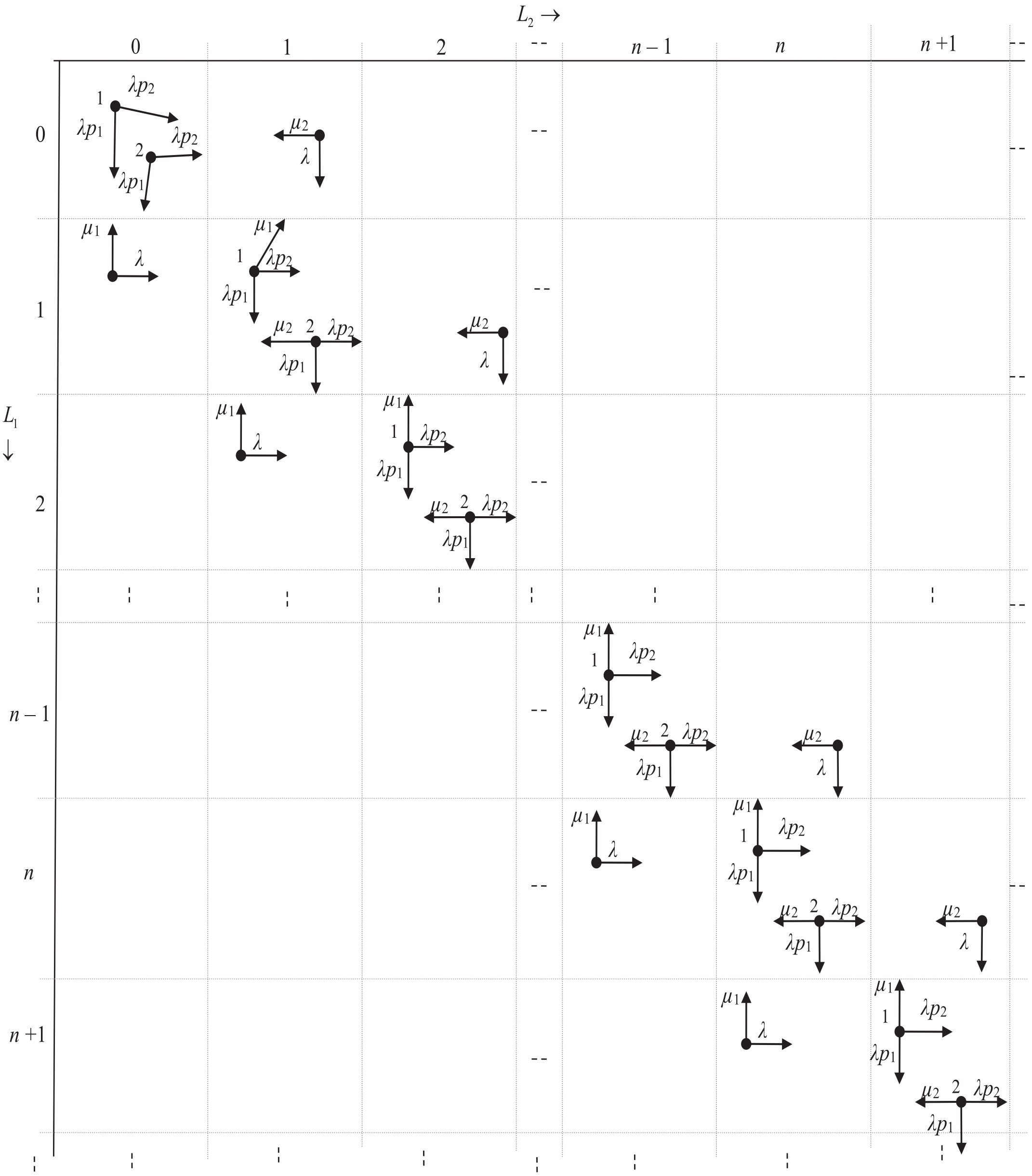

Fig. 1. Transition rate diagram of $(L_1(t),\ L_2(t))$.

JSQ-SLQ implies that the random process $D(t)$ may assume the values (-1), (0) or (1). In order to indicate the position of the server, we split state 0 into two states: 0_1 and 0_2, where 0_i indicates that the server attends Q_i, $i = 1, 2$. In this case, an arriving customer joins Q_i w.p. p_i. Note that $D = 1$ implies that the server is at Q_1, and since $L_1 > L_2$, an arriving customer joins Q_2, while $D = (-1)$ implies that $L_1 < L_2$, the server is at Q_2, and an arriving customer joins Q_1. We formulate the above non regular polling system as a two dimensional continuous time Markovian process, with state space $\{(n,\ d)\}$, for $n \geq 0$ and $d \in \mathfrak{D} = \{1, 0_1, 0_2, -1\}$. In the sequel, we discuss the stability condition of the system. Assuming that the stability condition holds, the system's steady state joint probability distribution function is denoted by $P_{n,d} = \mathbb{P}(L_1 = n, D = d)$. A transition rate diagram of the process $(L_1(t),\ L_2(t))$ is depicted in Fig. 1, from which the states of the resulting process $(L_1(t),\ D(t))$ are readily concluded. The numbers 1 or 2 in each square on the diagonal indicate the position of the server, i.e. at Q_1 or at Q_2, respectively.

3. Steady-state analysis using probability generating functions

In this section we derive the steady-state probability distribution function of the unbounded two-dimensional process defining the states of the system, i.e. the joint distribution of $(L_1,\ D)$. We use an unconventional construction of the probability generating functions (PGFs) and utilize their properties, as described below.

3.1. Balance equations and PGFs

Writing the balance equations for all n along each diagonal of Fig. 1, we obtain:

When $d = 1$,

$$(\lambda + \mu_1)P_{n,1} = \lambda p_1 (P_{n-1,0_1} + P_{n-1,0_2}) + \mu_2 P_{n,0_2}, \quad n \geq 1. \tag{1}$$

when $d = 0_1$,

$$\lambda P_{0,0_1} = \mu_1 P_{1,1}, \tag{2}$$

$$(\lambda+\mu_1)P_{n,0_1} = \lambda P_{n,1} + \mu_1 P_{n+1,1}, \;\; n \geq 1. \tag{3}$$

For $d = 0_2$,

$$\lambda P_{0,0_2} = \mu_2 P_{0,-1}, \tag{4}$$

$$(\lambda+\mu_2)P_{n,0_2} = \lambda P_{n-1,-1} + \mu_2 P_{n,-1}, \;\; n \geq 1. \tag{5}$$

Finally, for $d = -1$,

$$(\lambda+\mu_2)P_{n,-1} = \lambda p_2 (P_{n,0_1} + P_{n,0_2}) + \mu_1 P_{n+1,0_1}, \;\; n \geq 0. \tag{6}$$

For each $d \in \mathfrak{D} = \{1, 0_1, 0_2, -1\}$, define the conditional probability generating function of the number of customers in Q_1 as:

$$G_d(z) = \sum_n P_{n,d} z^n, \quad d \in \mathfrak{D}.$$

Multiplying Eq. (1) by z^n and summing over $n \geq 1$, we get

$$(\lambda+\mu_1)G_1(z) = \lambda p_1 z G_{0_1}(z) + (\lambda p_1 z + \mu_2)G_{0_2}(z) - \mu_2 P_{0,0_2}. \tag{7}$$

Repeating this process for $d = 0_1$ and $d = 0_2$, while using Eqs. (2)–(5), we obtain,

$$(\lambda+\mu_1)zG_{0_1}(z) = (\lambda z+\mu_1)G_1(z) + \mu_1 z P_{0,0_1}, \tag{8}$$

$$(\lambda+\mu_2)G_{0_2}(z) = (\lambda z+\mu_2)G_{-1}(z) + \mu_2 P_{0,0_2}. \tag{9}$$

Last, from Eq. (6) we derive

$$(\lambda+\mu_2)zG_{-1}(z) = (\lambda p_2 z+\mu_1)G_{0_1}(z) + \lambda p_2 z G_{0_2}(z) - \mu_1 P_{0,0_1}. \tag{10}$$

The set of Eqs. (7)–(10) can be written in a matrix form as

$$A(z)\cdot\vec{G}(z) = \vec{P}(z), \tag{11}$$

where

$$A(z) = \begin{pmatrix} \lambda+\mu_1 & -\lambda p_1 z & -(\lambda p_1 z+\mu_2) & 0 \\ -(\lambda z+\mu_1) & (\lambda+\mu_1)z & 0 & 0 \\ 0 & 0 & \lambda+\mu_2 & -(\lambda z+\mu_2) \\ 0 & -(\lambda p_2 z+\mu_1) & -\lambda p_2 z & (\lambda+\mu_2)z \end{pmatrix}.$$

$\vec{G}(z) = \left(G_1(z), G_{0_1}(z), G_{0_2}(z), G_{-1}(z)\right)^T$ is a 4-dimensional column vector of the desired PGF's, and $\vec{P}(z) = \left(-\mu_2 P_{0,0_2},\ \mu_1 z P_{0,0_1},\ \mu_2 P_{0,0_2},\ -\mu_1 P_{0,0_1}\right)^T$ is a vector containing the two unknown, so-called 'boundary probabilities', $P_{0,0_1}$ and $P_{0,0_2}$.

To explicitly obtain $G_d(z)$ we use Cramer's rule and write $G_d(z) = \frac{|A_d(z)|}{|A(z)|}$, $d \in \mathfrak{D} = \{1, 0_1, 0_2, -1\}$, where $|A|$ is the determinant of a matrix A, and $A_d(z)$ is the matrix obtained from $A(z)$ by replacing the corresponding column of the latter matrix by $\vec{P}(z)$. Note that the PGFs $G_d(z)$, $d \in \mathfrak{D}$, are expressed in terms of the two unknown boundary probabilities, $P_{0,0_1}$ and $P_{0,0_2}$, appearing in $\vec{P}(z)$. Two equations are required to calculate the latter probabilities. First, by the normalization condition, we have

$$\sum_{d\in\mathfrak{D}} G_d(1) = \sum_{d\in\mathfrak{D}} \lim_{z\to 1} \frac{|A_d(z)|}{|A(z)|} = 1. \tag{12}$$

The second relation between $P_{0,0_1}$ and $P_{0,0_2}$ is derived from the matrix $A(z)$. Since $G_d(z)$ is a (partial) probability generating function defined for all $|z| < 1$, each root of $|A(z)|$ is a root of $|A_d(z)|$. The determinant $|A(z)|$ is a 3-rd degree polynomial, and can be expressed as $|A(z)| = (1-z)h(z)$, where

$$\begin{aligned} h(z) = z^2\Big[\lambda^4 + \lambda^3(\mu_1(1+p_2) + \mu_2(1+p_1)) + \lambda^2(\mu_1^2 p_2 + \mu_2^2 p_1)\Big] \\ - z\Big[\lambda\mu_1\mu_2(\mu_1(1+p_1) + \mu_2(1+p_2)) + \mu_1^2\mu_2^2\Big] - \mu_1^2\mu_2^2. \end{aligned} \tag{13}$$

The quadratic polynomial $h(z)$ possesses 2 roots, denoted by z_1 and z_2, that can be expressed explicitly by solving a square root formula. Since $h(-1) > 0$ and $h(0) < 0$, then $z_1 \in (-1, 0)$ and can be used to obtain the required second relation. Since $h(\infty) = \infty$, the interval containing the other root, z_2, is determined by the system parameters (i.e. λ, μ_1, μ_2 and p_1) and may be either in (0,1] or in $(1, \infty)$. If $h(1) > 0$ then $z_2 \in (0, 1)$, and its use leads to a system of 3 equations in the two boundary probabilities, implying that there is no solution and the system is un-stable. Hence, the condition $h(1) < 0$, which means that $z_2 \in (1, \infty)$ is the system's stability condition. Note that

$$h(1) = (\lambda+\mu_1)(\lambda+\mu_2)\left[\lambda^2 + \lambda(\mu_1 p_2 + \mu_2 p_1) - 2\mu_1\mu_2\right],$$

and $h(1) < 0$ simplifies to the inequality

$$\lambda < \frac{-(\mu_1 p_2 + \mu_2 p_1) + \sqrt{(\mu_1 p_2 + \mu_2 p_1)^2 + 8\mu_1\mu_2}}{2}. \tag{14}$$

It will be verified again, when applying matrix geometric analysis in Section 4, that Eq. (14) defines the system's stability condition.

From all the above, $P_{0,0_1}$ and $P_{0,0_2}$ can be derived explicitly, which provides us with closed-form expressions for the PGFs, $G_d(z)$, for all $d \in \mathfrak{D}$. Specifically, explicit calculation of the determinants $|A_d(z)|$, for all $d \in \mathfrak{D}$ results in:

$$\begin{aligned} |A_1(z)| = -\lambda\mu_2 z(1-z)[P_{0,0_1}\mu_1(p_1 z(\lambda+\mu_2) + \lambda z + \mu_2) \\ + P_{0,0_2} z(\lambda+\mu_1)(\lambda+\mu_2 p_1)], \end{aligned} \tag{15}$$

$$\begin{aligned} |A_{0_1}(z)| = -(1-z)\Big[P_{0,0_1}\mu_1\big(\mu_1\mu_2^2 - \lambda^2 z^2(\lambda+\mu_1 p_2) \\ + \mu_2 z((1+p_1)\lambda\mu_1 + \mu_2(\lambda+\mu_1))\big) \\ + P_{0,0_2}\lambda\mu_2 z(\mu_2 p_1 + \lambda)(\lambda z + \mu_1)\Big], \end{aligned} \tag{16}$$

$$\begin{aligned} |A_{0_2}(z)| = -(1-z)[P_{0,0_1}\lambda\mu_1 z(\lambda z+\mu_2)(\lambda+\mu_1 p_2) \\ + P_{0,0_2}\mu_2\big(\mu_1^2\mu_2 - \lambda^2 z^2(\lambda+\mu_2 p_1) \\ + \mu_1 z((1+p_2)\lambda\mu_2 + \mu_1(\lambda+\mu_2))\big)], \end{aligned} \tag{17}$$

$$\begin{aligned} |A_{-1}(z)| = -\lambda\mu_1(1-z)[P_{0,0_1} z(\lambda+\mu_2)(\lambda+\mu_1 p_2) \\ + P_{0,0_2}\mu_2(p_2 z(\lambda+\mu_1) + \lambda z + \mu_1)]. \end{aligned} \tag{18}$$

Substituting z_1 in any of the above determinants (15)-(18) leads to the same equation, which provides us with one equation in the two boundary probabilities. The second equation is obtained from Eq. (12), which, after some algebra, leads to

$$\begin{aligned} \sum_{d\in\mathfrak{D}} G_d(1) = \\ \frac{P_{0,0_1}\mu_1(\lambda+\mu_2)\left[\lambda^2 + \lambda(\mu_1 p_2 + \mu_2 p_1) + 2\mu_2(\lambda+\mu_1)\right]}{(\lambda+\mu_1)(\lambda+\mu_2)\left[2\mu_1\mu_2 - \lambda(\mu_1 p_2 + \mu_2 p_1) - \lambda^2\right]} \\ + \frac{P_{0,0_2}\mu_2(\lambda+\mu_1)\left[\lambda^2 + \lambda(\mu_1 p_2 + \mu_2 p_1) + 2\mu_1(\lambda+\mu_2)\right]}{(\lambda+\mu_1)(\lambda+\mu_2)\left[2\mu_1\mu_2 - \lambda(\mu_1 p_2 + \mu_2 p_1) - \lambda^2\right]} = 1. \end{aligned} \tag{19}$$

Now, using $|A_d(z_1)| = 0$ for any $d \in \mathfrak{D}$ and Eq. (19) results in an explicit solution for $P_{0,0_1}$ and $P_{0,0_2}$ (in terms of the root z_1), given by

$$P_{0,0_1} = \frac{h(1)\beta_2(z_1)}{\alpha_1\beta_2(z_1) + \alpha_2\beta_1(z_1)},$$

$$P_{0,0_2} = \frac{h(1)\beta_1(z_1)}{\alpha_1\beta_2(z_1) + \alpha_2\beta_1(z_1)},$$

where

$$\begin{aligned} h(1) &= (\lambda+\mu_1)(\lambda+\mu_2)\left[\lambda^2 + \lambda(\mu_1 p_2 + \mu_2 p_1) - 2\mu_1\mu_2\right], \\ \alpha_1 &= \mu_1(\lambda+\mu_2)\left[\lambda^2 + \lambda(\mu_1 p_2 + \mu_2 p_1) + 2\mu_2(\lambda+\mu_1)\right], \\ \alpha_2 &= \mu_2(\lambda+\mu_1)\left[\lambda^2 + \lambda(\mu_1 p_2 + \mu_2 p_1) + 2\mu_1(\lambda+\mu_2)\right], \\ \beta_1(z_1) &= \mu_1\left[p_1 z_1(\lambda+\mu_2) + \lambda z_1 + \mu_2\right] - z_1(\lambda+\mu_2)(\lambda+\mu_1 p_2), \\ \beta_2(z_1) &= \mu_2\left[p_2 z_1(\lambda+\mu_1) + \lambda z_1 + \mu_1\right] - z_1(\lambda+\mu_1)(\lambda+\mu_2 p_1). \end{aligned}$$

3.2. Performance measures

In this section we derive the first and second moments of the queue lengths, as well as the correlation coefficient between them. We also calculate the proportion of time the server is idle, the mean sojourn time of an arbitrary customer in each queue, and the Laplace Stieltjes transforms (LSTs) of the sojourn times. Numerical results are presented in Section 6.2.

Define, respectively, the marginal probabilities of D and of L_1 as

$$P_{\bullet d} = \mathbb{P}(D=d) = \sum_{n=0}^{\infty} P_{n,d} = G_d(1), \quad d \in \mathfrak{D},$$

$$P_{n\bullet} = \mathbb{P}(L_1 = n) = \sum_{d\in\mathfrak{D}} P_{n,d}, \quad n \geq 0. \tag{20}$$

Then,

$$\begin{aligned}\mathbb{E}[D] &= 1\cdot G_1(1) + 0\cdot\Big(G_{0_1}(1)+G_{0_2}(1)\Big) + (-1)\cdot G_{-1}(1)\\ &= P_{\bullet 1} - P_{\bullet(-1)},\end{aligned}$$

$$\mathbb{E}[L_1] = \sum_{n=0}^{\infty} nP_{n\bullet} = \sum_{d\in\mathfrak{D}} G'_d(1),$$

$$\mathbb{E}[L_2] = \mathbb{E}[L_1] - \mathbb{E}[D].$$

Furthermore,

$$\begin{aligned}Cov(L_1,L_2) &= \mathbb{E}[L_1L_2] - \mathbb{E}[L_1]\mathbb{E}[L_2] = \mathbb{E}[L_1(L_1-D)] - \mathbb{E}[L_1]\mathbb{E}[L_2]\\ &= \mathbb{E}[L_1^2] - \mathbb{E}[L_1D] - \mathbb{E}[L_1]\mathbb{E}[L_2],\end{aligned}$$

where

$$\mathbb{E}[L_1^2] = \sum_{d\in\mathfrak{D}} G''_d(1) + \mathbb{E}[L_1],$$

$$\mathbb{E}[L_1D] = \sum_{d\in\mathfrak{D}}\sum_n ndP_{n,d} = \sum_n nP_{n,1} - \sum_n nP_{n,-1} = G'_1(1) - G'_{-1}(1).$$

Also, the variance of L_i, for $i = 1, 2$, is given by

$$Var(L_1) = \mathbb{E}[L_1^2] - (\mathbb{E}[L_1])^2,$$

$$\begin{aligned}Var(L_2) &= \mathbb{E}[L_2^2] - (\mathbb{E}[L_2])^2 = \mathbb{E}[(L_1-D)^2] - (\mathbb{E}[L_2])^2\\ &= \mathbb{E}[L_1^2] - 2\mathbb{E}[L_1D] + \mathbb{E}[D^2] - (\mathbb{E}[L_2])^2,\end{aligned}$$

where $\mathbb{E}[D^2] = P_{\bullet 1} + P_{\bullet(-1)}$.

From all the above, the correlation coefficient between L_1 and L_2, denoted by $Cor(L_1,\ L_2)$, can be explicitly calculated, using $Cor(L_1,L_2) = \frac{Cov(L_1,L_2)}{\sqrt{Var(L_1)Var(L_2)}}$ (for numerical results, see Section 6.2).

Let λ^i_{eff} denote the effective arrival rate to Q_i, i.e.

$$\lambda^1_{eff} = \lambda\Big(p_1(1-P_{\bullet 1}-P_{\bullet(-1)}) + P_{\bullet(-1)}\Big) = \lambda\Big(p_1(1-P_{\bullet 1}) + p_2P_{\bullet(-1)}\Big),$$

$$\lambda^2_{eff} = \lambda\Big(p_2(1-P_{\bullet 1}-P_{\bullet(-1)}) + P_{\bullet 1)}\Big) = \lambda\Big(p_2(1-P_{\bullet(-1)}) + p_1P_{\bullet 1}\Big).$$

Clearly, $\lambda^1_{eff} + \lambda^2_{eff} = \lambda$. Defining $\rho_i = \frac{\lambda^i_{eff}}{\mu_i}$, some algebra confirms that

$$\mathbb{P}(\text{Server is idle}) = P_{0,0_1} + P_{0,0_2} = 1 - \rho_1 - \rho_2.$$

Define W_i as the sojourn time of a customer in Q_i. Then, by Little's Law,

$$\mathbb{E}[W_i] = \frac{\mathbb{E}[L_i]}{\lambda^i_{eff}}.$$

Furthermore, the PGF's of the number of customers in Q_i (for $i = 1, 2$), denoted by $\hat{L_i}(z)$, are given by

$$\hat{L_1}(z) = G_1(z) + G_{0_1}(z) + G_{0_2}(z) + G_{-1}(z), \tag{21}$$

$$\hat{L_2}(z) = \frac{1}{z}G_1(z) + G_{0_1}(z) + G_{0_2}(z) + zG_{-1}(z). \tag{22}$$

Eq. (21) follows directly from Eq. (20), while Eq. (22) is a consequence of

$$\mathbb{P}(L_2 = k) = \begin{cases} P_{0,0_1} + P_{0,0_2} + P_{1,1}, & k = 0,\\ P_{k-1,-1} + P_{k,0_1} + P_{k,0_2} + P_{k+1,1}, & k \geq 1.\end{cases} \tag{23}$$

Hence,

$$\begin{aligned}\hat{L_2}(z) &= \mathbb{E}[z^{L_2}] = (P_{0,0_1} + P_{0,0_2} + P_{1,1})z^0\\ &\quad + \sum_{k=1}^{\infty}\Big(P_{k-1,-1} + P_{k,0_1} + P_{k,0_2} + P_{k+1,1}\Big)z^k\\ &= \frac{1}{z}G_1(z) + G_{0_1}(z) + G_{0_2}(z) + zG_{-1}(z).\end{aligned} \tag{24}$$

4. Matrix geometric method

4.1. Definitions and notations

An alternative approach to analyze the combined JSQ-SLQ model is by constructing a Quasi Birth and Death (QBD) process, with 4 phases, where phase d corresponds to $D = d$, for $d \in \mathfrak{D}$, and with an infinite number of levels, where each level corresponds to L_1, the total number of customers in Q_1. For $n \geq 1$ define $\mathcal{S}_n$ to be the set of states $\mathcal{S}_n = \{(n,1), (n,0_1), (n,0_2), (n,-1)\}$, and arrange the system's states in the order

$$\mathcal{S} = \Big\{(0,0_1), (0,0_2), (0,-1); \mathcal{S}_1; \mathcal{S}_2; \ldots; \mathcal{S}_n \ldots\Big\}$$

The infinitesimal generator of the QBD, denoted by Q, is given by

$$Q = \begin{pmatrix} B_1 & B_0 & 0 & \cdots & \cdots & \cdots\\ B_2 & A_1 & A_0 & 0 & \cdots & \cdots\\ 0 & A_2 & A_1 & A_0 & 0 & \cdots\\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots \end{pmatrix},$$

where,

$$B_0 = \begin{pmatrix} \lambda p_1 & 0 & 0 & 0\\ \lambda p_1 & 0 & 0 & 0\\ 0 & 0 & \lambda & 0\end{pmatrix}, \quad B_1 = \begin{pmatrix} -\lambda & 0 & \lambda p_2\\ 0 & -\lambda & \lambda p_2\\ 0 & \mu_2 & -(\lambda+\mu_2)\end{pmatrix},$$

$$B_2 = \begin{pmatrix} \mu_1 & 0 & 0\\ 0 & 0 & \mu_1\\ 0 & 0 & 0\\ 0 & 0 & 0\end{pmatrix}.$$

and

$$A_0 = \begin{pmatrix} 0 & 0 & 0 & 0\\ \lambda p_1 & 0 & 0 & 0\\ \lambda p_1 & 0 & 0 & 0\\ 0 & 0 & \lambda & 0\end{pmatrix},$$

$$A_1 = \begin{pmatrix} -(\lambda+\mu_1) & \lambda & 0 & 0\\ 0 & -(\lambda+\mu_1) & 0 & \lambda p_2\\ \mu_2 & 0 & -(\lambda+\mu_2) & \lambda p_2\\ 0 & 0 & \mu_2 & -(\lambda+\mu_2)\end{pmatrix},$$

$$A_2 = \begin{pmatrix} 0 & \mu_1 & 0 & 0\\ 0 & 0 & 0 & \mu_1\\ 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\end{pmatrix}.$$

4.2. Stability condition

Define the matrix $A = A_0 + A_1 + A_2$. We get

$$A = \begin{pmatrix} -(\lambda+\mu_1) & \lambda+\mu_1 & 0 & 0\\ \lambda p_1 & -(\lambda+\mu_1) & 0 & \lambda p_2+\mu_1\\ \lambda p_1+\mu_2 & 0 & -(\lambda+\mu_2) & \lambda p_2\\ 0 & 0 & \lambda+\mu_2 & -(\lambda+\mu_2)\end{pmatrix},$$

The matrix A is the infinitesimal generator matrix of the process describing the evolution of D, given that Q_1 is not empty. Let $\vec{\pi}$ be the stationary vector of the matrix A, i.e. $\vec{\pi}A = \vec{0}$ and $\vec{\pi} \cdot \vec{e} = 1$ (where $\vec{e}$ is a 4-dimensional column vector with all its entries equal to 1). It then follows that

$$\vec{\pi} = \left(\frac{\lambda p_1 + \mu_2}{2(\lambda + \mu_1 + \mu_2)} \quad \frac{\lambda p_1 + \mu_2}{2(\lambda + \mu_1 + \mu_2)} \quad \frac{\lambda p_2 + \mu_1}{2(\lambda + \mu_1 + \mu_2)} \quad \frac{\lambda p_2 + \mu_1}{2(\lambda + \mu_1 + \mu_2)} \right).$$

The stability condition (see Neuts, 1981) is

$$\vec{\pi} A_0 \vec{e} < \vec{\pi} A_2 \vec{e},$$

which, after some algebra, translates here into

$$\lambda^2 + \lambda(\mu_1 p_2 + \mu_2 p_1) - 2\mu_1\mu_2 < 0,$$

or, equivalently,

$$\lambda < \frac{-(\mu_1 p_2 + \mu_2 p_1) + \sqrt{(\mu_1 p_2 + \mu_2 p_1)^2 + 8\mu_1\mu_2}}{2}. \tag{25}$$

Indeed, the stability condition (25) obtained by the matrix geometric method is equivalent to the condition (14) derived when analyzing the system via the PGF's method.

In the symmetric case, where $\mu_1 = \mu_2 = \mu$, the stability condition (14) translates into $\lambda < \mu$, for any value of p_1. This occurs when both service rates are equal and the system can be looked upon as a single $M(\lambda)/M(\mu)/1$ queue, for which the known stability condition is $\lambda < \mu$.

4.3. Calculation of the equilibrium distribution

For $n \geq 0$ define the steady-state probability vector $\vec{P}_n$, as follows:

$$\vec{P}_n = \begin{cases} (P_{0,0_1}, P_{0,0_2}, P_{0,-1}), & n = 0, \\ (P_{n,1}, P_{n,0_1}, P_{n,0_2}, P_{n,-1}), & n \geq 1. \end{cases}$$

From Neuts (1981),

$$\vec{P}_n = \vec{P}_1 R^{n-1}, \quad n \geq 1,$$

where R is the minimal non-negative solution of the matrix quadratic equation

$$A_0 + RA_1 + R^2 A_2 = 0. \tag{26}$$

The vectors $\vec{P}_0$, $\vec{P}_1$, can be found by solving the following linear system of equations:

$$\vec{P}_0 B_1 + \vec{P}_1 B_2 = \vec{0},$$
$$\vec{P}_0 B_0 + \vec{P}_1 A_1 + \vec{P}_1 R A_2 = \vec{0},$$
$$\vec{P}_0 \vec{e}_0 + \vec{P}_1 [\mathbf{I} - R]^{-1} \vec{e} = 1,$$

where $\vec{e}_0$ is a 3-dimensional vector of 1's and $\mathbf{I}$ is a 4×4 identity matrix.

The mean total number of customers in Q_1, $\mathbb{E}[L_1]$ is given by

$$\mathbb{E}[L_1] = \sum_{n=1}^{\infty} n\vec{P}_n \vec{e} = \sum_{n=1}^{\infty} n\vec{P}_1 R^{n-1} \vec{e} = \vec{P}_1 [\mathbf{I} - R]^{-2} \vec{e}. \tag{27}$$

4.4. Characterization of the rate matrix R

The matrix $R = [r_{i,j}]$ for $i, j = 1, 2, 3, 4$, can be calculated by using well-known algorithms, see e.g. Neuts (1981), Latouche and Ramaswami (1999) and Artalejo and Gómez-Corral (2008), or, by solving (in our case) a system of 16 non-linear equations with 16 variables. However, we are able to characterize some of the properties of R in this model. Following Ch. 6.2 in Latouche and Ramaswami (1999), the rate matrix R can be represented as $R = A_0 N$, where the element N_{ij} of the matrix N is the expected number of visits to state (n, j), starting from state (n, i), before the first visit to any of the states in levels lower than n. In our context, L_1 represents the levels, and the index j refers to the phases (represented by D). Without calculating N, since the entries of the first row of A_0 are all zeros, all elements in the first row of R are zeros as well. That is, $r_{1,j} = 0, \; j = 1, 2, 3, 4$. Furthermore, since the second and third rows of A_0 are equal, the second and third rows of R will also be equal, namely $r_{2,j} = r_{3,j}, \; j = 1, 2, 3, 4$. In addition, from explicitly writing Eq. (26), each element of R can be expressed in terms of only two elements, $r_{2,1}$ and $r_{2,2}$. These observations reduce the calculation efforts considerably.

5. Special case: $p_1 = 1$

Assume that the service rates are observable by the customers, so that whenever $L_1 = L_2$, an arriving customer always joins the queue with the faster service rate. Without loss of generality, assume that $\mu_1 > \mu_2$, implying that $p_1 = 1$. When $L_1 \neq L_2$, an arriving customer will join the shortest queue. The server's switching policy remains the same, i.e. serve the longest queue.

The stability condition given in Eq. (14) becomes

$$\lambda < \frac{-\mu_2 + \sqrt{\mu_2^2 + 8\mu_1\mu_2}}{2}. \tag{28}$$

The expressions for all the performance measures calculated in Section 3.2 are slightly modified when $p_1 = 1$ (and $p_2 = 0$). In Section 6.2 we provide numerical results for various values of p_1, including $p_1 = 1$.

6. Numerical results and comparison with an *M*/*G*/1 queue

In this section we first discuss a related $M/G/1$ queue and then present numerical results of the JSQ-SLQ system's performance measures for a set of parameter values. The results are than compared with those of the $M/G/1$ model.

6.1. A corresponding M/G/1 queue

Consider a single server queueing system with a Poisson arrival stream with rate λ and service time B, defined as

$$B \sim \begin{cases} exp(\mu_1) & w.p. \; p_1 \\ exp(\mu_2) & w.p. \; p_2 \end{cases},$$

so that

$$\mathbb{E}[B] = \frac{p_1}{\mu_1} + \frac{p_2}{\mu_2},$$
$$\mathbb{E}[B^2] = \frac{2p_1}{\mu_1^2} + \frac{2p_2}{\mu_2^2}.$$

Let W_q denote the waiting time of an arbitrary customer, W its total sojourn time in the system, and L the total number of customers in the system. Then, from the well-known Pollaczek–Khintchine formula, with $\rho_i = \frac{\lambda p_i}{\mu_i}$, we get

$$\mathbb{E}[W_q] = \frac{\lambda \mathbb{E}[B^2]}{2(1 - \lambda \mathbb{E}[B])} = \frac{\rho_1 \mu_2 + \rho_2 \mu_1}{\mu_1 \mu_2 (1 - \rho_1 - \rho_2)},$$

and

$$\mathbb{E}[W] = \mathbb{E}[W_q] + \mathbb{E}[B],$$
$$\mathbb{E}[L] = \lambda \mathbb{E}[W].$$

In Section 6.2 below, where numerical results are presented, the performance measures of the above $M/G/1$ queue are compared with those of the combined JSQ-SLQ model.

Table 1
Numerical results for $\lambda = 4,\ \mu_2 = 5,\ p_1 = 0.2,\ p_2 = 0.8$.

values of μ_1	$\mathbb{E}[L_1]$	$\mathbb{E}[L_2]$	$\mathbb{E}[W_1]$	$\mathbb{E}[W_2]$	λ^1_{eff}	λ^2_{eff}	ρ^1_{eff}	ρ^2_{eff}	$Cor(L_1, L_2)$
3.3	9.49	9.53	6.52	3.75	1.46	2.54	0.44	0.51	0.9975
3.5	6.23	6.28	4.24	2.48	1.47	2.53	0.42	0.51	0.9943
4	3.46	3.53	2.31	1.41	1.50	2.50	0.37	0.50	0.9836
4.5	2.46	2.55	1.62	1.03	1.52	2.48	0.34	0.49	0.9708
5	1.95	2.05	1.27	0.83	1.54	2.46	0.31	0.49	0.9575
6	1.42	1.55	0.91	0.64	1.57	2.43	0.26	0.49	0.9324
8	0.98	1.14	0.61	0.48	1.60	2.40	0.20	0.48	0.8927
20	0.51	0.73	0.3	0.32	1.67	2.33	0.08	0.46	0.8153
100	0.34	0.61	0.20	0.27	1.71	2.29	0.02	0.45	0.8040
500	0.31	0.59	0.18	0.26	1.72	2.28	0.003	0.45	0.8097

Table 2
Numerical results for $\lambda = 4,\ \mu_2 = 5,\ p_1 = p_2 = 0.5$.

values of μ_1	$\mathbb{E}[L_1]$	$\mathbb{E}[L_2]$	$\mathbb{E}[W_1]$	$\mathbb{E}[W_2]$	λ^1_{eff}	λ^2_{eff}	ρ^1_{eff}	ρ^2_{eff}	$Cor(L_1, L_2)$
3.3	61.46	61.39	32.99	28.72	1.87	2.13	0.56	0.43	0.9999
3.5	12.52	12.46	6.64	5.89	1.88	2.12	0.53	0.42	0.9985
4	4.35	4.31	2.25	2.08	1.93	2.07	0.48	0.41	0.9888
4.5	2.71	2.69	1.37	1.32	1.97	2.03	0.44	0.41	0.9738
5	2	2	1	1	2	2	0.4	0.4	0.9565
6	1.35	1.38	0.66	0.71	2.05	1.95	0.34	0.39	0.9203
8	0.87	0.92	0.41	0.49	2.12	1.88	0.26	0.38	0.8568
20	0.37	0.51	0.16	0.29	2.28	1.72	0.11	0.35	0.7130
100	0.20	0.39	0.08	0.24	2.38	1.62	0.02	0.32	0.6973
500	0.17	0.37	0.07	0.23	2.41	1.59	0.005	0.32	0.7183

Table 3
Numerical results for $\lambda = 4,\ \mu_2 = 5,\ p_1 = 0.8,\ p_2 = 0.2$.

values of μ_1	$\mathbb{E}[L_1]$	$\mathbb{E}[L_2]$	$\mathbb{E}[W_1]$	$\mathbb{E}[W_2]$	λ^1_{eff}	λ^2_{eff}	ρ^1_{eff}	ρ^2_{eff}	$Cor(L_1, L_2)$
3.5	130.99	130.83	57.23	76.46	2.29	1.71	0.65	0.34	0.9999
4	5.65	5.51	2.40	3.36	1.64	2.36	0.59	0.33	0.9933
4.5	2.97	2.86	1.23	1.80	1.59	2.41	0.54	0.32	0.9778
5	2.05	1.95	0.83	1.27	2.46	1.54	0.49	0.31	0.9575
6	1.30	1.22	0.51	0.83	2.54	1.46	0.42	0.29	0.9111
8	0.78	0.73	0.29	0.54	2.66	1.34	0.33	0.27	0.8215
20	0.28	0.30	0.09	0.29	2.94	1.06	0.15	0.21	0.5803
100	0.09	0.17	0.03	0.22	3.19	0.81	0.03	0.16	0.5326
500	0.06	0.15	0.02	0.21	3.26	0.74	0.006	0.15	0.5955

Table 4
Numerical results for $\lambda = 4,\ \mu_2 = 5,\ p_1 = 1,\ p_2 = 0$.

values of μ_1	$\mathbb{E}[L_1]$	$\mathbb{E}[L_2]$	$\mathbb{E}[W_1]$	$\mathbb{E}[W_2]$	λ^1_{eff}	λ^2_{eff}	ρ^1_{eff}	ρ^2_{eff}	$Cor(L_1, L_2)$
4	6.93	6.73	2.60	4.94	2.64	1.36	0.66	0.27	0.93
4.5	3.17	2.98	1.17	2.31	2.71	1.29	0.60	0.26	0.86
5	2.08	1.91	0.75	1.56	2.77	1.23	0.55	0.25	0.8
6	1.27	1.15	0.44	0.99	2.87	1.13	0.48	0.22	0.71
8	0.73	0.61	0.24	0.63	3.03	0.97	0.38	0.19	0.59
20	0.24	0.17	0.07	0.31	3.44	0.56	0.17	0.11	0.58
100	0.05	0.03	0.01	0.22	3.85	0.15	0.039	0.029	0.14
500	0.01	0.006	0.002	0.2	3.97	0.03	0.008	0.006	0.06

6.2. Numerical results

Tables 1–5 below present sets of numerical results, where the calculated performance measures in all tables are $\mathbb{E}[L_i]$, $\mathbb{E}[W_i]$, λ^i_{eff}, ρ^i_{eff} $(i = 1, 2)$ and $Cor(L_1, L_2)$. The tables maintain the same parameter values: $\lambda = 4$ and $\mu_2 = 5$, but differ by the values of the parameter p_1, where $p_1 = 0.2, 0.5, 0.8, 1$ in Tables 1–4, respectively, while $p_1 = \frac{\mu_1}{\mu_1+\mu_2}$ in Table 5. In each table, μ_1 is a variable that its value varies between 3.3 and 8. It is seen that the combined operating policy, namely JSQ with SLQ, greatly achieves its goal of balancing mean queue lengths, as well as mean waiting times. Table 2 exhibits that even if μ_1 and μ_2 take significantly different values, the ratio $\mathbb{E}[L_1]/\mathbb{E}[L_2]$ remains quite stable. For instance, when $\mu_1 = 3.3,\ \mu_2 = 5$, or when $\mu_1 = 8$ and $\mu_2 = 5$, similar values for the ratios are obtained: $\mathbb{E}[L_1]/\mathbb{E}[L_2] = 1.00114$, and $\mathbb{E}[L_1]/\mathbb{E}[L_2] = 0.94565$, respectively. The corresponding ratios between the mean waiting times are 1.14868 and 0.83673. Tables 1 and 3 show that even when $p_1 = 0.2$ and $p_2 = 0.8$ (and vice versa), while $\mu_2 = 5$ and μ_1 varies between $\mu_1 = 3.5$ to $\mu_1 = 8$, the difference in mean queue lengths is negligible. However, the difference between mean waiting times is high. Table 5 exhibits a stronger balancing result. If the joining probabilities are taken as the relative ratios of the service rates, i.e $p_i = \frac{\mu_i}{\mu_1+\mu_2},\ i = 1, 2,$ causing higher proportion of customers to join the queue with the faster service, both mean queue lengths are significantly reduced, where the maximum difference is 3.5%.

Table 5
Numerical results for $\lambda = 4$, $\mu_2 = 5$, $p_1 = \frac{\mu_1}{\mu_1+\mu_2}$, $p_2 = \frac{\mu_2}{\mu_1+\mu_2}$.

values of μ_1	$\mathbb{E}[L_1]$	$\mathbb{E}[L_2]$	$\mathbb{E}[W_1]$	$\mathbb{E}[W_2]$	λ^1_{eff}	λ^2_{eff}	ρ^1_{eff}	ρ^2_{eff}	$Cor(L_1, L_2)$
3.3	22.00	21.97	12.75	9.65	1.72	2.27	0.52	0.45	0.9995
3.5	9.75	9.72	5.53	4.35	1.76	2.24	0.50	0.45	0.9976
4	4.16	4.14	2.25	1.93	1.85	2.15	0.46	0.43	0.9878
4.5	2.69	2.67	1.39	1.29	1.93	2.07	0.43	0.41	0.9735
5	2	2	1	1	2	2	0.4	0.4	0.9565
6	1.34	1.35	0.63	0.72	2.12	1.87	0.35	0.37	0.9187
8	0.82	0.85	0.35	0.51	2.22	1.67	0.29	0.33	0.8428
20	0.28	0.30	0.09	0.29	2.94	1.06	0.15	0.21	0.5803
100	0.060	0.067	0.01	0.21	3.68	0.32	0.03	0.06	0.3762
500	0.01	0.01	0.003	0.02	3.93	0.07	0.007	0.01	0.3395

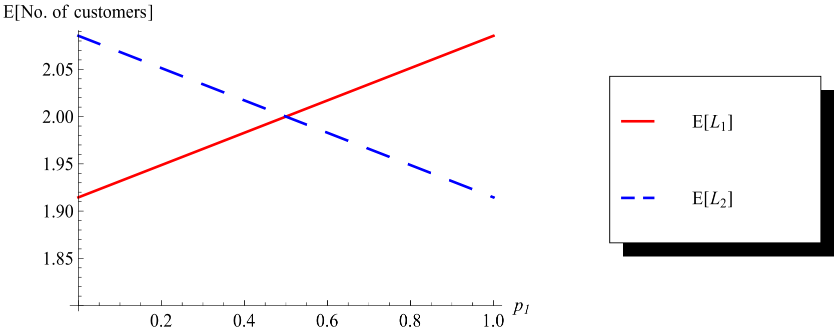

Fig. 2. $\mathbb{E}[L_1]$ and $\mathbb{E}[L_2]$ as a function of p_1, for $\lambda = 4$, $\mu_1 = \mu_2 = 5$.

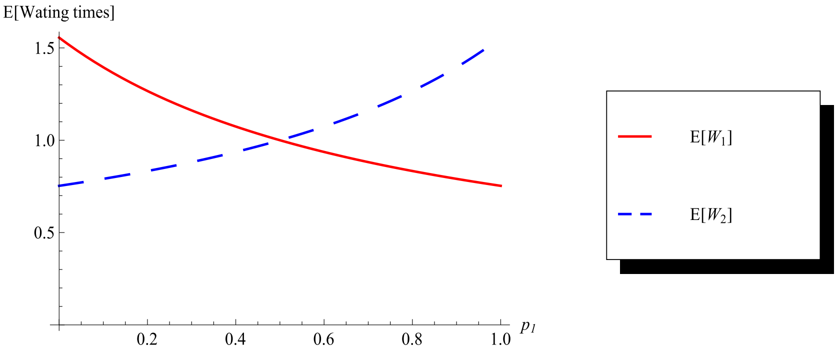

Fig. 3. $\mathbb{E}[W_1]$ and $\mathbb{E}[W_2]$ as a function of p_1, for $\lambda = 4$, $\mu_1 = \mu_2 = 5$.

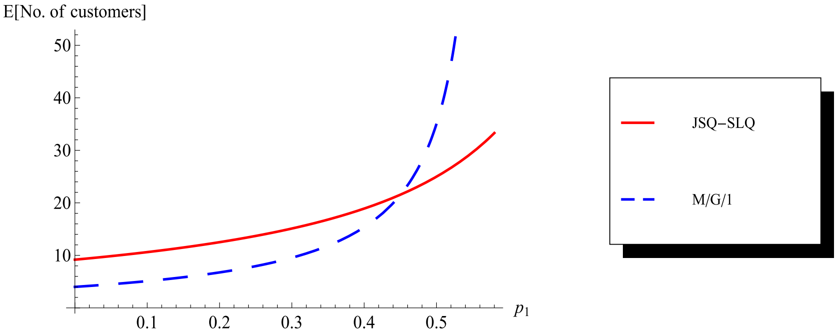

Fig. 4. $\mathbb{E}[L]$ for the JSQ-SLQ and for the $M/G/1$ models as a function of p_1, for $\lambda = 4$, $\mu_1 = 3.5$, $\mu_2 = 5$.

Furthermore, in all tables, as μ_1 increases, the fraction of time the server resides in both queues (i.e. $\rho^1_{eff} + \rho^2_{eff}$) reduces considerably.

Figs. 2 and 3 below show graphically mean queue sizes and mean waiting times as functions of p_1 for the given parameter values. Note that, in deviation from regular queues, where increasing arrival rate increases mean queue size and mean waiting time, larger p_1 in the JSQ-SLQ model (implying more customers joining Q_1 when the queues are equal), reduces mean queue size and mean waiting time in Q_1. This is a consequence of the SLQ regime, that directs the server to the longer queue (Q_1), thus reducing mean queue size and waiting time there. Figs. 4–6 show, both for

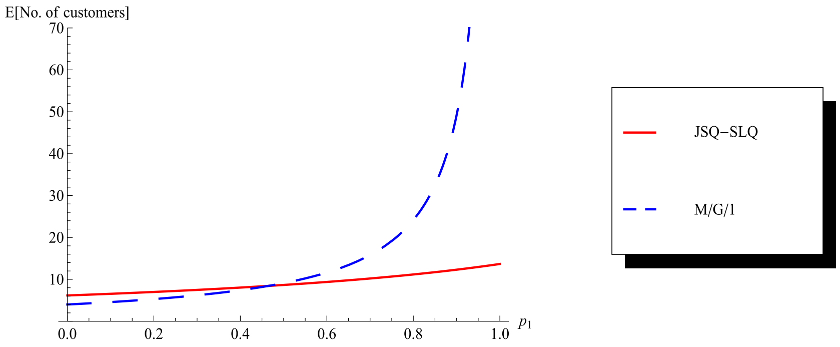

Fig. 5. $\mathbb{E}[L]$ for the JSQ-SLQ and the $M/G/1$ models as a function of p_1, for $\lambda = 4,\ \mu_1 = 4,\ \mu_2 = 5$.

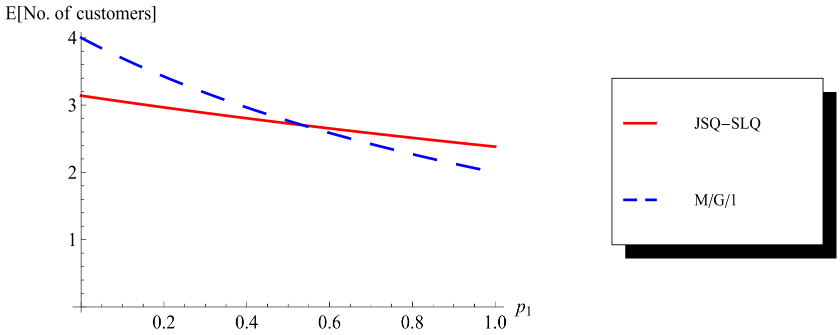

Fig. 6. $\mathbb{E}[L]$ for the JSQ-SLQ and the $M/G/1$ models as a function of p_1, for $\lambda = 4,\ \mu_1 = 6,\ \mu_2 = 5$.

Table 6
Numerical results of $\mathbb{E}[L]$ for the $M/G/1$ model, when $\lambda = 4$ and $\mu_2 = 5$.

Values of μ_1	$p_1 = 0.2$	$p_1 = 0.5$	$p_1 = 0.8$	$p_1 = \frac{\mu_1}{\mu_1+\mu_2}$
3.3	7.74	–	–	27.79
3.5	6.75	35.03	–	16.48
4	5.29	9.1	24.16	8.09
4.5	4.49	5.44	6.77	5.34
5	4	4	4	4
6	3.42	2.77	2.27	2.68
8	2.90	1.92	1.31	1.65

the JSQ-SLQ and for the $M/G/1$ models, the value of the mean total number of customers in the system, i.e. $\mathbb{E}[L]$, as a function of p_1. In all figures $\lambda = 4$ and $\mu_2 = 5$. In Fig. 4 $\mu_1 = 3.5$. It is seen that for small values of p_1, $\mathbb{E}[L]$ in the $M/G/1$ model is smaller than $\mathbb{E}[L]$ in the studied JSQ-SLQ model, but not drastically. However, as p_1 increases, the $M/G/1$ system becomes unstable at $p_1 = 0.58$, while the JSQ-SLQ system becomes unstable only when $p_1 = 0.84$. That is, the JSQ-SLQ policy helps to regulate the system and keeps it stable, even if the service rate μ_1 is slow. In Fig. 5 is it seen that for $p_1 < 0.5$, there is a slight advantage for the $M/G/1$ model vs. the JSQ-SLQ model (in terms of $\mathbb{E}[L]$). However, as p_1 increases, the JSQ-SLQ outperforms the $M/G/1$ model, as there are significant differences between the corresponding values of $\mathbb{E}[L]$. Note that for $p_1 = 1$ the $M/G/1$ model becomes unstable, while the JSQ-SLQ model remains stable. In Fig. 6, when $\mu_1 = 6$, it is shown that the differences between $\mathbb{E}[L]$ in both models are negligible. Clearly, for any value of p_1, both models are stable as $\lambda < Min(\mu_1,\ \mu_2)$ (Table 6).

7. Concluding remarks

This paper combines two different queueing regimes, usually treated separately and known as 'Join the Shortest Queue' (JSQ) and 'Serve the Longest Queue' (SLQ), into a unified system: each arriving customer joins the shortest queue, while the server always attends the longest queue. Both regimes aim at minimizing the difference between the queue lengths. The resulting non-conventional two-dimensional continuous-time Markov process describing the system is investigated via both probability generating functions and matrix geometric methods. By applying an non-usual approach we are able to fully analyze a non-symmetric un-bounded two-dimensional process without resorting to a complicated boundary-value problem analysis. The system's performance measures are analytically derived and its stability condition is determined.

The paper presents numerical results, exhibiting how the mean queue lengths, mean waiting times and loads are affected by the system's parameter values. Furthermore, the combined JSQ-SLQ system is compared with a corresponding $M/G/1$ queue. The numerical results show that in some range of the system's parameters the performance measures of the JSQ-SLQ model and the $M/G/1$ queue are not drastically different. However, a change in the system's parameters causes the $M/G/1$ queue to become unstable faster than the presented model, since the JSQ-SLQ policy helps to regulate the system.

References

Adan, I.J., Boxma, O.J., Kapodistria, S., Kulkarni, V.G., 2016. The shorter queue polling model. Ann. Oper. Res. 241 (1–2), 167–200.

Adan, I.J., Kapodistria, S., van Leeuwaarden, J.S., 2013. Erlang arrivals joining the shorter queue. Queueing Syst. 74 (2–3), 273–302.

Adan, I.J., Wessels, J., Zijm, W., 1991. Analysis of the asymmetric shortest queue problem. Queueing Syst. 8 (1), 1–58.

Adan, I.J., Wessels, J., Zijm, W., 1991. Analysis of the asymmetric shortest queue problem with threshold jockeying. Commun. Stat. 7 (4), 615–627.

Artalejo, J.R., Gómez-Corral, A., 2008. Retrial Queueing Systems: A Computational Approach. Springer.

Avrachenkov, K., Nain, P., Yechiali, U., 2014. A retrial system with two input streams and two orbit queues. Queueing Syst. 77 (1), 1–31.

Baharian, G., Tezcan, T., 2011. Stability analysis of parallel server systems under longest queue first. Math. Methods Oper. Res. 74 (2), 257.

Blanc, J.P.C., 2009. Bad luck when joining the shortest queue. Eur. J. Oper. Res. 195 (1), 167–173.

Boon, M.A., Van der Mei, R., Winands, E.M., 2011. Applications of polling systems. Surv. Oper. Res. Manag.Sci. 16 (2), 67–82.

Bright, L., Taylor, P.G., 1995. Calculating the equilibrium distribution in level dependent quasi-birth-and-death processes. Stoch. Models 11 (3), 497–525.

Browne, S., Yechiali, U., 1989. Dynamic priority rules for cyclic-type queues. Adv. Appl. Probab. 21 (2), 432–450.

Cohen, J., 1987. A two-queue, one-server model with priority for the longer queue. Queueing Syst. 2 (3), 261–283.

Cohen, J., 1998. Analysis of the asymmetrical shortest two-server queueing model. Int. J. Stoch. Anal. 11 (2), 115–162.

Dester, P.S., Fricker, C., Tibi, D., 2017. Stationary analysis of the shortest queue problem. Queueing Syst. 87 (3–4), 211–243.

Flatto, L., 1989. The longer queue model. Probab. Eng. Inf. Sci. 3 (4), 537–559.

Foley, R.D., McDonald, D.R., 2001. Join the shortest queue: stability and exact asymptotics. Ann. Appl. Probab. 569–607.

Gupta, V., Balter, M.H., Sigman, K., Whitt, W., 2007. Analysis of join-the-shortest-queue routing for web server farms. Perform. Eval. 64 (9–12), 1062–1081.

Halfin, S., 1985. The shortest queue problem. J. Appl. Probab. 22 (4), 865–878.

Hanukov, G., Yechiali, U., 2019. Explicit solutions for continuous-time QBD processes by using relations between matrix geometric analysis and the probability generating functions method. Technical report, Dept. of Statistics and Operations Research, Tel-Aviv University. Submitted for publication in Probability in the Engineering and Informational Sciences.

Hordijk, A., Koole, G., 1990. On the optimality of the generalized shortest queue policy. Probab. Eng. Inf. Sci. 4 (4), 477–487.

Houtum, G.-J.V., Adan, I., Der wal, J., 1997. The symmetric longest queue system. Stoch. Models 13 (1), 105–120.

Jolles, A., Perel, E., Yechiali, U., 2018. Alternating Server with Non-Zero Switch-Over Times and Opposite-Queue Threshold-Based Switching Policy. Working paper

Knessl, C., Yao, H., 2013. On the nonsymmetric longer queue model: joint distribution, asymptotic properties, and heavy traffic limits. Adv. Oper. Res. 2013.

Latouche, G., Ramaswami, V., 1999. Introduction to matrix analytic methods in stochastic modeling. Siam.

Maguluri, S.T., Hajek, B., Srikant, R., 2014. The stability of longest-queue-first scheduling with variable packet sizes. IEEE Trans. Automat. Control 59 (8), 2295–2300.

Menich, R., 1987. Optimally of shortest queue routing for dependent service stations. In: Decision and Control, 1987. 26th IEEE Conference on, vol. 26. IEEE, pp. 1069–1072.

Menich, R., Serfozo, R.F., 1991. Optimality of routing and servicing in dependent parallel processing systems. Queueing Syst. 9 (4), 403–418.

Neuts, M.F., 1981. Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach. The Johns Hopkins University Press.

Paz, N., Yechiali, U., 2014. An M/M/1 queue in random environment with disasters. Asia-Pac. J. Oper. Res. 31 (03), 1450016.

Pedarsani, R., Walrand, J., 2016. Stability of multiclass queueing networks under longest-queue and longest-dominating-queue scheduling. J. Appl. Probab. 53 (2), 421–433.

Perel, E., Yechiali, U., 2008. Queues where customers of one queue act as servers of the other queue. Queueing Syst. 60 (3–4), 271–288.

Perel, E., Yechiali, U., 2017. Two-queue polling systems with switching policy based on the queue that is not being served. Stoch. Models 33 (3), 430–450.

Perel, N., Yechiali, U., 2013. The israeli queue with priorities. Stoch. Models 29 (3), 353–379.

Perel, N., Yechiali, U., 2014. The israeli queue with retrials. Queueing Syst. 78 (1), 31–56.

Phung-Duc, T., 2017. Exact solutions for m/m/c/setup queues. Telecommun. Syst. 64 (2), 309–324.

Ravid, R., Boxma, O.J., Perry, D., 2013. Repair systems with exchangeable items and the longest queue mechanism. Queueing Syst. 73 (3), 295–316.

Takagi, H., 1986. Analysis of Polling Systems. MIT press.

Turner, S.R., 2000. A join the shorter queue model in heavy traffic. J. Appl. Probab. 37 (1), 212–223.

Van Houdt, B., van Leeuwaarden, J.S., 2011. Triangular m/g/1-type and tree-like quasi-birth-death markov chains. INFORMS J. Comput. 23 (1), 165–171.

Van Leeuwaarden, J., Winands, E., 2006. Quasi-birth-and-death processes with an explicit rate matrix. Stoch. Models 22 (1), 77–98.

Van Leeuwaarden, J.S., Squillante, M.S., Winands, E.M., 2009. Quasi-birth-and-death processes, lattice path counting, and hypergeometric functions. J. Appl. Probab. 46 (2), 507–520.

Winston, W., 1977. Optimality of the shortest line discipline. J. Appl. Probab. 14 (1), 181–189.

Yao, H., Knessl, C., 2005. On the infinite server shortest queue problem: symmetric case. Stoch. Models 21 (1), 101–132.

Yao, H., Knessl, C., 2006. On the infinite server shortest queue problem: non-symmetric case. Queueing Syst. 52 (2), 157–177.

Yechiali, U., 1993. Analysis and control of polling systems. Perform. Eval. Comput.Commun. Syst. 630–650.

Zheng, Y.-S., Zipkin, P., 1990. A queueing model to analyze the value of centralized inventory information. Oper. Res. 38 (2), 296–307.