Mathematical Methods of Operations Research
https://doi.org/10.1007/s00186-023-00848-7

ORIGINAL ARTICLE

The non-preemptive 'Join the Shortest Queue–Serve the Longest Queue' service system with or without switch-over times

Efrat Perel[1] · Nir Perel[1] · Uri Yechiali[2]

Received: 30 May 2023 / Revised: 26 November 2023 / Accepted: 6 December 2023

Abstract

A 2-queue system with a single-server operating according to the combined 'Join the Shortest Queue–Serve the Longest Queue' regime is analyzed. Both cases, with or without server's switch-over times, are investigated under the non-preemptive discipline. Instead of dealing with a state space comprised of two un-bounded dimensions, a non-conventional formulation is constructed, leading to an alternative two-dimensional state space, where only one dimension is infinite. As a result, the system is defined as a quasi birth and death process and is analyzed via both the probability generating functions method and the matrix geometric formulation. Consequently, the system's two-dimensional probability mass function is derived, from which the system's performance measures, such as mean queue sizes, mean sojourn times, fraction of time the server resides in each queue, correlation coefficient between the queue sizes, and the probability mass function of the difference between the queue sizes, are obtained. Extensive numerical results for various values of the system's parameters are presented, as well as a comparison between the current non-preemptive model and its twin system of preemptive service regime. One of the conclusions is that, depending on the variability of the various parameters, the preemptive regime is not necessarily more efficient than the non-preemptive one. Finally, economic issues are discussed and numerical comparisons are presented, showing the advantages and disadvantages of each regime.

Keywords Queueing · Join the Shortest Queue · Serve the Longest Queue · Polling · Probability generating functions · Matrix geometric

✉ Nir Perel
nirp@afeka.ac.il

[1] School of Industrial Engineering and Management, Afeka Tel Aviv Academic College of Engineering, Tel Aviv, Israel

[2] Department of Statistics and Operations Research, School of Mathematical Sciences, Tel Aviv University, Tel Aviv, Israel

Published online: 26 February 2024

1 Introduction

1.1 Background and contribution

The aim of the combined ‘Join the Shortest Queue’ (JSQ) and ‘Serve the Longest Queue’ (SLQ) service system is to equalize queue sizes. Each of the regimes, JSQ or SLQ, has been studied separately in the literature. Under the JSQ regime, the system is comprised of multiple separated queues, each with its specific service rate, while a newly arriving customer joins the shortest queue. In the SLQ model, a single server attends several queues, and always attempts to serve customers from the longest queue.

The combined JSQ–SLQ 2-queue Markovian system operating under the preemptive regime discipline with zero switch-over times has been recently introduced and analyzed in Perel et al. (2020). The investigation was further extended in Perel et al. (2022) for a 3-queue system.

The current paper broadens and generalizes the analysis of the 2-queue combined JSQ–SLQ system in three directions: (i) by analyzing the non-preemptive policy; (ii) by considering non-zero server's switch-over times, and (iii) by dealing with economic dichotomies to determine which and when one regime is preferable upon the other.

The non-preemptive policy implies that the server's switching decisions are made only upon service completions. Switching then occurs to a non-served queue only if its size is larger than the size of the queue attended by the server. Two versions are investigated: (i) non-preemptive with zero switch-over times, and (ii) non-preemptive with non-zero switch-over times. The JSQ–SLQ system with non-preemptive switching policy differs from the classical multi-class preemptive or non-preemptive priority models where class priority is fixed. In contrast, in the current model, the priority levels change dynamically, affected by changes in queue sizes.

An example of the combined JSQ–SLQ model, taken from the healthcare domain, is given in Perel et al. (2020), describing a medical clinic with a single operating physician and several treatment rooms, each with its dedicated medical assisting staff. A newly arriving patient is directed to the treatment room with the shortest queue. When the physician becomes available, s/he consistently visits the room having the longest queue. Service rates may differ between the treatment rooms.

Another example may be taken from the area of road transportation. Consider a road with two junctions apart from each other. There are two possible routes between the two junctions. At the first one, a traffic navigation application directs each arriving vehicle according to the JSQ policy. At the second junction, a traffic control mechanism gives extra ‘green-light time’ to the route with the longest queue.

A third example where the combined JSQ–SLQ model is applicable rises from the area of self-service cashiers in a supermarket. A paying customer joins the shortest queue, while a dedicated supermarket employee assists customers from the longest queue.

Main contributions A conventional formulation of the combined 2-queue JSQ–SLQ model leads to a two dimensional Markovian queueing system, where each dimension represents the queue size of the corresponding queue. Evidently, this 2-dimensional setting involves two un-bounded queues. For example, Flatto (1989), Avrachenkov et al. (2014) and Adan et al. (2016) applied boundary value problem technique to solve such problems, while Bright and Taylor (1995) applied a truncation method. In contrast, by using an unorthodox approach, we are able to derive the equilibrium joint probability mass function of the queue lengths. The innovation in our analysis is that rather than defining an un-bounded 2 dimensional state space describing the queue sizes, the analysis is based on a novel formulation that transforms the state-space into two dimensions, one finite, the other infinite, thus enabling the use of probability generating functions (PGFs) technique combined with a Matrix Geometric analysis. A second contribution is the analysis of the non-preemptive case, while a third contribution is the introduction of server's switch-over times. A fourth contribution is an economic analysis to determine which and when one regime is preferable upon the other.

1.2 Related work

Fixed priority models, as well as polling systems, have been studied extensively in the queueing literature (see e.g. Conway et al. (2003), Takagi (1986), Kella and Yechiali (1988), Yechiali (1993), Boon et al. (2011), along with the extensive references therein). Browne and Yechiali (1989) investigated a polling-type system with server's dynamic switching rules. Recently, Perel and Yechiali (2017) and Jolles et al. (2018) investigated a single-server two-queue systems where the server's switching decisions are threshold-based, depending on the evolving queue sizes.

A system with two $M/G/1$-type queues under the SLQ regime and non-preemptive discipline was analyzed by Cohen (1987). Flatto (1989) investigated the SLQ system with two identical queues and server's preemptive switching policy. A SLQ system with N symmetric queues and non-preemptive policy was analyzed by van Houtum et al. (1997). A Markovian non-symmetric 2-queue system was studied by Knessl and Yao (2013) under heavy traffic regime. A wireless network with SLQ mechanism was studied by Maguluri et al. (2014), while Pedarsani and Walrand (2016) studied an SLQ model for a multi-class network.

Queueing systems operating under the JSQ policy have also been studied extensively in the literature. Winston (1977) studied a system with Poisson arrivals and a finite number of identical servers, each serving its own queue, and proved that the discounted number of jobs completing service by some time t is maximized under the JSQ policy. This result was further extended by Hordijk and Koole (1990), who considered general arrival process, batch arrivals and finite buffers. Halfin (1985) derived bounds for the probability distribution of the number of customers in a JSQ system comprised of two identical servers. In Adan et al. (1991a), a JSQ system comprised of two non-symmetric servers was analyzed by using the compensation method (iterative approach), while a similar model with jockeying between the queues was studied in Adan et al. (1991b) via a matrix geometric approach. Furthermore, for the 2-queue JSQ system with Poisson arrival and two non-symmetric Exponential servers, Cohen (1998) derived explicitly the bivariate probability generating function of the stationary joint distribution of the queue lengths. This model was further analyzed in Adan et al. (2016), by using the compensation method and by solving a boundary value problem. van Houtum et al. (2001) studied a production system consisting of a group of parallel machines (servers) and multiple job types, where upon arrival, each job joins the shortest queue among all queues capable of serving it. Using numerical methods and truncation, the authors derive upper and lower bounds for the mean waiting time. Yao and Knessl (2005, 2006) studied a system comprised of two $M/M/\infty$ queues, in which new arrivals join the shortest queue. They applied asymptotic analysis for the bivariate generating function of the number of customers in each queue. A JSQ system operating under the Halfin-Whitt regime was studied by Eschenfeldt and Gamarnik (2018) and by Braverman (2020), where steady-state characteristics were obtained. A JSQ model with a large number of queues was studied in Dawson et al. (2019) and limiting results of various performance measures were derived. Dimitriou (2021) explored the stability and tail asymptotics of a Markovian single server retrial system with two infinite capacity orbits, where arriving customers that find the server busy join the shortest orbit queue. It should be emphasized that in the majority of the aforementioned JSQ models, a dedicated server is assigned to each queue. In deviation, in the current study, a unified JSQ–SLQ model is analyzed to examine a single-server polling-type system, where the server's transitions are from a shorter queue to a longer one, while incoming customers opt for the shortest queue.

Order of the paper and results This paper is organized as follows. Sections 2, 3 and 4 present and analyze the non-preemptive model with zero switch-over times. The system is first formulated in Sect. 2, while in Sect. 3 the partial probability generating functions of the system's states are derived, along with calculations of system's performance measures. In Sect. 4, the system's stability condition is derived via the matrix geometric method. Section 5 analyzes the case with non-zero switch-over times. Extensive numerical results are presented in Sect. 6, and the two regimes - preemptive vs. non-preemptive - are compared. Various insights are then drawn, one of which is that the preemptive regime is not necessarily more efficient than the non-preemptive one. A detailed economic analysis follows, assessing the relative advantages of the two regimes and identifies the parameter values under which one regime performs better than the other. Section 7 concludes the paper.

2 Zero switch-over times

Consider a single server polling-type system comprised of two asymmetrical queues, denoted by Q_1 and Q_2. The arrival process of customers to the system is Poisson with rate λ. Service duration of an arbitrary customer in Q_i is exponentially distributed with mean $1/\mu_i, i = 1, 2$. An arriving customer follows the 'Join the Shortest Queue' (JSQ) policy, i.e. s/he always joins the shortest queue, while if both queues are with equal lengths, the customer joins Q_i w.p. $p_i \geq 0$, where $p_1 + p_2 = 1$. The server resides in each queue according to a *non-preemptive* 'Serve the Longest Queue' (SLQ) policy. That is, if service is rendered to one of the queues, and the number of customers in the un-served queue exceeds the number of customers in the served queue, the server first finishes the current service, and only then switches to the other (un-served) queue, provided that the number of customers in the un-served queue is still greater than the number of customers in the served queue. Otherwise, the server remains in the current queue. Also, 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 next service completion, the customer will join Q_1 or Q_2 with probability p_1 and p_2, respectively, independently of the servers' position.

Let $L_i(t)$ denote the number of customers present in Q_i $(i = 1, 2)$ at time $t > 0$, and let $I(t)$ denote the server's position at that time. $I(t) = i$ implies that the server resides in Q_i. Under stability (see stability condition in the sequel), let $L_i = \lim_{t\to\infty} L_i(t)$ and $I = \lim_{t\to\infty} I(t)$. Define $D(t) = (L_1(t) - L_2(t))_{I(t)}$ and $D = (L_1 - L_2)_I$. In deviation from the preemptive JSQ–SLQ regime, where $L_1 - L_2$ may assume only the values (-1), (0) and (1), under the non-preemptive combined JSQ–SLQ policy, $L_1 - L_2$ assumes the values (-2), (-1), (0), (1) or (2). As a result from the model definition, $L_1 - L_2 = 2$ necessarily implies that $I = 1$, while $L_1 - L_2 = -2$ directly leads to $I = 2$. However, when $L_1 - L_2 = -1, 0, 1$, the position of the server, I, may assume any of the values 1 or 2, which leads to the form $1_i, 0_i, -1_i$, for $i = 1, 2$. Hence, overall, there are 8 values for D, as follows:

- $D = 2$: $L_1 - L_2 = 2$, server is at Q_1, an arriving customer joins Q_2.
- $D = 1_i$: $L_1 - L_2 = 1$, server is at Q_i $(i = 1, 2)$, an arriving customer joins Q_2.
- $D = 0_i$: $L_1 = L_2$, server is at Q_i $(i = 1, 2)$, an arriving customer joins Q_i w.p. p_i $(p_1 + p_2 = 1)$.
- $D = -1_i$: $L_1 - L_2 = -1$, server is at Q_i $(i = 1, 2)$, an arriving customer joins Q_1.
- $D = -2$: $L_1 - L_2 = -2$, server is at Q_2, an arriving customer joins Q_1.

The above system is now formulated as a two dimensional continuous time Markovian process, having state space $\{(n, d)\}$, for $n \geq 0$ and $d \in \mathfrak{D} = \{2, 1_1, 1_2, 0_1, 0_2, -1_1, -1_2, -2\}$, where n denotes the number of customers in Q_1. In a stable system, the steady state joint probability mass 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 construction of Fig. 1 deserves a further explanation. It is comprised of squares along five diagonals. Each square represents a possible state of (L_1, L_2). A square along the main diagonal has two possible states (n, n): one for the case where $D = 0_1$ (both queues equal in size and the server resides in Q_1), and $D = 0_2$ (the server resides in Q_2). Furthermore, each square on the diagonal above the main diagonal represents the cases where $D = -1_i$, where the index i $(i = 1, 2)$ indicates the position of the server. Finally, each square on the diagonal above the latter (the most north-east diagonal) represents the case where $D = -2$, while the diagonal below the main and the one below it represent the cases where $D = 1_i$ and $D = 2$, respectively.

3 Probability generating functions

This section is devoted to the derivation of the steady-state probability mass function of the 2-dimensional process (L_1, D). Following Perel et al. (2020), we use an unconventional construction of the probability generating functions (PGFs) and utilize

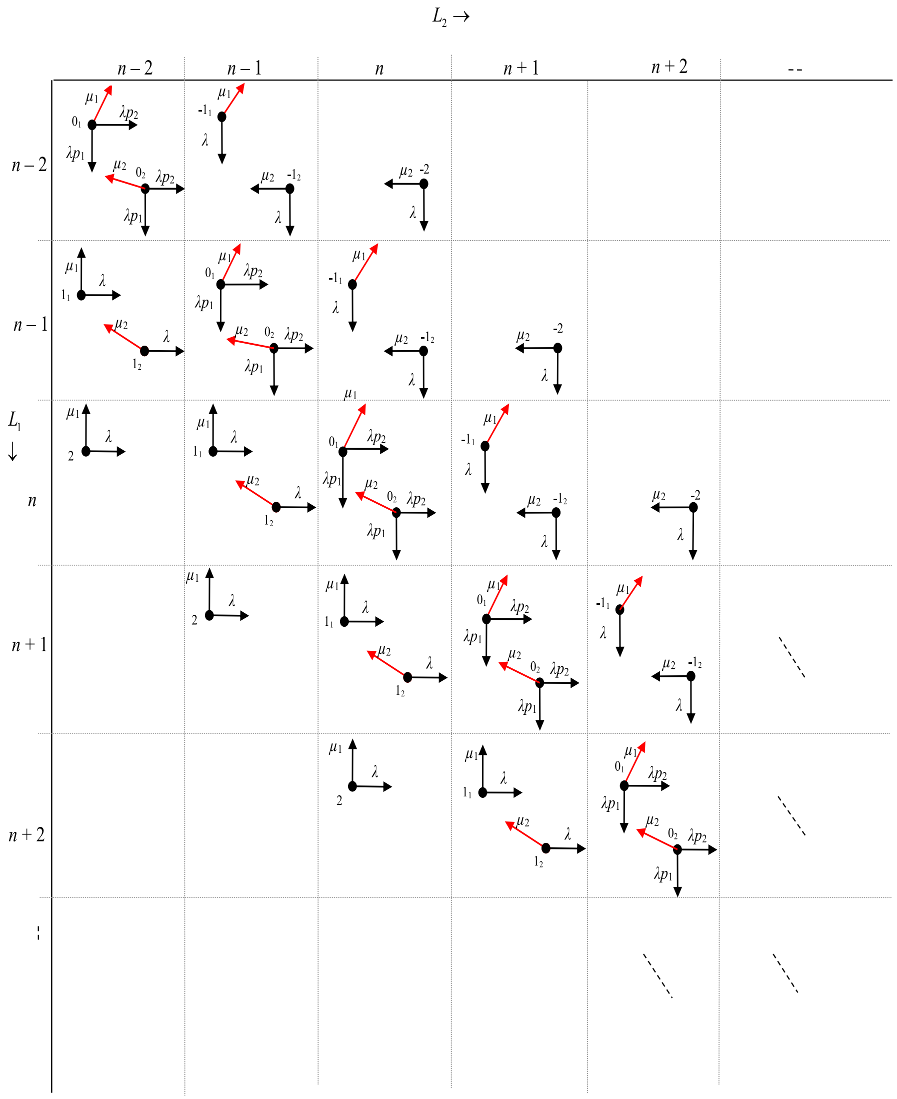


Fig. 1 Transition rate diagram of $(L_1(t), L_2(t))$

their properties. Specifically, instead of deriving PGFs along the horizontal axis, as is commonly done, we derive PGFs along the diagonals of Fig. 1.

3.1 Steady-state equations and corresponding PGFs

For each diagonal of Fig. 1, we write the balance equations for all states n along the diagonal. We first obtain the equations describing the states when the server resides in Q_1, as follows: (i) When $d = 2$ (lower diagonal),

$$(\lambda + \mu_1)P_{n,2} = \mu_2 P_{n,1_2},\ n \geq 2. \tag{1}$$

(ii) When $d = 1_1$ (states 1_1 along the diagonal below the main),

$$(\lambda + \mu_1)P_{1,1_1} = \lambda p_1 \left(P_{0,0_1} + P_{0,0_2}\right) + \mu_1 P_{2,2} + \mu_2 P_{1,0_2}, \tag{2}$$

$$(\lambda + \mu_1)P_{n,1_1} = \lambda p_1 P_{n-1,0_1} + \lambda P_{n,2} + \mu_1 P_{n+1,2} + \mu_2 P_{n,0_2},\ n \geq 2. \tag{3}$$

(iii) For $d = 0_1$ (states 0_1 along the main diagonal),

$$\lambda P_{0,0_1} = \mu_1 P_{1,1_1}, \tag{4}$$

$$(\lambda + \mu_1)P_{1,0_1} = \lambda P_{1,1_1} + \mu_1 P_{2,1_1}, \tag{5}$$

$$(\lambda + \mu_1)P_{n,0_1} = \lambda P_{n,1_1} + \lambda P_{n-1,-1_1} + \mu_1 P_{n+1,1_1},\ n \geq 2. \tag{6}$$

(iv) For $d = -1_1$ (states -1_1 along the diagonal above the main)

$$(\lambda + \mu_1)P_{n,-1_1} = \lambda p_2 P_{n,0_1}, n \geq 1. \tag{7}$$

In a similar manner, the balance equations for the case when the server resides in Q_2 are given by:

(v) When $d = 1_2$ (states 1_2 along the diagonal below the main),

$$(\lambda + \mu_2)P_{n,1_2} = \lambda p_1 P_{n-1,0_2},\ n \geq 2. \tag{8}$$

(vi) When $d = 0_2$ (states 0_2 along the main diagonal),

$$\lambda P_{0,0_2} = \mu_2 P_{0,-1_2}, \tag{9}$$

$$(\lambda + \mu_2)P_{1,0_2} = \lambda P_{0,-1_2} + \mu_2 P_{1,-1_2}, \tag{10}$$

$$(\lambda + \mu_2)P_{n,0_2} = \lambda P_{n-1,-1_2} + \lambda P_{n,1_2} + \mu_2 P_{n,-1_2},\ n \geq 2. \tag{11}$$

(vii) For $d = -1_2$ (states -1_2 along the diagonal above the main),

$$(\lambda + \mu_2)P_{0,-1_2} = \lambda p_2 \left(P_{0,0_1} + P_{0,0_2}\right) + \mu_2 P_{0,-2} + \mu_1 P_{1,0_1}, \tag{12}$$

$$(\lambda + \mu_2)P_{n,-1_2} = \lambda p_2 P_{n,0_2} + \lambda P_{n-1,-2} + \mu_2 P_{n,-2} + \mu_1 P_{n+1,0_1},\ n \geq 1. \tag{13}$$

$(viii)$ For $d = -2$ (upper diagonal),

$$(\lambda + \mu_2)P_{n,-2} = \mu_1 P_{n+1,-1_1},\ n \geq 0. \tag{14}$$

For each $d \in \mathfrak{D}$, the conditional PGF of the number of customers in Q_1 is defined as follows:

$$G_d(z) = \sum_n P_{n,d} z^n, \quad d \in \mathfrak{D},\ |z| \leq 1$$

Multiplying each of the Eqs. (1)–(14) by z^n (for the appropriate n), and summing over $n \geq 1$, we get a set of 8 equations for the probability generating functions, $((G_2(z), G_{1_1}(z), G_{0_1}(z), G_{-1_1}(z), G_{1_2}(z), G_{0_2}(z), G_{-1_2}(z), G_{-2}(z))$, as follows:

$$(\lambda + \mu_1)G_2(z) = \mu_2 G_{1_2}(z), \tag{15}$$

$$\begin{aligned}(\lambda + \mu_1)zG_{1_1}(z) = {} & (\lambda z + \mu_1)G_2(z) + \lambda p_1 z^2 G_{0_1}(z) + \mu_2 z G_{0_2}(z) \\ & - (\mu_2 - \lambda p_1)zP_{0,0_2},\end{aligned} \tag{16}$$

$$(\lambda + \mu_1)zG_{0_1}(z) = (\lambda z + \mu_1)G_{1_1}(z) + \lambda z^2 G_{-1_1}(z) + \mu_1 z P_{0,0_1}, \tag{17}$$

$$(\lambda + \mu_1)G_{-1_1}(z) = \lambda p_2 G_{0_1}(z) - \lambda p_2 P_{0,0_1}, \tag{18}$$

$$(\lambda + \mu_2)G_{1_2}(z) = \lambda p_1 z G_{0_2}(z) - \lambda p_1 P_{0,0_2}, \tag{19}$$

$$(\lambda + \mu_2)G_{0_2}(z) = \lambda G_{1_2}(z) + (\lambda z + \mu_2)G_{-1_2}(z) + \mu_2 P_{0,0_2}, \tag{20}$$

$$\begin{aligned}(\lambda + \mu_2)zG_{-1_2}(z) = {} & \mu_1 G_{0_1}(z) + \lambda p_2 z G_{0_2}(z) + (\lambda z + \mu_2)zG_{-2}(z) \\ & - (\mu_1 - \lambda p_2)zP_{0,0_1},\end{aligned} \tag{21}$$

$$(\lambda + \mu_2)zG_{-2}(z) = \mu_1 G_{-1_1}(z). \tag{22}$$

The set of 8 Eqs. (15)–(22) is condensed into a matrix form:

$$A(z) \cdot \vec{G}(z) = \vec{P}(z), \tag{23}$$

where

$$A(z) = \begin{pmatrix} \lambda + \mu_1 & 0 & 0 & 0 & -\mu_2 & 0 & 0 & 0 \\ -(\lambda z + \mu_1) & (\lambda + \mu_1)z & -\lambda p_1 z^2 & 0 & 0 & -\mu_2 z & 0 & 0 \\ 0 & (-\lambda z + \mu_1) & (\lambda + \mu_1)z & -\lambda z^2 & 0 & 0 & 0 & 0 \\ 0 & 0) & -\lambda p_2 & \lambda + \mu_1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \lambda + \mu_2 & -\lambda p_1 z & 0 & 0 \\ 0 & 0 & 0 & 0 & -\lambda & \lambda + \mu_2 & -(\lambda z + \mu_2) & 0 \\ 0 & 0 & -\mu_1 & 0 & 0 & -\lambda p_2 z & (\lambda + \mu_2)z & -(\lambda z + \mu_2)z \\ 0 & 0 & 0 & -\mu_1 & 0 & 0 & 0 & (\lambda + \mu_2)z \end{pmatrix},$$

$\vec{G}(z) = \left(G_2(z),\ G_{1_1}(z),\ G_{0_1}(z),\ G_{-1_1}(z),\ G_{1_2}(z),\ G_{0_2}(z),\ G_{-1_2}(z),\ G_{-2}(z)\right)^T$ is an 8-dimensional column vector of the desired PGFs, and

$$\begin{aligned}\vec{P}(z) = {} & ((0, -(\mu_2 - \lambda p_1 z)zP_{0,0_2}, \mu_1 z P_{0,0_1}, -\lambda p_2 P_{0,0_1}, -\lambda p_1 z P_{0,0_2}, \mu_2 P_{0,0_2}, \\ & - (\mu_1 - \lambda p_2 z)P_{0,0_1}, 0)^T\end{aligned}$$

is a column vector containing the two unknown ‘boundary probabilities’, $P_{0,0_1}$ and $P_{0,0_2}$.

By Cramer’s rule, $G_d(z) = \frac{|A_d(z)|}{|A(z)|}$, $d \in \mathfrak{D} = \{2, 1_1, 1_2, 0_1, 0_2, -1_1, -1_2, -2\}$, 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 $\vec{P}(z)$. Each of the PGFs $G_d(z)$, $d \in \mathfrak{D}$, is a function of the probabilities $P_{0,0_1}$ and $P_{0,0_2}$, appearing in $\vec{P}(z)$. One equation for the calculation of the boundary probabilities, is the normalization equation,

$$\sum_{d\in\mathfrak{D}} G_d(1) = \sum_{d\in\mathfrak{D}} \lim_{z\to 1} \frac{\mid A_d(z) \mid}{\mid A(z) \mid} = 1, \tag{24}$$

while a second relation between $P_{0,0_1}$ and $P_{0,0_2}$ is derived as follows. Since $G_d(z)$ is defined for all $|z| \le 1$, each root of $|A(z)|$ is a root of $|A_d(z)|$. The determinant $|A(z)|$ is a 6-degree polynomial which can be expressed as $|A(z)| = z^2(1-z)h(z)$, where

$$h(z) = -\, z^3\alpha_3 + z^2\alpha_2 - z\alpha_1 - \alpha_0, \tag{25}$$

and

$$\begin{aligned}
\alpha_0 &= \mu_1^2\mu_2^2\left[(\lambda+\mu_1)^2(\lambda+\mu_2)^2+\lambda(\lambda+\mu_1)(\lambda+\mu_2)(\mu_1p_1+\mu_2p_2)+\lambda^2\mu_1\mu_2p_1p_2\right],\\
\alpha_1 &= \lambda\mu_1\mu_2(\lambda+\mu_1)^2(\lambda+\mu_2)^2(\mu_1+\mu_2) + \mu_1^2\mu_2^2(\lambda+\mu_1)^2(\lambda+\mu_2)^2\\
&\quad + \lambda^2\mu_1^2\mu_2^2(\lambda+\mu_1)(\lambda+\mu_2)+\lambda^2\mu_1\mu_2(\lambda+\mu_1)(\lambda+\mu_2)(\mu_1p_1+\mu_2p_2)(\mu_1+\mu_2)\\
&\quad + \lambda\mu_1^2\mu_2^2(\lambda+\mu_1)(\lambda+\mu_2)(\mu_1p_1+\mu_2p_2)+\lambda^2\mu_1^3\mu_2^3p_1p_2\\
&\quad + 2\lambda^3\mu_1^2\mu_2^2p_1p_2(\mu_1+\mu_2),\\
\alpha_2 &= \lambda^3(\lambda+\mu_1)^2(\lambda+\mu_2)^2(\mu_1(1+p_2)+\mu_2(1+p_1)) + \lambda^4(\lambda+\mu_1)^2(\lambda+\mu_2)^2\\
&\quad + \lambda^2(\lambda+\mu_1)^2(\lambda+\mu_2)^2(\mu_1^2+\mu_2^2) + \lambda^4\mu_1\mu_2(\lambda+\mu_1)(\lambda+\mu_2)\\
&\quad + 2\lambda^5\mu_1\mu_2p_1p_2(\mu_1+\mu_2) + \lambda^6\mu_1\mu_2p_1p_2,\\
\alpha_3 &= \lambda^4\left[(\lambda+\mu_1)^2(\lambda+\mu_2)^2 - \lambda^2\mu_1\mu_2p_1p_2\right].
\end{aligned}$$

Note that the elements α_i, $i = 0, \ldots, 3$, are all positive and symmetric in μ_1 and μ_2 and in p_1 and p_2, as expected. The cubic polynomial $h(z)$ possesses 3 roots, denoted by z_k, $k = 1, 2, 3$, that can be expressed explicitly by solving a cubic formula. Since $h(z) > 0$ for all $z < -1$ and $h(0) < 0$, then there is at least one root in the interval $(-1, 0)$. By applying Descartes’ rule of signs on $h(-z)$ (see e.g. Curtiss 1918), we conclude that $h(z)$ possesses a single root in $(-1, 0)$, which we use as a second relation between the looked for boundary probabilities, $P_{0,0_1}$ and $P_{0,0_2}$. It can be verified that, under stability, the two other roots are greater than 1, and therefore not relevant. Furthermore, the necessary and sufficient condition for the stability is is $h(1) < 0$, and it is shown in Sect. 4.2 below, by using matrix geometric analysis. From all the above, $P_{0,0_1}$ and $P_{0,0_2}$ can be calculated, which provides us with expressions for the PGFs, $G_d(z)$, for all $d \in \mathfrak{D}$.

3.2 Performance measures

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

$$\begin{aligned}
P_{\bullet d} &= \mathbb{P}(D = d) = \sum_{n=0}^{\infty} P_{n,d} = G_d(1), \quad d \in \mathfrak{D}, \\
P_{n\bullet} &= \mathbb{P}(L_1 = n) = \sum_{d \in \mathfrak{D}} P_{n,d}\ , \quad n \geq 0.
\end{aligned}$$

Then,

$$\begin{aligned}
\mathbb{E}[L_1] &= \sum_{n=0}^{\infty} n P_{n\bullet} = \sum_{d \in \mathfrak{D}} G_d'(1), \\
\mathbb{E}[D] &= \mathbb{E}[L_1 - L_2] = \sum_{d \in \mathfrak{D}} d G_d(1) \\
&\quad = 2G_2(1) + G_{1_1}(1) + G_{1_2}(1) - G_{-1_1}(1) - G_{-1_2}(1) - 2G_{-2}(1) \\
&\quad = 2(P_{\bullet 2} - P_{\bullet(-2)}) + (P_{\bullet 1_1} - P_{\bullet(-1_1)}) + (P_{\bullet 1_2} - P_{\bullet(-1_2)}), \\
\mathbb{E}[L_2] &= \mathbb{E}[L_1] - \mathbb{E}[D].
\end{aligned}$$

Furthermore,

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

where

$$\begin{aligned}
\mathbb{E}[L_1^2] &= \sum_{d \in \mathfrak{D}} G_d''(1) + \mathbb{E}[L_1], \\
\mathbb{E}[L_1 D] &= \sum_{d \in \mathfrak{D}} \sum_{n} n d P_{n,d} = 2\sum_{n} n P_{n,2} - 2\sum_{n} n P_{n,-2} + \sum_{n} n P_{n,1_1} - \sum_{n} n P_{n,-1_1} \\
&\quad + \sum_{n} n P_{n,1_2} - \sum_{n} n P_{n,-1_2} \\
&= 2\left[G_2'(1) - G_{-2}'(1)\right] + G_{1_1}'(1) - G_{-1_1}'(1) + G_{1_2}'(1) - G_{-1_2}'(1).
\end{aligned}$$

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

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

where $\mathbb{E}[D^2] = \sum_{d \in \mathfrak{D}} d^2 P_{\bullet d} = 4(P_{\bullet 2} + P_{\bullet(-2)}) + P_{\bullet 1_1} + P_{\bullet(-1_1)} + P_{\bullet 1_2} + P_{\bullet(-1_2)}$.

As a result, 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 Sect. 6).

The effective arrival rate to Q_i $(i = 1, 2)$, λ^i_{eff}, is given by

$$\lambda^1_{eff} = \lambda\Big(p_1(P_{\bullet 0_1} + P_{\bullet 0_2}) + P_{\bullet(-1_1)} + P_{\bullet(-1_2)} + P_{\bullet(-2)}\Big),$$
$$\lambda^2_{eff} = \lambda\Big(p_2(P_{\bullet 0_1} + P_{\bullet 0_2}) + P_{\bullet 1_1} + P_{\bullet 1_2} + P_{\bullet 2}\Big).$$

Clearly, $\lambda^1_{eff} + \lambda^2_{eff} = \lambda$. The effective rate of work flowing into Q_i is $\rho^i_{eff} = \frac{\lambda^i_{eff}}{\mu_i}$. We thus have,

$$\mathbb{P}(\text{Server is idle}) = P_{0,0_1} + P_{0,0_2} = 1 - \rho^1_{eff} - \rho^2_{eff}.$$

By Little's Law, the mean sojourn time of a customer in Q_i, for $i = 1, 2$, is given by

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

4 Matrix geometric

4.1 Formulation

The system is formulated as a Quasi Birth and Death (QBD) process with 8 phases and an un-bounded number of levels, where phase d corresponds to $D = d$, for $d \in \mathfrak{D}$, and each level n corresponds to $L_1 = n$, the total number of customers in Q_1. The connections between the PGF method and the matrix geometric approach have been investigated in several papers (see e.g. Perel and Yechiali 2013a, b; Paz and Yechiali 2014; Perel and Yechiali 2017; Phung-Duc 2017; Armony et al. 2019; Hanukov and Yechiali 2021).

For $n \geq 2$ define $\mathcal{S}_n$ to be the set of states

$\mathcal{S}_n = \{(n, 2), (n, 1_1), (n, 0_1), (n, -1_1), (n, 1_2), (n, 0_2), (n, -1_2), (n, -2)\}$, and arrange the system's states in the order

$$\mathcal{S} = \Big\{(0, 0_1), (0, 0_2), (0, -1_2), (0, -2); (1, 1_1), (1, 0_1), (1, -1_1), (1, 0_2), (1, -1_2),$$
$$(1, -2); \mathcal{S}_2; \mathcal{S}_3; \ldots; \mathcal{S}_n \ldots\Big\}.$$

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

$$Q = \begin{pmatrix} B_1^0 & B_0^0 & 0 & \cdots & \cdots & \cdots & \cdots & \cdots \\ B_2^1 & B_1^1 & B_0^1 & 0 & \cdots & \cdots & \cdots & \cdots \\ 0 & B_2 & A_1 & A_0 & 0 & \cdots & \cdots & \cdots \\ 0 & 0 & A_2 & A_1 & A_0 & 0 & \cdots & \cdots \\ 0 & 0 & 0 & A_2 & A_1 & A_0 & 0 & \cdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \end{pmatrix},$$

where

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

$$B_2^1 = \begin{pmatrix} \mu_1 & 0 & 0 & 0 \\ 0 & 0 & \mu_1 & 0 \\ 0 & 0 & 0 & \mu_1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

$$B_1^1 = \begin{pmatrix} -(\lambda+\mu_1) & \lambda & 0 & 0 & 0 & 0 \\ 0 & -(\lambda+\mu_1) & \lambda p_2 & 0 & 0 & 0 \\ 0 & 0 & -(\lambda+\mu_1) & 0 & 0 & 0 \\ \mu_2 & 0 & 0 & -(\lambda+\mu_2) & \lambda p_2 & 0 \\ 0 & 0 & 0 & \mu_2 & -(\lambda+\mu_2) & 0 \\ 0 & 0 & 0 & 0 & \lambda & -(\lambda+\mu_2) \end{pmatrix},$$

$$B_0^1 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \lambda p_1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \lambda & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \lambda p_1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \lambda & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \lambda & 0 \end{pmatrix}, \quad B_2 = \begin{pmatrix} \mu_1 & 0 & 0 & 0 & 0 & 0 \\ 0 & \mu_1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \mu_1 & 0 \\ 0 & 0 & 0 & 0 & 0 & \mu_1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

and

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

$$A_0 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \lambda p_1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \lambda & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \lambda p_1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \lambda & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \lambda & 0 \end{pmatrix}, \quad A_2 = \begin{pmatrix} 0 & \mu_1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \mu_1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \mu_1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mu_1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

4.2 Stability condition

Define the matrix $A = A_0 + A_1 + A_2$. Then,

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

The matrix A is the infinitesimal generator matrix of the process describing the evolution of D, given that $L_1 \geq 2$. 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 an 8-dimensional column vector with all its entries equal to 1). From Neuts (1981), we have that the stability condition is

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

which, after tedious algebra, translates into a 5-degree polynomial in λ,

$$\begin{aligned} g(\lambda) = & \lambda^5 \left(\mu_1(p_1-2) + \mu_2(p_2-2)\right) \\ & + \lambda^4 \left(p_1\mu_1^2 + p_2\mu_2^2 - 3\mu_1^2 - 3\mu_2^2 - 4\mu_1\mu_2 - 2p_1p_2\mu_1\mu_2\right) \\ & - \lambda^3 \left(\mu_1^3 + \mu_2^3 + \mu_1^2\mu_2 + \mu_1\mu_2^2 + p_1\mu_1\mu_2^2 + p_2\mu_1^2\mu_2\right) \\ & + \lambda^2 \left(p_1\mu_1^3\mu_2 + p_2\mu_1\mu_2^3 + 7\mu_1^2\mu_2^2 - (p_1^2 + p_2^2)\mu_1^2\mu_2^2\right) \\ & + \lambda \left(3\mu_1^3\mu_2^2 + 3\mu_1^2\mu_2^3 + 2p_1\mu_1^3\mu_2^2 + 2p_2\mu_1^2\mu_2^3\right) + 2\mu_1^3\mu_2^3 > 0. \end{aligned} \tag{26}$$

Note that Eq. (26) is symmetric in all of the system's parameters, i.e. λ, μ_1, μ_2, p_1, p_2. It can also be verified that Eq. (26) implies $h(1) < 0$, where $h(z)$ is defined in Eq. (25). In order to explore the roots of $g(\lambda)$, we utilize again Descartes' rule of signs. Since the coefficients of λ^5, λ^4 and λ^3 are negative, while all other coefficients of λ^i, for $i = 0, 1, 2$ are positive, there is a single change of signs between the coefficients. Therefore, there is a single positive root of this polynomial, which we denote by $\lambda_0 = f(\mu_1, \mu_2, p_1, p_2)$. Since $g(0) = 2\mu_1^3\mu_2^3 > 0$, and $g(\infty) < 0$, the system is stable iff

$$\lambda < \lambda_0. \tag{27}$$

In the symmetric case, where $\mu_1 = \mu_2 = \mu$, the stability condition (26) translates into

$$\mu(\mu - \lambda)(\mu + \lambda)\left(3\lambda^3 + 9\lambda^2\mu + 2p_1p_2\lambda^2\mu + 8\lambda\mu^2 + 2\mu^3\right) > 0,$$

or, equivalently,

$$\lambda < \mu.$$

This holds for any value of p_1, since when both service rates are equal, 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 The equilibrium distribution

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

$$\vec{P}_n = \begin{cases} \left(P_{0,0_1}, P_{0,0_2}, P_{0,-1_2}, P_{0,-2}\right), & n = 0, \\ \left(P_{1,1_1}, P_{1,0_1}, P_{1,-1_1}, P_{1,0_2}, P_{1,-1_2}, P_{1,-2}\right), & n = 1, \\ \left(P_{n,2}, P_{n,1_1}, P_{n,0_1}, P_{n,-1_1}, P_{n,1_2}, P_{n,0_2}, P_{n,-1_2}, P_{n,-2}\right), & n \geq 2 \end{cases}$$

From Neuts (1981),

$$\vec{P}_n = \vec{P}_2 R^{n-2}, \quad n \geq 2,$$

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

$$A_0 + RA_1 + R^2A_2 = 0. \tag{28}$$

The vectors $\vec{P}_0$, $\vec{P}_1$ and $\vec{P}_2$ are determined by the following linear system:

$$\begin{aligned} &\vec{P}_0 B_1^0 + \vec{P}_1 B_2^1 = \vec{0}, \\ &\vec{P}_0 B_0^0 + \vec{P}_1 B_1^1 + \vec{P}_2 B_2 = \vec{0}, \\ &\vec{P}_1 B_0^1 + \vec{P}_2 A_1 + \vec{P}_2 R A_2 = \vec{0}, \\ &\vec{P}_0 \vec{e}_0 + \vec{P}_1 \vec{e}_1 + \vec{P}_2 [\mathbf{I} - R]^{-1} \vec{e} = 1, \end{aligned} \tag{29}$$

where $\vec{e}_0$ and $\vec{e}_1$ are 4-dimensional and 6-dimensional, respectively, vectors of 1's. Equation (29) is the normalization equation.

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

$$\begin{aligned}\mathbb{E}[L_1] &= \sum_{n=1}^{\infty} n\vec{P}_n\vec{e} = \vec{P}_1\vec{e}_1 + \sum_{n=2}^{\infty} n\vec{P}_2 R^{n-2}\vec{e}\\ &= \vec{P}_1\vec{e}_1 + \sum_{n=2}^{\infty}(n-1)\vec{P}_2 R^{n-2}\vec{e} + \sum_{n=2}^{\infty}\vec{P}_2 R^{n-2}\vec{e}\\ &= \vec{P}_1\vec{e}_1 + \vec{P}_2\left([\mathbf{I}-R]^{-2}+[\mathbf{I}-R]^{-1}\right)\vec{e}.\end{aligned}$$

4.4 Characterization of the rate matrix *R*

Equation (28) for the calculation of the matrix $R=\left[r_{i,j}\right]$ for $i, j = 1,\ldots,8$ involve 64 non-linear equations with 64 variables. Algorithms for calculatiog R can be found in Neuts (1981), Latouche and Ramaswami (1999), Artalejo and Gómez-Corral (2008), and Harchol-Balter (2013). However, it is possible to characterize some of the properties of R. 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. We recall that 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, second and fifth rows of A_0 are all zeros, all elements in the corresponding rows of R are zeros as well. That is, $r_{i,j} = 0$ for $i = 1, 2, 5$ and $j = 1,\ldots,8$. 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. (28), 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. We also refer to Hanukov and Yechiali (2021), where it is shown that when the matrices A_0, A_1, and A_2 are all lower (or all upper) triangular, the elements of R can be explicitly calculated and the stability condition is readily obtained.

5 Non-zero switch-over times

In this section we present the analysis for the 2-queue JSQ–SLQ system with non-zero switch-over times. Specifically, we assume that the server's switching time from Q_1 (Q_2) to Q_2 (Q_1) is exponentially distributed with mean $1/\gamma_2$ ($1/\gamma_1$). We also assume that once the server switches from Q_i to Q_j, it completes the switch and the service in Q_j, regardless the size of Q_i. After completing the service in Q_i, the server continues to operate according to the non-preemptive SLQ policy. The existence of switch-over times enlarges the number of possible states, such that the set $\mathfrak{D}$ consists now 16 states. Recall that the states in $\mathfrak{D}$ describe the difference $L_1 - L_2$, as well as the location of the server, which, in the case with switch-over times, is not only a specific queue, but may also describe a transition phase of the server from Q_i to Q_j. The various states are given as follows:

- $D = 2$: Server is at Q_1, an arriving customer joins Q_2.
- $D = 2_{s_1}$: Server switches to Q_1, an arriving customer joins Q_2.
- $D = 1_i$: Server is at Q_i $(i = 1, 2)$, an arriving customer joins Q_2.
- $D = 1_{s_i}$: Server switches to Q_i $(i = 1, 2)$, an arriving customer joins Q_2.
- $D = 0_i$: Server is at Q_i $(i = 1, 2)$, an arriving customer joins Q_i w.p. p_i $(p_1+p_2 = 1)$.
- $D = 0_{s_i}$: Server switches to Q_i $(i = 1, 2)$, an arriving customer joins Q_i w.p. p_i $(p_1 + p_2 = 1)$.
- $D = -1_i$: Server is at Q_i $(i = 1, 2)$, an arriving customer joins Q_1.
- $D = -1_{s_i}$: Server switches to Q_i $(i = 1, 2)$, an arriving customer joins Q_1.
- $D = -2$: Server is at Q_2, an arriving customer joins Q_1.
- $D = -2_{s_2}$: Server switches to Q_2, an arriving customer joins Q_1.

Similarly, as done in Sect. 3, we write the balance equations of the corresponding system. The balance equations for the case when the server resides in Q_1 or switches to Q_1 are given by

When $d = 2$,

$$(\lambda + \mu_1)P_{n,2} = \gamma_1 P_{n,2_{s_1}},\ n \geq 2. \tag{30}$$

When $d = 2_{s_1}$,

$$(\lambda + \gamma_1)P_{n,2_{s_1}} = \mu_2 P_{n,1_2},\ n \geq 2. \tag{31}$$

When $d = 1_1$,

$$(\lambda + \mu_1)P_{1,1_1} = \lambda p_1 P_{0,0_1} + \mu_1 P_{2,2} + \gamma_1 P_{1,1_{s_1}}, \tag{32}$$

$$(\lambda + \mu_1)P_{n,1_1} = \lambda p_1 P_{n-1,0_1} + \lambda P_{n,2} + \mu_1 P_{n+1,2} + \gamma_1 P_{n,1_{s_1}},\ n \geq 2. \tag{33}$$

When $d = 1_{s_1}$,

$$(\lambda + \gamma_1)P_{1,1_{s_1}} = \lambda p_1 P_{0,0_2} + \mu_2 P_{1,0_2}, \tag{34}$$

$$(\lambda + \gamma_1)P_{n,1_{s_1}} = \lambda P_{n,2_{s_1}} + \mu_2 P_{n,0_2} + \lambda p_1 P_{n-1,0_{s_1}},\ n \geq 2. \tag{35}$$

For $d = 0_1$,

$$\lambda P_{0,0_1} = \mu_1 P_{1,1_1}, \tag{36}$$

$$(\lambda + \mu_1)P_{1,0_1} = \lambda P_{1,1_1} + \mu_1 P_{2,1_1} + \gamma_1 P_{1,0s_1}, \tag{37}$$

$$(\lambda + \mu_1)P_{n,0_1} = \lambda P_{n,1_1} + \lambda P_{n-1,-1_1} + \mu_1 P_{n+1,1_1} + \gamma_1 P_{n,0s_1},\ n \geq 2. \tag{38}$$

For $d = 0_{s_1}$,

$$(\lambda + \gamma_1)P_{1,0_{s_1}} = \lambda P_{1,1_{s_1}}, \tag{39}$$

$$(\lambda + \gamma_1)P_{n,0_{s_1}} = \lambda P_{n,1_{s_1}} + \lambda P_{n-1,-1_{s_1}},\ n \geq 2. \tag{40}$$

For $d=-1_1$,

$$(\lambda+\mu_1)P_{n,-1_1}=\lambda p_2 P_{n,0_1}+\gamma_1 P_{n,-1_{s_1}}, n\geq 1. \tag{41}$$

For $d=-1_{s_1}$,

$$(\lambda+\gamma_1)P_{n,-1_{s_1}}=\lambda p_2 P_{n,0_{s_1}}, n\geq 1. \tag{42}$$

Furthermore, the balance equations for the case when the server resides in Q_2 or switches to Q_2 are given by:

When $d=1_{s_2}$,

$$(\lambda+\gamma_2)P_{n,1_{s_2}}=\lambda p_1 P_{n-1,0_{s_2}},\ n\geq 2. \tag{43}$$

When $d=1_2$,

$$(\lambda+\mu_2)P_{n,1_2}=\lambda p_1 P_{n-1,0_2}+\gamma_2 P_{n,1_{s_2}},\ n\geq 2. \tag{44}$$

When $d=0_{s_2}$,

$$(\lambda+\gamma_2)P_{1,0_{s_2}}=\lambda P_{0,-1_{s_2}}, \tag{45}$$

$$(\lambda+\gamma_2)P_{n,0_{s_2}}=\lambda P_{n-1,-1_{s_2}}+\lambda P_{n,1_{s_2}},\ n\geq 2. \tag{46}$$

When $d=0_2$,

$$\lambda P_{0,0_2}=\mu_2 P_{0,-1_2}, \tag{47}$$

$$(\lambda+\mu_2)P_{1,0_2}=\lambda P_{0,-1_2}+\mu_2 P_{1,-1_2}+\gamma_2 P_{1,0_{s_2}}, \tag{48}$$

$$(\lambda+\mu_2)P_{n,0_2}=\lambda P_{n-1,-1_2}+\lambda P_{n,1_2}+\mu_2 P_{n,-1_2}+\gamma_2 P_{n,0_{s_2}},\ n\geq 2. \tag{49}$$

For $d=-1_{s_2}$,

$$(\lambda+\gamma_2)P_{0,-1_{s_2}}=\lambda p_2 P_{0,0_1}+\mu_1 P_{1,0_1}, \tag{50}$$

$$(\lambda+\gamma_2)P_{n,-1_{s_2}}=\lambda p_2 P_{n,0_{s_2}}+\lambda P_{n-1,-2_{s_2}}+\mu_1 P_{n+1,0_1},\ n\geq 1. \tag{51}$$

For $d=-1_2$,

$$(\lambda+\mu_2)P_{0,-1_2}=\lambda p_2 P_{0,0_2}+\mu_2 P_{0,-2}+\gamma_2 P_{0,-1_{s_2}}, \tag{52}$$

$$(\lambda+\mu_2)P_{n,-1_2}=\lambda p_2 P_{n,0_2}+\lambda P_{n-1,-2}+\mu_2 P_{n,-2}+\gamma_2 P_{n,-1_{s_2}},\ n\geq 1. \tag{53}$$

For $d=-2_{s_2}$,

$$(\lambda+\gamma_2)P_{n,-2_{s_2}}=\mu_1 P_{n+1,-1_1},\ n\geq 0. \tag{54}$$

For $d = -2$,

$$(\lambda + \mu_2)P_{n,-2} = \gamma_2 P_{n,-2_{s_2}},\ n \geq 0. \tag{55}$$

The set of 16 PGFs is derived as follows:

$$(\lambda + \mu_1)G_2(z) = \gamma_1 G_{2_{s_1}}(z), \tag{56}$$

$$(\lambda + \gamma_1)G_{2_{s_1}}(z) = \mu_2 G_{1_2}(z), \tag{57}$$

$$(\lambda + \mu_1)zG_{1_1}(z) = \lambda p_1 z^2 G_{0_1}(z) + (\lambda z + \mu_1)G_2(z) + \gamma_1 z G_{1_{s_1}}(z) \tag{58}$$

$$(\lambda + \gamma_1)G_{1_{s_1}}(z) = \lambda G_{2_{s_1}}(z) + \mu_2 G_{0_2}(z) + \lambda p_1 z G_{0_{s_1}}(z) - (\mu_2 - \lambda p_1 z)P_{0,0_2} \tag{59}$$

$$(\lambda + \mu_1)zG_{0_1}(z) = (\lambda z + \mu_1)G_{1_1}(z) + \lambda z^2 G_{-1_1}(z) + \gamma_1 z G_{0_{s_1}}(z) + \mu_1 z P_{0,0_1}, \tag{60}$$

$$(\lambda + \gamma_1)G_{0_{s_1}}(z) = \lambda G_{1_{s_1}}(z) + \lambda z G_{-1_{s_1}}(z), \tag{61}$$

$$(\lambda + \mu_1)G_{-1_1}(z) = \lambda p_2 G_{0_1}(z) + \gamma_1 G_{-1_{s_1}}(z) - \lambda p_2 P_{0,0_1}, \tag{62}$$

$$(\lambda + \gamma_1)G_{-1_{s_1}}(z) = \lambda p_2 G_{0_{s_1}}(z), \tag{63}$$

$$(\lambda + \gamma_2)G_{1_{s_2}}(z) = \lambda p_1 z G_{0_{s_2}}(z), \tag{64}$$

$$(\lambda + \mu_2)G_{1_2}(z) = \lambda p_1 z G_{0_2}(z) + \gamma_2 G_{1_{s_2}}(z) - \lambda p_1 z P_{0,0_2}, \tag{65}$$

$$(\lambda + \gamma_2)G_{0_{s_2}}(z) = \lambda z G_{-1_{s_2}}(z) + \lambda G_{1_{s_2}}(z), \tag{66}$$

$$(\lambda + \mu_2)G_{0_2}(z) = \lambda G_{1_2}(z) + (\lambda z + \mu_2)G_{-1_2}(z) + \gamma_2 G_{0_{s_2}}(z) + \mu_2 P_{0,0_2}, \tag{67}$$

$$(\lambda + \gamma_2)zG_{-1_{s_2}}(z) = \mu_1 G_{0_1}(z) + \lambda p_2 z G_{0_{s_2}}(z) + \lambda z^2 G_{-2_{s_2}}(z) - (\mu_1 - \lambda p_2 z)P_{0,0_1}, \tag{68}$$

$$(\lambda + \mu_2)G_{-1_2}(z) = \lambda p_2 G_{0_2}(z) + (\lambda z + \mu_2)G_{-2}(z) + \gamma_2 G_{-1_{s_2}}(z), \tag{69}$$

$$(\lambda + \gamma_2)zG_{-2_{s_2}}(z) = \mu_1 G_{-1_1}(z), \tag{70}$$

$$(\lambda + \mu_2)G_{-2}(z) = \gamma_2 G_{-2_{s_2}}(z). \tag{71}$$

Note that in order to fully obtain the above PGFs that satisfy Eqs. (56)–(71), only two probabilities, $P_{0,0_1}$ and $P_{0,0_2}$, need to ba calculated. The derivation of these probabilities is done similarly to the way described in Sect. 3.1. Furthermore, all performance measures calculated in Sect. 3.2 for the case without switch-over times can be derived in a similar manner for the model with switch-over times. Therefore, we omit these calculations from the paper. Nevertheless, in Sect. 6 we present numerical results for various performance measures of this model.

6 Numerical results

In this section we compare the performance measures of the current *non-preemptive* JSQ–SLQ model with the corresponding measures of the *preemptive* JSQ–SLQ queueing system studied in Perel et al. (2020). For each model, we present numerical results of the system's performance measures for a wide range of parameters. Note that all calculations are based on the analytical results and derived in the paper. There are two sets of tables: Tables 1, 2, 3, 4 and 5, comparing various performance measures, and Tables 6, 7, 8, 9 and 10, dealing with the probability distribution of D. Furthermore, numerical results for the case with server's switch-over times are provided in Tables 12 and 13.

6.1 Comparison between preemptive and non-preemptive JSQ–SLQ models without switch-over times

Tables 1, 2, 3, 4 and 5 present sets of results, where the calculated 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, 2, 3 and 4, respectively, while $p_1 = \frac{\mu_1}{\mu_1+\mu_2}$ in Table 5. In each table, the values of μ_1 vary between 3.3 and 100. For each value of μ_1, the first row describes results for the preemptive model, while the second row presents results for the non-preemptive case. Furthermore, for each case, we present the stability condition in terms of λ, according to the values of μ_1, μ_2, p_1 and p_2 (see Eq. 27).

6.1.1 Insights from Tables 1, 2, 3, 4 and 5

1. In all tables, when $\mu_2 = 5$, $\mathbb{E}[L_1]$ and $\mathbb{E}[L_2]$, as well as $\mathbb{E}[W_1]$ and $\mathbb{E}[W_2]$, are all decreasing functions of μ_1.
2. When $p_1 \leq 0.5$ (Tables 1 and 2) and for small values of μ_1, the values for $\mathbb{E}[L_i]$ and $\mathbb{E}[W_i]$ $(i = 1, 2)$ in the non-preemptive model are significantly larger than the corresponding measures in the preemptive case. However, as μ_1 increases, the differences between $\mathbb{E}[L_i]$, $i = 1, 2$, and between $\mathbb{E}[W_i]$ in both models become negligible.
3. When $p_1 > 0.5$ (Tables 3 and 4) and for small values of μ_1, the results for $\mathbb{E}[L_i]$ and for $\mathbb{E}[W_i]$ in the non-preemptive model are smaller than the corresponding results in the preemptive case. This can be explained since whenever $p_1 > 0.5$, in case when $L_1 = L_2$, arriving customers are more likely to join Q_1 rather than Q_2. In the preemptive case, when μ_1 is relatively small (large mean service time), the server switches to Q_2 immediately when $L_2 > L_1$, even before completing service at Q_1. While the server resides at Q_2, customers arrive at Q_1, causing L_1 to exceed L_2, so the server immediately switches back to Q_1, and so on. Note that μ_2 is not relatively large, therefore, the server rapidly alternates between the queues, significant amount of times, without completing service. On the other hand, in the non-preemptive case, the server first completes service, and only then, if needed, switches to the other queue. Intuitively, one may think that switching over in the middle of a service of a job (under a non-resume policy) may always be worse than the non-preemptive case. However, the above results show that this is not always the case, and, depending on the parameter values, switching may prove beneficial.

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

μ_1	$\mathbb{E}[L_1]$	$\mathbb{E}[L_2]$	$\mathbb{E}[W_1]$	$\mathbb{E}[W_2]$	λ_{eff}^1	λ_{eff}^2	ρ_{eff}^1	ρ_{eff}^2	$Cor(L_1, L_2)$	stab. cond
3.3	9.49	9.53	6.52	3.75	1.46	2.54	0.44	0.51	0.9975	$\lambda < 4.2$
	22.52	22.67	13.14	9.92	1.71	2.29	0.52	0.46	0.9990	$\lambda < 4.08$
3.5	6.23	6.28	4.24	2.48	1.47	2.53	0.42	0.51	0.9943	$\lambda < 4.31$
	9.22	9.38	5.42	4.08	1.70	2.30	0.49	0.46	0.9945	$\lambda < 4.21$
4	3.46	3.53	2.31	1.41	1.50	2.50	0.37	0.50	0.9836	$\lambda < 4.56$
	3.85	4.00	2.29	1.72	1.68	2.32	0.42	0.46	0.9715	$\lambda < 4.50$
4.5	2.46	2.55	1.62	1.03	1.52	2.48	0.34	0.49	0.9708	$\lambda < 4.79$
	2.52	2.67	1.51	1.14	1.66	2.34	0.37	0.47	0.9403	$\lambda < 4.76$
5	1.95	2.05	1.27	0.83	1.54	2.46	0.31	0.49	0.9575	$\lambda < 5$
	1.92	2.08	1.16	0.88	1.65	2.35	0.33	0.47	0.9084	$\lambda < 5$
8	0.98	1.14	0.61	0.48	1.60	2.40	0.20	0.48	0.8927	$\lambda < 5.98$
	0.95	1.12	0.58	0.48	1.63	2.37	0.20	0.47	0.7926	$\lambda < 6.04$
20	0.51	0.73	0.3	0.32	1.67	2.33	0.08	0.46	0.8153	$\lambda < 8$
	0.54	0.75	0.33	0.32	1.65	2.35	0.08	0.47	0.7381	$\lambda < 7.93$
100	0.34	0.61	0.20	0.27	1.71	2.29	0.02	0.45	0.8040	$\lambda < 10.88$
	0.40	0.65	0.24	0.28	1.67	2.33	0.02	0.47	0.7512	$\lambda < 10.01$

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

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}	$Cor(L_1, L_2)$	stab. cond
3.3	61.46	61.39	32.99	28.72	1.87	2.13	0.56	0.43	0.9999	$\lambda < 4.03$
	6013.3	6013.3	3099.1	2919.6	1.94	2.06	0.59	0.41	1.00	$\lambda < 4.00$
3.5	12.52	12.46	6.64	5.89	1.88	2.12	0.53	0.42	0.9985	$\lambda < 4.16$
	14.90	14.88	7.65	7.25	1.95	2.05	0.56	0.41	0.9977	$\lambda < 4.13$
4	4.35	4.31	2.25	2.08	1.93	2.07	0.48	0.41	0.9888	$\lambda < 4.46$
	4.45	4.48	2.26	2.18	1.97	2.03	0.49	0.41	0.9763	$\lambda < 4.45$
4.5	2.71	2.69	1.37	1.32	1.97	2.03	0.44	0.41	0.9738	$\lambda < 4.74$
	2.71	2.71	1.37	1.35	1.98	2.02	0.44	0.40	0.9421	$\lambda < 4.73$
5	2	2	1	1	2	2	0.4	0.4	0.9565	$\lambda < 5$
	2	2	1	1	2	2	0.4	0.4	0.9042	$\lambda < 5$
8	0.87	0.92	0.41	0.49	2.12	1.88	0.26	0.38	0.8568	$\lambda < 6.27$
	0.91	0.94	0.44	0.49	2.07	1.93	0.26	0.39	0.7493	$\lambda < 6.20$
20	0.37	0.51	0.16	0.29	2.28	1.72	0.11	0.35	0.7130	$\lambda < 9.21$
	0.45	0.55	0.21	0.30	2.19	1.81	0.11	0.36	0.6596	$\lambda < 8.47$
100	0.20	0.39	0.08	0.24	2.38	1.62	0.02	0.32	0.6973	$\lambda < 14.85$
	0.29	0.44	0.13	0.26	2.30	1.70	0.02	0.34	0.6698	$\lambda < 11.12$

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

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}	$Cor(L_1, L_2)$	stab. cond
3.5	130.99	130.83	57.23	76.46	2.29	1.71	0.65	0.34	0.9999	$\lambda < 4.01$
	35.19	34.98	16.18	19.16	2.17	1.83	0.62	0.37	0.9996	$\lambda < 4.05$
4	5.65	5.51	2.40	3.36	1.64	2.36	0.59	0.33	0.9933	$\lambda < 4.36$
	5.17	4.99	2.31	2.83	2.24	1.76	0.56	0.35	0.9820	$\lambda < 4.40$
4.5	2.97	2.86	1.23	1.80	1.59	2.41	0.54	0.32	0.9778	$\lambda < 4.69$
	2.92	2.74	1.27	1.61	2.30	1.70	0.51	0.34	0.9473	$\lambda < 4.71$
5	2.05	1.95	0.83	1.27	2.46	1.54	0.49	0.31	0.9575	$\lambda < 5$
	2.08	1.92	0.87	1.16	2.34	1.66	0.47	0.33	0.9056	$\lambda < 5$
8	0.78	0.73	0.29	0.54	2.66	1.34	0.33	0.27	0.8215	$\lambda < 6.57$
	0.87	0.77	0.34	0.53	2.54	1.46	0.32	0.29	0.7074	$\lambda < 6.35$
20	0.28	0.30	0.09	0.29	2.94	1.06	0.15	0.21	0.5803	$\lambda < 10.70$
	0.36	0.34	0.13	0.30	2.85	1.15	0.14	0.23	0.5473	$\lambda < 9.04$
100	0.09	0.17	0.03	0.22	3.19	0.81	0.03	0.16	0.5326	$\lambda < 21.82$
	0.17	0.21	0.05	0.24	3.12	0.88	0.03	0.18	0.5190	$\lambda < 12.27$

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

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}	$Cor(L_1, L_2)$	stab. cond
4.2	4.68	4.48	1.75	3.36	2.67	1.33	0.63	0.27	0.9907	$\lambda < 4.44$
	4.23	3.93	1.72	2.55	2.45	1.55	0.58	0.31	0.9744	$\lambda < 4.98$
4.5	3.17	2.98	1.17	2.31	2.71	1.29	0.60	0.26	0.9807	$\lambda < 4.66$
	3.06	2.77	1.22	1.86	2.51	1.49	0.57	0.30	0.9526	$\lambda < 4.69$
5	2.08	1.91	0.75	1.56	2.77	1.23	0.55	0.25	0.9593	$\lambda < 5$
	2.13	1.87	0.83	1.31	2.58	1.42	0.52	0.28	0.9097	$\lambda < 5$
8	0.73	0.61	0.24	0.63	3.03	0.97	0.38	0.19	0.8027	$\lambda < 6.79$
	0.84	0.65	0.29	0.59	2.89	1.11	0.36	0.22	0.6783	$\lambda < 6.47$
20	0.24	0.17	0.07	0.31	3.44	0.56	0.17	0.11	0.4857	$\lambda < 11.86$
	0.30	0.20	0.09	0.32	3.37	0.63	0.17	0.13	0.4085	$\lambda < 9.44$
100	0.05	0.03	0.01	0.22	3.85	0.15	0.039	0.029	0.2878	$\lambda < 29.22$
	0.06	0.04	0.02	0.24	3.83	0.17	0.04	0.03	0.2056	$\lambda < 13.07$

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

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}	$Cor(L_1, L_2)$	stab. cond
3.3	22.00	21.97	12.75	9.65	1.72	2.27	0.52	0.45	0.9995	$\lambda < 4.09$
	67.08	67.11	35.94	31.45	1.87	2.13	0.56	0.42	0.9998	$\lambda < 4.03$
3.5	9.75	9.72	5.53	4.35	1.76	2.24	0.50	0.45	0.9976	$\lambda < 4.20$
	12.69	12.71	6.75	5.99	1.87	2.13	0.54	0.42	0.9968	$\lambda < 4.16$
4	4.16	4.14	2.25	1.93	1.85	2.15	0.46	0.43	0.9878	$\lambda < 4.48$
	4.34	4.35	2.26	2.09	1.91	2.09	0.48	0.42	0.9753	$\lambda < 4.46$
4.5	2.69	2.67	1.39	1.29	1.93	2.07	0.43	0.41	0.9735	$\lambda < 4.74$
	2.69	2.70	1.38	1.32	1.96	2.04	0.43	0.41	0.9417	$\lambda < 4.74$
5	2	2	1	1	2	2	0.4	0.4	0.9565	$\lambda < 5$
	2	2	1	1	2	2	0.4	0.4	0.9042	$\lambda < 5$
8	0.82	0.85	0.35	0.51	2.22	1.67	0.29	0.33	0.8428	$\lambda < 6.38$
	0.89	0.88	0.40	0.50	2.25	1.75	0.28	0.35	0.7332	$\lambda < 6.26$
20	0.28	0.30	0.09	0.29	2.94	1.06	0.15	0.21	0.5803	$\lambda < 10.70$
	0.36	0.35	0.13	0.30	2.85	1.15	0.14	0.23	0.5457	$\lambda < 9.04$
100	0.060	0.067	0.01	0.21	3.68	0.32	0.03	0.06	0.3762	$\lambda < 27.21$
	0.09	0.08	0.02	0.24	3.65	0.35	0.03	0.07	0.3295	$\lambda < 12.88$

Table 6 The probability mass function of D, for $\lambda = 4$, $\mu_2 = 5$, $p_1 = 0.2$, $p_2 = 0.8$

μ_1	2	1_1	0_1	-1_1	1_2	0_2	-1_2	-2
3.3	0.0123	0.1782	0.2367	0.1002	0.0179	0.2172	0.2007	0.0368
3.5	0.0118	0.1732	0.2301	0.0901	0.0177	0.2348	0.2073	0.035
4	0.0108	0.1620	0.2181	0.0706	0.0173	0.2694	0.2204	0.0314
5	0.0093	0.1438	0.2062	0.0467	0.0168	0.3141	0.2372	0.0259
8	0.0067	0.1083	0.2023	0.0189	0.0161	0.3722	0.2587	0.0168
20	0.0032	0.055	0.2261	0.0029	0.0154	0.4164	0.2747	0.0063
100	0.0007	0.0129	0.26	0.0001	0.0151	0.431	0.2792	0.001

Table 7 The probability mass function of D, for $\lambda = 4$, $\mu_2 = 5$, $p_1 = p_2 = 0.5$

μ_1	2	1_1	0_1	-1_1	1_2	0_2	-1_2	-2
3.3	0.0276	0.2243	0.2639	0.0723	0.0403	0.1815	0.1636	0.0265
3.5	0.0263	0.219	0.2611	0.0655	0.0395	0.1952	0.1679	0.0255
4	0.0236	0.2071	0.2574	0.0522	0.0376	0.2226	0.1763	0.0232
5	0.0195	0.1871	0.2582	0.0352	0.0352	0.2582	0.1871	0.0195
8	0.0131	0.1458	0.2804	0.0143	0.0316	0.3027	0.1994	0.0127
20	0.0058	0.0782	0.3526	0.0019	0.0280	0.3264	0.2027	0.0044
100	0.0012	0.0192	0.4347	0.00005	0.026	0.3219	0.1964	0.00055

Table 8 The probability mass function of D, for $\lambda = 4$, $\mu_2 = 5$, $p_1 = 0.8$, $p_2 = 0.2$

μ_1	2	1_1	0_1	-1_1	1_2	0_2	-1_2	-2
3.5	0.0379	0.2707	0.2904	0.0301	0.0568	0.1654	0.137	0.0117
4	0.0329	0.2583	0.2968	0.0244	0.0526	0.1836	0.1405	0.0109
5	0.0259	0.2372	0.3141	0.0168	0.0467	0.2062	0.1438	0.0093
8	0.0156	0.1913	0.3728	0.0069	0.0374	0.2271	0.1427	0.0062
20	0.0057	0.1089	0.5192	0.0009	0.0272	0.2116	0.1245	0.002
100	0.001	0.0282	0.6821	0.00001	0.0203	0.16999	0.0983	0.0002

Table 9 The probability mass function of D, for $\lambda = 4$, $\mu_2 = 5$, $p_1 = 1$, $p_2 = 0$

μ_1	2	1_1	0_1	-1_1	1_2	0_2	-1_2	-2
4.2	0.0354	0.2924	0.3296	0.00	0.0581	0.1646	0.1199	0.00
4.5	0.0324	0.2858	0.3386	0.00	0.0552	0.1686	0.1194	0.00
5	0.0283	0.2754	0.3541	0.00	0.051	0.1731	0.1181	0.00
8	0.015	0.2272	0.4439	0.00	0.036	0.1723	0.1056	0.00
20	0.0038	0.1349	0.6602	0.00	0.0183	0.1164	0.0664	0.00
100	0.0002	0.0365	0.9083	0.00	0.0046	0.0323	0.0181	0.00

Table 10 The probability mass function of D, for $\lambda = 4$, $\mu_2 = 5$, $p_1 = \frac{\mu_1}{\mu_1+\mu_2}$, $p_2 = \frac{\mu_2}{\mu_1+\mu_2}$

μ_1	2	1_1	0_1	-1_1	1_2	0_2	-1_2	-2
3.3	0.0227	0.208	0.255	0.0831	0.0332	0.1924	0.1751	0.0305
3.5	0.0224	0.205	0.2522	0.0739	0.0336	0.2057	0.1785	0.0287
4	0.0214	0.1983	0.2502	0.0563	0.0343	0.2306	0.1838	0.0251
5	0.0195	0.1871	0.2582	0.0352	0.0352	0.2582	0.1871	0.0195
8	0.0145	0.1623	0.3141	0.0118	0.0349	0.2745	0.1774	0.0105
20	0.0057	0.1089	0.5192	0.0009	0.0272	0.2116	0.1245	0.002
100	0.0005	0.0343	0.8487	0.00	0.0094	0.06846	0.0386	0.00004

4. As μ_1 increases, the maximum value of λ ensuring stability increases. However, for $p_1 \leq 0.5$, the stability condition is more strict in the non-preemptive case than it is in the preemptive one.
5. For $\mu_1 \leq 5$, as p_1 increases, $\mathbb{E}[L_1]$ and $\mathbb{E}[L_2]$ increase, as higher proportion of the arrivals joins Q_1 when $L_1 = L_2$. If the server renders service in Q_1, the non-preemptive policy forces the server to remain there until service completion, and if $\mu_1 < \mu_2$, the number of customers arriving to the system during this service time, increases. However, when $\mu_1 > \mu_2$, as p_1 increases, the values for $\mathbb{E}[L_1]$ and $\mathbb{E}[L_2]$ decrease.
6. The correlation coefficient between L_1 and L_2 is always positive. This follows from the JSQ policy. Furthermore, as μ_1 increases, the correlation coefficient between L_1 and L_2 decreases. To explain this phenomenon, consider first small values of μ_1, where the server is busy in Q_1 for a long period of time. Therefore, since customers follow the JSQ policy, the number of customers in both queues increases simultaneously. However, for large values of μ_1, the behavior of L_2 is less affected by L_1, since the server resides in Q_1 a short amount of time.
7. Table 5 exhibits a balancing result between $\mathbb{E}[L_1]$ and $\mathbb{E}[L_2]$, in both the preemptive and the non-preemptive cases. When the joining probabilities are relative to the service rates ratio, i.e $p_i = \frac{\mu_i}{\mu_1+\mu_2}$, $i = 1, 2$, both mean queue lengths are significantly reduced.
8. All tables show that, when μ_1 increases, the proportion of time the server spends in both queues (i.e. $\rho^1_{eff} + \rho^2_{eff}$) reduces significantly.
9. In the symmetric case, i.e. when $p_1 = p_2 = 0.5$ and $\mu_1 = \mu_2$, the obtained results correspond to an $M/M/1$ system, as expected (see Tables 2 and 5 for the case $\mu_1 = 5$).

Figures 2 and 3 depict the impact of p_1 on the sum $\mathbb{E}[L_1] + \mathbb{E}[L_2]$, for the case $\lambda = 4$, and $\mu_2 = 5$, where in Fig. 2 $\mu_1 = 3.5$ while in Fig. 3 $\mu_1 = 6.5$. Obviously, when $\mu_1 < \mu_2$, that is, service time is longer in Q_1 than in Q_2 (Fig. 2), $\mathbb{E}[L_1]+\mathbb{E}[L_2]$ grows as p_1 increases. This occurs since when p_1 is high, customers tend to join the slower queue when both queue sizes are equal, which increases the total number of customers in the system. On the other hand, when $\mu_1 > \mu_2$ (Fig. 3), $\mathbb{E}[L_1] + \mathbb{E}[L_2]$ decreases as p_1 increases, since joining the faster queue with high probability when the queue sizes are equal, reduces the total number of customers in the system. Moreover,

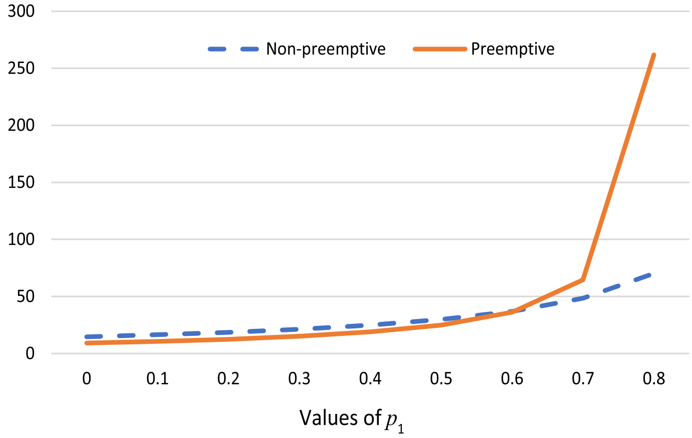


Fig. 2 The impact of p_1 on $\mathbb{E}[L_1] + \mathbb{E}[L_2]$ for $\lambda = 4$, $\mu_1 = 3.5$ and $\mu_2 = 5$

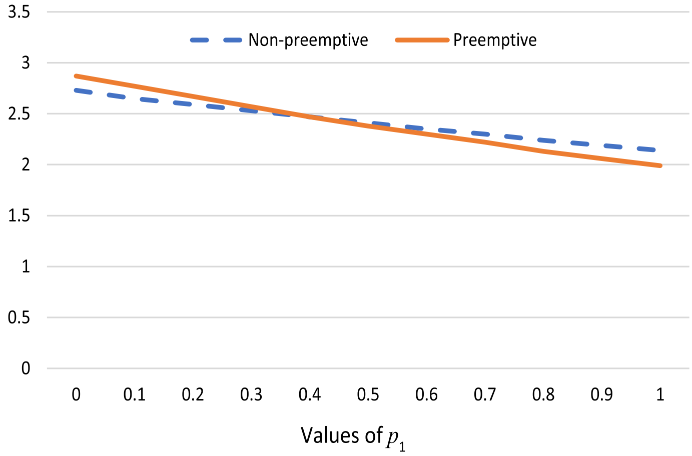


Fig. 3 The impact of p_1 on $\mathbb{E}[L_1] + \mathbb{E}[L_2]$ for $\lambda = 4$, $\mu_1 = 6.5$ and $\mu_2 = 5$

in Fig. 2 ($\mu_1 < \mu_2$), when $p_1 < 0.6$, the sum of queue sizes is almost equal in both preemptive and non-preemptive regimes, while when p_1 increases beyond 0.6, the total queue under the preemptive regime becomes much higher than that under the non-preemptive regime. In Fig. 3 ($\mu_1 > \mu_2$), there is no much difference between the sum of the queue sizes as p_1 increases, but for lower values of p_1 the non-preemptive case is better than the preemptive regime and for p_1 larger, the opposite holds.

6.1.2 Insights from Tables 6, 7, 8, 9 and 10

Tables 6, 7, 8, 9 and 10 deal with the distribution of D, where $\lambda = 4$, $\mu_2 = 5$ while p_1 and μ_1 change.

1. In all tables, as μ_1 increases, the probabilities $\mathbb{P}(D = 2)$, $\mathbb{P}(D = -2)$ and $\mathbb{P}(D = 1_1)$ decrease drastically. Indeed, as service rate in Q_1 increases, the total number

of customers in both queues decreases, as well as the gap between the queue sizes. Furthermore, $\mathbb{P}(D = 1_2)$ also decreases when μ_1 increases, where in Tables 6, 7 and 8 the decrease is more moderate than in Tables 9 and 10.

2. In all tables, the probability $\mathbb{P}(L_1 = L_2) = \mathbb{P}(D = 0_1) + \mathbb{P}(D = 0_2)$ increases as μ_1 increases, since a fast service rate enables the system to be more balanced.
3. When $p_1 \geq p_2$ (Tables 7, 8, 9, 10), the probability $\mathbb{P}(D = 0_1)$ increases as μ_1 increases. Indeed, when service rate in Q_1 is high, as well as the proportion of customers joining Q_1 when $L_1 = L_2$, the server spends more time in Q_1 and queue sizes tend to be more balanced. However, when $p_1 < p_2$ as in Table 6, $\mathbb{P}(D = 0_1)$ has no drastic changes for different values of μ_1, while $\mathbb{P}(D = 0_2)$ increases significantly as μ_1 becomes large.
4. For the symmetric case (Tables 7 and 10 for $\mu_1 = 5$) we get, as expected, that $\mathbb{P}(\text{server at } Q_1) = \mathbb{P}(\text{server at } Q_2) = 0.5$, where

$$\mathbb{P}(\text{server at } Q_1) = \mathbb{P}(D = 2) + \mathbb{P}(D = 1_1) + \mathbb{P}(D = 0_1) + \mathbb{P}(D = -1_1),$$

$$\mathbb{P}(\text{server at } Q_2) = \mathbb{P}(D = 1_2) + \mathbb{P}(D = 0_2) + \mathbb{P}(D = -1_2) + \mathbb{P}(D = -2).$$

5. In Table 9, where $p_1 = 1$, we have that $\mathbb{P}(D = -2) = 0$, since in case when $L_1 = L_2$, arriving customers will always join Q_1, so L_2 can never exceed L_1 by more than one customer.

6.2 Economic comparison between the preemptive and non-preemptive regimes

Let C denote the total operational cost rate of the JSQ–SLQ queueing systems with zero-switch-over times (either under the preemptive policy or under the non-preemptive policy). Specifically, C is the sum of the expected total cost per unit time incurred by customers' sojourn times in the system, and the cost per unit time resulting from the server's switch-overs. Define $\mathbb{E}[S_{ij}]$ to be the expected number of switches per unit time from Q_i to Q_j, $i \neq j$, and let $\mathbb{E}[Switch] = \mathbb{E}[S_{12}] + \mathbb{E}[S_{21}]$ denote the mean total number of server's switches per unit time. In steady state, we have $\mathbb{E}[S_{12}] = \mathbb{E}[S_{21}]$. Then, the expected total operational cost per unit time is

$$\mathbb{E}[C] = c\,(\mathbb{E}[L_1] + \mathbb{E}[L_2]) + s\mathbb{E}[Switch],$$

where c is the cost rate for a customer's sojourn in the system, and s is the cost per server's switch. Without loss of generality, we set $c = 1$ and write

$$\mathbb{E}[C] = (\mathbb{E}[L_1] + \mathbb{E}[L_2]) + s\mathbb{E}[Switch]. \tag{72}$$

The measures $\mathbb{E}[S_{ij}]$ for $i \neq j$ are calculated as follows.

For the preemptive regime,

$$\mathbb{E}[S_{12}] = \lambda p_2 G_{0_1}(1) + \mu_1 \left(G_{0_1}(1) - P_{0,0_1}\right),$$

Table 11 Values of $\mathbb{E}[Switch]$ for the preemptive and the non-preemptive regimes, where $\lambda = 4$ and μ_2=5

μ_1	3.5	4	4.5	5	5.5	6	6.5	7	7.5	8	8.5	9	9.5	10
$p_1 = 0.2$	2.89	2.83	2.77	2.73	2.7	2.67	2.65	2.93	2.61	2.59	2.58	2.57	2.56	2.55
	2.23	2.24	2.25	2.26	2.26	2.27	2.27	2.27	2.27	2.28	2.28	2.28	2.28	2.28
$p_1 = 0.5$	2.98	2.94	2.91	2.89	2.87	2.84	2.83	2.81	2.79	2.78	2.77	2.75	2.74	2.73
	2.24	2.28	2.31	2.33	2.35	2.36	2.37	2.37	2.38	2.38	2.38	2.38	2.38	2.38
$p_1 = 0.8$	2.81	2.75	2.76	2.73	2.71	2.68	2.65	2.62	2.59	2.56	2.54	2.51	2.49	2.46
	2.2	2.23	2.25	2.26	2.26	2.25	2.24	2.23	2.22	2.2	2.19	2.18	2.16	2.14

$$\mathbb{E}[S_{21}] = \lambda p_1 G_{0_2}(1) + \mu_2 \left(G_{0_2}(1) - P_{0,0_2}\right),$$

while for the non-preemptive regime,

$$\begin{aligned}\mathbb{E}[S_{12}] &= \lambda p_2 P_{0,0_1} + \mu_1 \left(G_{0_1}(1) + G_{-1_1}(1) - P_{0,0_1}\right),\\ \mathbb{E}[S_{21}] &= \lambda p_1 P_{0,0_2} + \mu_2 \left(G_{0_2}(1) + G_{1_2}(1) - P_{0,0_1}\right).\end{aligned}$$

6.2.1 Switching rates - numerical results and insights

The values of $\mathbb{E}[Switch]$, for both the preemptive and non-preemptive regimes, for $\lambda = 4$ and μ_2=5, where p_1 assumes the values 0.2, 0.5 and 0.8 and μ_1 varies between 3.5 to 10, are given in Table 11. For each value of p_1, the top row represents the preemptive case while the bottom row refers to the non-preemptive case.

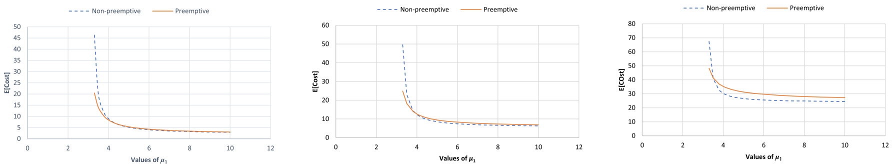
Insights from Table 11 It is evident from the table that the switching rate in the non-preemptive regime is smaller than the corresponding one in the preemptive case. This is a direct result of the non-preemptive policy. However, as μ_1 increases, the gaps between the switch-over rates between the two regimes decrease. Furthermore, Table 11 shows that for the preemptive case, the switching rate is a decreasing function of μ_1, while for the non-preemptive case, it moderately increases when $p_1 = 0.2$ and $p_1 = 0.5$, and concave when $p_1 = 0.8$.

6.2.2 Expected total operational cost—results and insights

Figures 4, 5 and 6 depict $\mathbb{E}[C]$ (Eq. 72) as a function of μ_1, both for the preemptive and the non-preemptive regimes, where in each figure $\lambda = 4$, $\mu_2 = 5$ while p_1 assumes the values 0.2, 0.5 and 0.8, respectively. Furthermore, for each figure, the left sub-figures describe the case where switch-over cost is $s = 0.5$, the mid sub-figures refer to the case where $s = 2$, while the right sub-figures describe the case where switch-over cost is $s = 10$.

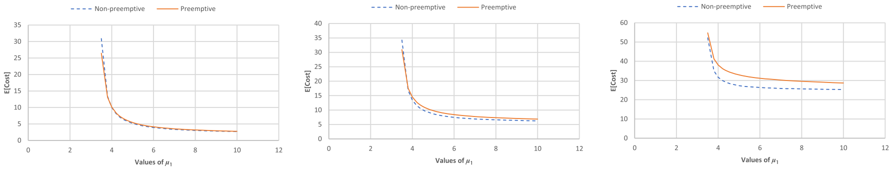
Insights from Figs. 4, 5 and 6:

1. In all 3 figures, the systems' operational cost is a decreasing function of μ_1.
2. Figure 4a and b show that when $p_1 = 0.2$, for small values of μ_1, the cost rate in the preemptive case is much lower than in the non-preemptive case, since there are significant gaps between the number of customers in the system, i.e. between $\mathbb{E}[L_1] + \mathbb{E}[L_2]$, as shown in Table 1. However, as μ_1 increases, the graphs of



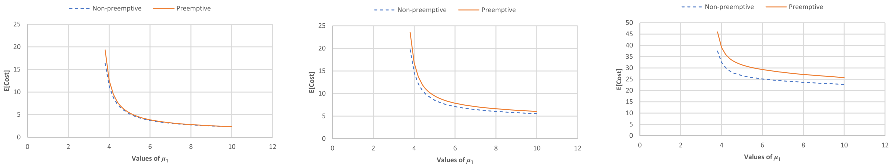
(a) Switch-over cost $s = 0.5$ (b) Switch-over cost $s = 2$ (c) Switch-over cost $s = 10$

Fig. 4 Expected operational costs of the preemptive and non-preemptive regimes, for $\lambda = 4$, $\mu_2 = 5$ and $p_1 = 0.2$



(a) Switch-over cost $s = 0.5$ (b) Switch-over cost $s = 2$ (c) Switch-over cost $s = 10$

Fig. 5 Expected operational costs of the preemptive and non-preemptive regimes, for $\lambda = 4$, $\mu_2 = 5$ and $p_1 = 0.5$



(a) Switch-over cost $s = 0.5$ (b) Switch-over cost $s = 2$ (c) Switch-over cost $s = 10$

Fig. 6 Expected operational costs of the preemptive and non-preemptive regimes, for $\lambda = 4$, $\mu_2 = 5$ and $p_1 = 0.8$

the costs of the two regimes interlace, and the operational cost rate of the non-preemptive case becomes lower than the preemptive one. Furthermore, when the penalty on a server's switch is $s = 10$ (Fig. 4c), the differences between the costs are greater, but with smaller values for the non-preemptive case. The above findings result from the values of $\mathbb{E}[switch]$, given in Table 11.

3. When $p_1 = 0.5$ and $s = 0.5$ (Fig. 5a), the graphs of operational cost rates of both regimes seem to be quite identical. As the penalty for a switch rises, see Fig. 5b and c, the expected operational cost of the non-preemptive regime becomes lower than the preemptive one.
4. Figure 6a–c refer to the case where $p_1 = 0.8$, i.e., that when $L_1 = L_2$, a newly arriving customer joins Q_1 with high probability. As a result, the operational cost rate of the non-preemptive system is smaller than the corresponding cost in preemptive case, especially when $s = 10$.

6.3 Numerical results for non-zero switch-over times

Tables 12 and 13 present sets of results, where the calculated measures in both tables are again $\mathbb{E}[L_i]$, $\mathbb{E}[W_i]$, λ^i_{eff}, ρ^i_{eff} $(i = 1, 2)$ and $Cor(L_1, L_2)$, as well as the measure $\mathbb{P}(Switch)$, which is the fraction of time that the server is switching from Q_i to Q_j. In both tables, the parameters' values are: $\mu_2 = 5$, $p_1 = p_2 = 0.5$, $\gamma_2 = 3$, while $\lambda_1 = 2$ in Table 12 and $\lambda_1 = 3$ in Table 13. Furthermore, in Table 12 the values of μ_1 and γ_1 vary between 3 to 8, and 3 to 10, respectively, while in Table 13 the values of μ_1 and γ_1 vary between 5 to 10, and 10 to 14, respectively.

Insights from Tables 12 and 13:

1. In both tables, $\mathbb{E}[L_1]$ and $\mathbb{E}[L_2]$, as well as $\mathbb{E}[W_1]$ and $\mathbb{E}[W_2]$, are all decreasing functions of μ_1 and γ_1.

Table 12 Numerical results for the model with switch-over times, $\lambda = 2$, $\mu_2 = 5$, $p_1 = p_2 = 0.5$, $\gamma_2 = 3$

μ_1	γ_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}	$Cor(L_1, L_2)$	$\mathbb{P}(Switch)$
3	3	6.40	6.38	6.54	6.24	0.98	1.02	0.37	0.20	0.9912	0.41
	4	3.454	3.447	3.49	3.41	0.99	1.01	0.33	0.20	0.9727	0.36
	5	2.66	2.66	2.67	2.65	0.998	1.002	0.33	0.20	0.9573	0.33
	10	1.79	1.81	1.76	1.83	1.02	0.98	0.338	0.197	0.9205	0.27
4	3	2.72	2.71	2.74	2.68	0.99	1.01	0.25	0.20	0.9584	0.41
	4	1.88	1.88	1.87	1.89	0.997	1.003	0.25	0.20	0.9236	0.36
	5	1.56	1.57	1.54	1.59	1.01	0.99	0.253	0.198	0.8991	0.33
	10	1.13	1.57	1.17	1.10	1.03	0.97	0.26	0.19	0.8479	0.27
5	3	1.95	1.95	1.95	1.95	1	1	0.20	0.20	0.9291	0.41
	4	1.42	1.43	1.40	1.46	1.01	0.99	0.203	0.197	0.8853	0.36
	5	1.21	1.23	1.18	1.26	1.02	0.98	0.205	0.195	0.8563	0.33
	10	0.90	0.94	0.86	0.96	1.05	0.95	0.209	0.191	0.7993	0.27
6	3	1.62	1.63	1.61	1.64	1.007	0.993	0.18	0.20	0.9070	0.41
	4	1.21	1.23	1.18	1.26	1.02	0.98	0.17	0.196	0.8582	0.36
	5	1.03	1.06	0.998	1.07	1.03	0.97	0.1725	0.194	0.8267	0.33
	10	0.77	0.83	0.73	0.87	1.05	0.95	0.18	0.19	0.7679	0.27
8	3	1.32	1.34	1.30	1.36	1.02	0.98	0.13	0.20	0.8782	0.42
	4	0.99	1.03	0.97	1.07	1.03	0.97	0.13	0.19	0.8247	0.36
	5	0.86	0.90	0.82	0.94	1.05	0.95	0.13	0.19	0.7915	0.33
	10	0.64	0.71	0.60	0.76	1.07	0.93	0.13	0.19	0.7333	0.27
10	3	1.18	1.20	1.15	1.23	1.03	0.97	0.10	0.19	0.8613	0.42
	4	0.90	0.94	0.86	0.98	1.04	0.96	0.10	0.19	0.8060	0.36
	5	0.77	0.82	0.73	0.87	1.05	0.95	0.10	0.19	0.7726	0.33
	10	0.57	0.65	0.53	0.71	1.08	0.92	0.11	0.18	0.7168	0.27

Table 13 Numerical results for the model with switch-over times, $\lambda = 3$, $\mu_2 = 5$, $p_1 = p_2 = 0.5$, $\gamma_2 = 3$

μ_1	γ_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}	$Cor(L_1, L_2)$	$\mathbb{P}(Switch)$
5	10	37.62	37.67	23.83	26.50	1.58	1.42	0.32	0.28	0.9997	0.39
	12	15.81	15.87	9.97	11.23	1.41	1.59	0.32	0.28	0.9985	0.37
	14	11.11	11.18	6.98	7.94	1.41	1.59	0.32	0.28	0.9971	0.36
6	10	6.64	6.71	4.18	4.76	1.59	1.41	0.27	0.28	0.9921	0.39
	12	5.22	5.30	3.27	3.78	1.60	1.40	0.27	0.28	0.9878	0.38
	14	4.52	4.59	2.81	3.29	1.61	1.39	0.27	0.28	0.9842	0.37
8	5	14.11	14.16	9.00	9.88	1.57	1.43	0.19	0.27	0.9981	0.49
	8	3.91	3.97	2.44	2.84	1.60	1.40	0.20	0.28	0.9793	0.42
	10	3.08	3.16	1.91	2.27	1.61	1.39	0.20	0.28	0.9692	0.39
	12	2.69	2.77	1.66	2.00	1.62	1.38	0.20	0.28	0.9616	0.38
	14	2.46	2.54	1.51	1.85	1.63	1.37	0.20	0.27	0.9559	0.37
10	4	21.44	21.48	13.75	14.91	1.56	1.44	0.16	0.29	0.9992	0.53
	5	6.15	6.20	3.89	4.36	1.58	1.42	0.16	0.28	0.9908	0.49
	8	2.74	2.81	1.70	2.03	1.61	1.39	0.16	0.28	0.9627	0.42
	10	2.27	2.35	1.39	1.71	1.62	1.38	0.16	0.27	0.9501	0.39
	12	2.03	2.11	1.24	1.55	1.64	1.36	0.16	0.27	0.9412	0.38

2. In both tables, as μ_1 increases, the effective arrival rate to Q_1, i.e. λ^1_{eff}, increases. Clearly, a fast service rate in Q_1 shortens the queue length, and therefore rises the probability that a newly arrival will join Q_1.
3. The correlation coefficient between L_1 and L_2 is always positive, as obtained in Tables 1, 2, 3, 4 and 5 . This again follows from the JSQ policy, as explained in Sect. 6.1.
4. The proportion of time the server is switching between queues is a decreasing function of γ_1. However, it is less sensitive to the service rate μ_1.

7 Concluding remarks

The combined operating policy ‘Join the Shortest Queue–Serve the Longest Queue’ is analyzed under the *non-preemptive* service regime for a 2-queue Markovian system attended by a single server. The system is formulated in an innovative way, where, instead of defining an *un-bounded* 2-dimensional state space (L_1, L_2), where L_i represents the number of customers in Q_i, $i = 1, 2$, the system is characterized by the couple L_1 and $D = L_1 - L_2$. This leads to a 2-dimensional state space with finite, and small, dimension for D, and infinite dimension only for L_1. The resulting QBD process enables the combined use of PGFs method and matrix geometric analysis. The results of the *non-preemptive* service regime are compared numerically with the corresponding results of its twin *preemptive* service regime in numerous tables for a wide range of parameter values. Among many insights, it is shown that when the overall traffic intensity $\rho = \rho^1_{eff} + \rho^2_{eff}$ approaches 1, and $p_1 \leq p_2$, the values of $\mathbb{E}[L_i]$ and $\mathbb{E}[W_i]$ under the preemptive regime are smaller than their corresponding values under the non-preemptive discipline. The ratio changes when $p_1 > p_2$. When ρ is small, the differences between the performance measures under the two regimes are small. In terms of total operating cost (customers’ sojourn times and server’s switches), it is shown that there are cases where the non-preemptive regime is more efficient economically, while in other cases the preemptive regime is preferable.

Declarations

Data availability statements We do not analyze or generate any datasets, because our work proceeds within a theoretical and mathematical approach.

Conflict of interest All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

References

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

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

Adan IJ, Boxma OJ, Kapodistria S, Kulkarni VG (2016) The shorter queue polling model. Ann Oper Res 241(1–2):167–200

Armony M, Perel E, Perel N, Yechiali U (2019) Exact analysis for multiserver queueing systems with cross selling. Ann Oper Res 274(1):75–100

Artalejo JR, Gómez-Corral A (2008) Retrial queueing systems: a computational approach. Springer, Berlin

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

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

Braverman A (2020) Steady-state analysis of the join-the-shortest-queue model in the Halfin–Whitt regime. Math Oper Res 45(3):1069–1103

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

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

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

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

Conway R, Maxwell W, Miller L (2003) Theory of scheduling. Dover books on computer science series. Dover, New York

Curtiss D (1918) Recent extentions of descartes' rule of signs. Ann Math 19:251–278

Dawson DA, Tang J, Zhao YQ et al (2019) Performance analysis of joining the shortest queue model among a large number of queues. Asia-Pac J Oper Res 36(04):1–23

Dimitriou I (2021) Analysis of the symmetric join the shortest orbit queue. Oper Res Lett 49(1):23–29

Eschenfeldt P, Gamarnik D (2018) Join the shortest queue with many servers. the heavy-traffic asymptotics. Math Oper Res 43(3):867–886

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

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

Hanukov G, Yechiali U (2021) Explicit solutions for continuous-time QBD processes by using relations between matrix geometric analysis and the probability generating functions method. Probab Eng Inf Sci 35:565–580

Harchol-Balter M (2013) Performance modeling and design of computer systems: queueing theory in action. Cambridge University Press, Cambridge

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

Jolles A, Perel E, Yechiali U (2018) Alternating server with non-zero switch-over times and opposite-queue threshold-based switching policy. Perform Eval 126:22–38

Kella O, Yechiali U (1988) Priorities in m/g/1 queue with server vacations. Naval Res Logist 35(1):23–34

Knessl C, Yao H (2013) On the nonsymmetric longer queue model: joint distribution, asymptotic properties, and heavy traffic limits. Adv Oper Res. https://doi.org/10.1155/2013/680539

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

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

Neuts MF (1981) Matrix-geometric solutions in stochastic models: an algorithmic approach. The Johns Hopkins University Press, Baltimore

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 (2013a) On customers acting as servers. Asia-Pac J Oper Res 30(05):1350019

Perel N, Yechiali U (2013b) The Israeli queue with priorities. Stoch Model 29(3):353–379

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

Perel E, Perel N, Yechiali U (2020) A polling system with "join the shortest-serve the longest" policy. Comput Oper Res 114:104809

Perel E, Perel N, Yechiali U (2022) A 3-queue polling system with join the shortest-serve the longest policy. Indag Math 34:1101–1120

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

Takagi H (1986) Analysis of polling systems. MIT Press, Cambridge

van Houtum G-J, Adan IJ, Der wal J (1997) The symmetric longest queue system. Stoch Model 13(1):105–120

van Houtum G-J, Adan IJ, Wessels J, Zijm WH (2001) Performance analysis of parallel identical machines with a generalized shortest queue arrival mechanism. OR-Spektrum 23(3):411–427

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 Model 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. In: Donatiello L, Nelson R (eds) Performance evaluation of computer and communication systems. Springer, Berlin, pp 630–650

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.