HAL
open science

# Finite-Buffer Polling Systems with Threshold-Based Switching Policy

Konstantin Avrachenkov, Efrat Perel, Uri Yechiali

## To cite this version:

Konstantin Avrachenkov, Efrat Perel, Uri Yechiali. Finite-Buffer Polling Systems with Threshold-Based Switching Policy. TOP, 2016, 24 (3), pp.541 - 571. 10.1007/s11750-015-0408-6 . hal-01402886

**HAL Id: hal-01402886**

**https://inria.hal.science/hal-01402886**

Submitted on 25 Nov 2016

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Finite-Buffer Polling Systems with Threshold-Based Switching Policy

Kostia Avrachenkov[1] Efrat Perel[2,3] and Uri Yechiali[3]

[1]Inria, Sophia Antipolis, France.
[2]Afeka, Tel-Aviv Academic College of Engineering, Tel-Aviv, Israel.
[3]Department of Statistics and Operations Research,
School of Mathematical Sciences, Tel-Aviv University, Tel-Aviv, Israel.
[k.avrachenkov@sophia.inria.fr] [efratp@afeka.ac.il] [uriy@post.tau.ac.il]

**Abstract**

We consider a system of two separate finite-buffer $M/M/1$ queues served by a single server, where the switching mechanism between the queues is threshold-based, determined by the queue which is *not being served*. Applications may be found in data centers, smart traffic-light control and human behavior. Specifically, whenever the server attends queue $i$ ($Q_i$) and the number of customers in the other queue, $Q_j$ ($i, j = 1, 2;\ j \neq i$), reaches its threshold level, the server immediately switches to $Q_j$ whenever $Q_i$ is below its threshold. When a served $Q_i$ becomes empty we consider two scenarios: ($i$) non-work-conserving; and ($ii$) work-conserving. We present occasions where the non-work-conserving policy is more economical than the work-conserving policy when high switching costs are involved. An intrinsic feature of the process is an oscillation phenomenon: when the occupancy of $Q_i$ decreases the occupancy of the other queue increases. This fact is illustrated and discussed. By formulating the system as a three-dimensional continuous-time Markov chain we provide a probabilistic analysis of the system and investigate the effects of buffer sizes and arrival rates, as well as service rates, on the system's performance. Numerical examples are presented and extreme cases are investigated.

**Keywords** Polling Systems · Threshold Policy · Finite-Buffer Queues · Oscillations
**Mathematic Subject Classification (2000)** Primary: 60K25

# 1 Introduction

We study a two-queue finite-buffer polling-type system with a threshold-based switching policy (see Figure 1.1). However, in contrast to other models, switching instants from one queue to another are determined by the state of the queue which is *not being served*. That is, when the server attends queue $i$ ($Q_i$), $i = 1, 2$, it serves the customers there until the first moment thereafter when the number of customers in $Q_j$ ($j = 1, 2,\ j \neq i$) reaches its threshold level. At that instant the server immediately switches to $Q_j$ (preemptive policy), unless the number of customers in $Q_i$ is greater or equal to the threshold level there. In the latter case the server remains in $Q_i$ until the number of customers there is reduced below the threshold level, and only then it switches to $Q_j$. The server will remain in $Q_j$ following the same switching policy, and so on. When a served $Q_i$ becomes empty we consider two switching scenarios: ($i$) non-work-conserving: the server remains in $Q_i$ until $Q_j$ reaches its threshold level, and ($ii$) work-conserving: the server immediately switches to $Q_j$ if it is not empty. Although the work-conserving scenario seems more efficient, we present cases where it is less economical when high switching costs are involved. For each queue $i$, $i = 1, 2$, we assume that the queue capacity is $C_i < \infty$, so that a customer arriving at $Q_i$ and finds $C_i$ customers there, leaves the system never to return. Customers arrive at $Q_i$ according to a Poisson process with rate $\lambda_i$ and the service time there is exponentially distributed with mean $1/\mu_i$. The processes are independent. The threshold levels are $K_1 \leq C_1$ for $Q_1$, and $K_2 \leq C_2$ for $Q_2$. To keep the presentation more tractable, we provide analysis for the case where the threshold level in each queue equals the full capacity of the queue, namely, $K_1 = C_1$ and $K_2 = C_2$. The analysis when $K_1 < C_1$ and/or $K_2 < C_2$ can be performed in a similar manner, although the details become more cumbersome. Moreover, our numerical results show that for sufficiently large values of the thresholds and buffer capacities, the values of the buffer capacities have very little effect on the system performance. The cases when $C_1$ or/and $C_2$ are infinite are currently under investigation by the authors.

One application of such a model arises in data centers, where a rack of discs requires a special attention when the amount of recorded data exceeds a certain limit (threshold), causing an inefficient operation that calls for a clean up action. Another motivation for such systems is an automated traffic-light control that regulates the traffic of vehicles crossing an intersection. The traffic-light alternates right-of-way priority between two directions as follows: when one direction has the right-of way and the accumulating number of cars in the *other* direction reaches a threshold, the right-of-way is transferred to the latter direction, and vice versa. In fact, human-beings behave

in a similar manner: while working on a given task they let the load of other tasks pile up. Only when the amount of work of another task exceeds a threshold, they switch their attention to that task.

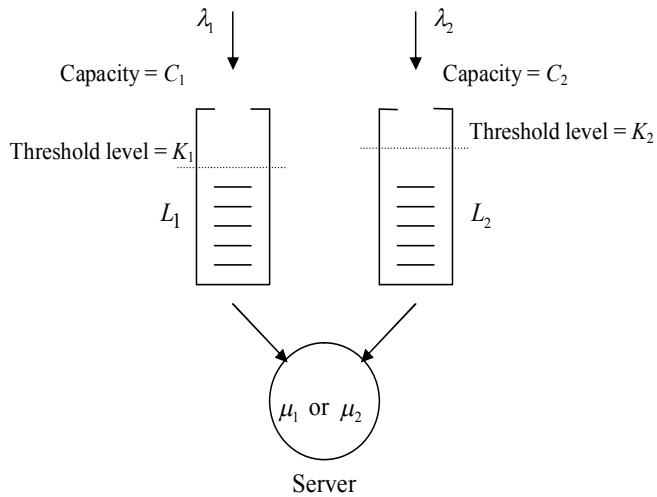


Figure 1.1: Two-queue polling system with threshold-based switching policy.

Polling systems have been studied extensively in the queueing literature mostly for unbounded-buffer systems (Boon et al. 2011; Takagi 1986; Yechiali 1993; and references there). Threshold based polling systems with unbounded buffers were also treated in the past (Lee and Sengupta 1993; Haverkort et al. 1994; Boxma et al. 1995a, 1995b; Lee 1996; Avram and Gómez-Corral 2006). Lee (1996) considered a single-server two-queue model, where a high priority queue is served exhaustively and a low priority queue follows a $k$-limited service policy. Lee and Sengupta (1993) analyzed a system where customers of each queue are served alternately unless the queue length of the high priority queue exceeds a certain threshold. Then, only the high priority customers are served until the queue length there reduces back to the threshold level. A variant of the latter model was studied by Haverkort et al. (1994): once the threshold is exceeded, the server serves the high priority queue exhaustively. Boxma et al. (1995a, 1995b) studied a similar model in which queues are being served exhaustively unless a certain threshold level is reached. In those studies service times are assumed to be exponentially distributed. In Boxma et al. (1995a) the service process at queue 2 is preemptively interrupted whenever the threshold at queue 1 is reached, while in Boxma et al. (1995b) it is nonpreemptively interrupted. The latter case was extended by Boxma and Down (1997) to the case where service times are generally distributed, and was

further examined by Avram and Gómez-Corral (2006) where the main objective is to investigate the solution of a dynamic programming optimality equation.

In the current work our aim is to determine the joint distribution function of the queue size process. To this end we formulate the system as a three-dimensional continuous-time Markov chain and study its steady-state behavior by applying two solution methods: $(i)$ Probability Generating Functions (PGFs), and $(ii)$ Matrix analytic approach. We present a full analysis of the non-work-conserving scenario, and discuss briefly the details of the work-conserving scenario. The two scenarios are compared numerically. Moreover, such a model exhibits oscillations, whose nature depends on the relative values of the various parameters: when the occupancy of $Q_i$ decreases the occupancy of the other queue increases. Vivid examples of oscillations between two queues are presented in Coffman, Puhalskii and Reiman (1995) (C.P.R) and in Arazi, Ben-Jacob and Yechiali (2005). However, there is an important difference between the work of C.P.R and the present model. In the former heavy load regime is needed for appearance of oscillations, whereas in this work, the oscillations manifest themselves when the values of the thresholds are sufficiently large irrespectively of the system load. Investigating these oscillations led to an observation that the dynamics of the system converges to a periodic behavior when the threshold values, $K_1$ for $Q_1$ and $K_2$ for $Q_2$, are large. We leave the probabilistic analysis of this convergence for future research. A recent work, still in progress, by Jonckheere, Nazarathy and Rojas-Nandayapa may provide a good base for the fluid analysis of the considered system.

The structure of the paper is as follows: In Section 2 we consider the non-work-conserving scenario. In Subsection 2.1 we define the system as a three-dimensional Markovian process and develop the system's balance equations. Consequently, in Subsection 2.3, we construct the corresponding marginal PGFs. The solution of the PGFs is obtained by solving two finite linear systems of the form $A(z)\vec{G}(z) = \vec{P}(z)$ and $B(w)\vec{F}(w) = \vec{\Pi}(w)$, where $\vec{G}(z)$ and $\vec{F}(w)$ are vector functions of the sought-for PGFs; $A(z)$ and $B(w)$ are finite square matrices with entries constructed from the parameters of the system; and $\vec{P}(z)$ and $\vec{\Pi}(w)$ are finite-dimensional vectors consisting of unknown boundary probabilities (Litvak and Yechiali 2003; Perel and Yechiali 2008, 2013). We show how to calculate the boundary probabilities determining the PGFs by deriving explicit *closed-form* combinatorial expressions for the determinants of $A(z)$ and $B(w)$. We are not aware of any previous relevant studies that obtained such explicit expressions. In Subsection 2.4 we use a matrix analytic approach to solve the two sets of balance equations constructed in Subsection 2.1. It follows that the generator matrix defining the process differs from the classical quasi birth-and-death (QBD) processes generators. Although the three-dimensional finite set of balance equations can

be solved numerically, we present an algorithmic-type method to compute the steady state probabilities. We use this approach since traditional matrix geometric methods (Neuts 1981; Latouche and Ramaswami 1999) usually address infinite-dimensional systems, while truncation methods for infinite-dimensional systems (Bright and Taylor 1995), are not relevant here. Our analysis leads to the calculation of certain matrices defined by combinatorial expressions resembling the combinatorial expressions derived for the determinants of $A(z)$ and $B(w)$ associated with the PGFs. Furthermore, we are able to reduce the Kolmogorov set of $2(K_1+1)\times(K_2+1)$ linear equations to a set of $K_1+K_2+2$ linear equations in the $K_1+K_2+1$ unknown boundary probabilities, by which all other probabilities are calculated. In Section 3 we discuss the second switching scenario, while in Section 4 we present numerical results. Section 5 deals with extreme cases. The oscillation phenomenon is presented and discussed in Section 6, where the two scenarios are also compared. Section 7 concludes the paper.

# 2 Scenario 1: Non-Work-Conserving Policy

## 2.1 Balance Equations

Let $L_i(t)$ denote the number of customers present in $Q_i$ at time $t$, $i=1,2$, and $L_i=\lim_{t\to\infty}L_i(t)$. Let $I(t)=1$ if the server attends $Q_1$ at time $t$, and $I(t)=2$ if the server attends $Q_2$ ($I=\lim_{t\to\infty}I(t)$). The triple $(L_1(t),L_2(t),I(t))$ defines a non reducible continuous-time finite Markov chain, with transition-rate diagram depicted in Figure 2.1 (the numbers 1 or 2 appearing next to each node indicate whether $I=1$, or $I=2$, respectively. Each box $(k,n)$ depicts both the state where $I(t)=1$ and the state where $I(t)=2$). Let $P_{kn}(i)=\mathbb{P}(L_1=k,L_2=n,I=i)$, where $0\le k\le K_1$, $0\le n\le K_2$, $i=1,2$, denote the steady state probabilities of the system state. Then, the sets of balance equations, for $I=1$ are given bellow. For $I=2$ the equations are symmetrical.

$\underline{I=1}$.

$$\begin{aligned}
&\underline{k=0}:\\
&n=0: \quad (\lambda_1+\lambda_2)P_{00}(1)=\mu_1P_{10}(1)\\
&1\le n\le K_2-1: \quad (\lambda_1+\lambda_2)P_{0n}(1)=\lambda_2P_{0,n-1}(1)+\mu_1P_{1n}(1) \qquad (2.1)\\
&\underline{1\le k\le K_1-1}:\\
&n=0: \quad (\lambda_1+\lambda_2+\mu_1)P_{k0}(1)=\lambda_1P_{k-1,0}(1)+\mu_1P_{k+1,0}(1)\\
&1\le n\le K_2-1: \quad (\lambda_1+\lambda_2+\mu_1)P_{kn}(1)=\lambda_1P_{k-1,n}(1)+\lambda_2P_{k,n-1}(1)+\mu_1P_{k+1,n}(1) \qquad (2.2)
\end{aligned}$$

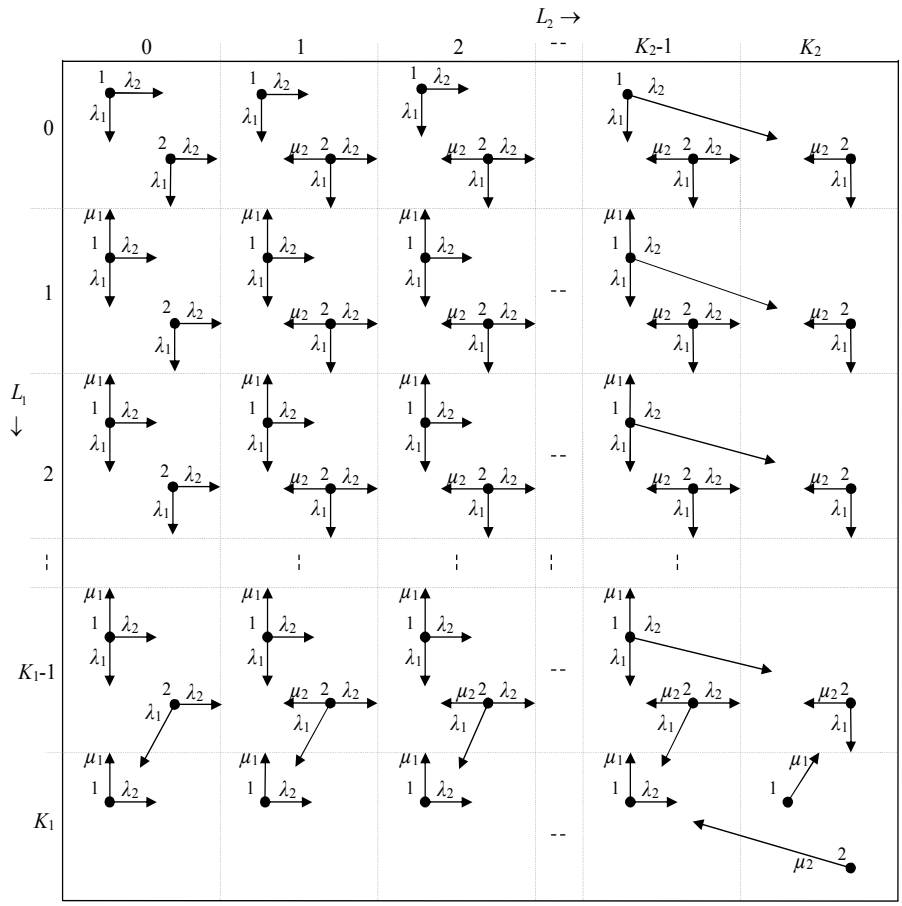


Figure 2.1: Transition rate diagram of $(L_1, L_2, I)$ for Scenario 1.

$$
\begin{aligned}
&\underline{k = K_1}: \\
&n = 0: \quad (\lambda_2 + \mu_1) P_{K_1 0}(1) = \lambda_1 P_{K_1 - 1, 0}(1) + \lambda_1 P_{K_1 - 1, 0}(2) \\
&1 \le n \le K_2 - 2: \quad (\lambda_2 + \mu_1) P_{K_1 n}(1) = \lambda_1 P_{K_1 - 1, n}(1) + \lambda_1 P_{K_1 - 1, n}(2) + \lambda_2 P_{K_1, n-1}(1) \\
&n = K_2 - 1: \quad (\lambda_2 + \mu_1) P_{K_1, K_2 - 1}(1) = \lambda_1 P_{K_1 - 1, K_2 - 1}(1) + \lambda_1 P_{K_1 - 1, K_2 - 1}(2) \\
&\qquad\qquad\qquad\qquad + \lambda_2 P_{K_1, K_2 - 2}(1) + \mu_2 P_{K_1 K_2}(2) \\
&n = K_2: \quad \mu_1 P_{K_1 K_2}(1) = \lambda_2 P_{K_1, K_2 - 1}(1)
\end{aligned} \tag{2.3}
$$

Define the following marginal probabilities

$$
\mathbb{P}(L_1 = k, I = i) \equiv P_{k\bullet}(i) = \sum_n P_{kn}(i) \ , \quad \text{for} \ \ 0 \le k \le K_1, \quad i = 1, 2,
$$

$$
\mathbb{P}(L_2 = n, I = i) \equiv P_{\bullet n}(i) = \sum_k P_{kn}(i) \ , \quad \text{for} \ \ 0 \le n \le K_2, \quad i = 1, 2,
$$

Summing equations (2.1) – (2.3) we arrive at

$$\mu_1 P_{K_1K_2}(1) + \lambda_2 \left(P_{\bullet K_2-1}(1) - P_{K_1,K_2-1}(1)\right) = \mu_2 P_{K_1K_2}(2) + \lambda_1 \left(P_{K_1-1\bullet}(2) - P_{K_1-1,K_2}(2)\right). \tag{2.4}$$

Indeed, equation (2.4) states that the average switching rate from state $I = 1$ to state $I = 2$ (left hand side of (2.4)) is equal to the average switching rate from state $I = 2$ to state $I = 1$ (right hand side of (2.4)).

## 2.2 Idleness and Carried Loads

Summing over $n$ each equation for $k$ in the set (2.1) – (2.3) for $I = 1$, together with the corresponding set for $I = 2$, we get, after rearranging terms,

$$\lambda_1 \mathbb{P}\left(L_1 = k\right) = \mu_1 P_{k+1\bullet}(1), \qquad 0 \le k \le K_1 - 1 \tag{2.5}$$

Similarly, for every $n$, $0 \le n \le K_2 - 1$, we have

$$\lambda_2 \mathbb{P}\left(L_2 = n\right) = \mu_2 P_{\bullet n+1}(2), \qquad 0 \le n \le K_2 - 1 \tag{2.6}$$

Now, summing (2.5) over all $k$, $0 \le k \le K_1 - 1$, and rearranging terms we obtain

$$P_{0\bullet}(1) = \mathbb{P}\left(I = 1\right) - \frac{\lambda_1^{eff}}{\mu_1} = \mathbb{P}\left(I = 1\right) - \rho_1^{eff}. \tag{2.7}$$

Similarly,

$$P_{\bullet 0}(2) = \mathbb{P}\left(I = 2\right) - \frac{\lambda_2^{eff}}{\mu_2} = \mathbb{P}\left(I = 2\right) - \rho_2^{eff}, \tag{2.8}$$

where, for every $i = 1, 2$, $\lambda_i^{eff} = \lambda_i \left(1 - \mathbb{P}\left(L_i = K_i\right)\right) = \lambda_i \left(1 - \mathbb{P}\left(\text{arriving customer is lost in } Q_i\right)\right)$ and $\rho_i^{eff} = \frac{\lambda_i^{eff}}{\mu_i}$. Notice that $\rho_i^{eff}$ is exactly the mean carried load at $Q_i$, $i = 1, 2$.
Summing (2.7) and (2.8) we obtain

$$P_{0\bullet}(1) + P_{\bullet 0}(2) = 1 - \left(\rho_1^{eff} + \rho_2^{eff}\right). \tag{2.9}$$

Equation (2.9) states that the proportion of time the server is idle, i.e. resides in an empty queue (LHS of (2.9)) equals one minus the proportion of time the server is busy (RHS of (2.9)).

## 2.3 Generating Functions

Define, for each $0 \le k \le K_1$, and $0 \le n \le K_2$, the probability generating functions (PGFs) for states $I = 1$ and $I = 2$, respectively: $G_k(z) = \sum_{n=0}^{K_2} P_{kn}(1) z^n$, $F_n(w) = \sum_{k=0}^{K_1} P_{kn}(2) w^k$. Multiplying by $z^n$ each equation for $n$ in the sets (2.1) – (2.3), summing over $n$, and rearranging terms we get

$\underline{k = 0}$ :

$$\left(\lambda_1 + \lambda_2(1 - z)\right) G_0(z) = \mu_1 G_1(z) - \lambda_2 P_{0,K_2-1}(1) z^{K_2} \tag{2.10}$$

$\underline{1 \le k \le K_1 - 2}$ :

$$(\lambda_1 + \mu_1 + \lambda_2(1-z))\, G_k(z) = \lambda_1 G_{k-1}(z) + \mu_1 G_{k+1}(z) - \lambda_2 P_{k,K_2-1}(1) z^{K_2} \tag{2.11}$$

$\underline{k = K_1 - 1}$ :

$$\begin{aligned}(\lambda_1 + \mu_1 + \lambda_2(1-z))\, G_{K_1-1}(z) &= \lambda_1 G_{K_1-2}(z) + \mu_1 G_{K_1}(z) - \lambda_2 P_{K_1-1,K_2-1}(1) z^{K_2} \\ &\quad - \mu_1 P_{K_1K_2}(1) z^{K_2}\end{aligned} \tag{2.12}$$

$\underline{k = K_1}$ :

$$\begin{aligned}(\mu_1 + \lambda_2(1-z))\, G_{K_1}(z) &= \lambda_1 G_{K_1-1}(z) + \lambda_1 \sum_{n=0}^{K_2-1} P_{K_1-1,n}(2) z^n + \lambda_2 P_{K_1K_2}(1)(1-z) z^{K_2} \\ &\quad + \mu_2 P_{K_1K_2}(2) z^{K_2-1}\end{aligned} \tag{2.13}$$

The sets (2.10) – (2.13), and the corresponding set for $F_n(w)$, $0 \le n \le K_2$, comprise two systems of linear equations of the form

$$A(z)\vec{G}(z) = \vec{P}(z)\ , \quad B(w)\vec{F}(w) = \vec{\Pi}(w),$$

where the column vectors $\vec{G}(z)$, $\vec{P}(z)$, $\vec{F}(w)$ and $\vec{\Pi}(w)$, and the matrices $A(z)$ and $B(w)$ are defined as follows:

$$\begin{aligned}&\vec{G}(z) = (G_0(z), G_1(z), ..., G_{K_1}(z))^t\,, \qquad \vec{P}(z) = (P_0(z), P_1(z), ..., P_{K_1}(z))^t\,, \\ &\vec{F}(w) = (F_0(w), F_1(w), ..., F_{K_2}(w))^t\,, \qquad \vec{\Pi}(w) = (\Pi_0(w), \Pi_1(w), ..., \Pi_{K_2}(w))^t\,,\end{aligned}$$

with

$$P_k(z) = \begin{cases} -\lambda_2 P_{k,K_2-1}(1) z^{K_2}, & 0 \le k \le K_1 - 2 \\ -\lambda_2 P_{K_1-1,K_2-1}(1) z^{K_2} - \mu_1 P_{K_1K_2}(1) z^{K_2}, & k = K_1 - 1 \\ \lambda_1 \sum_{n=0}^{K_2-1} P_{K_1-1,n}(2) z^n + \lambda_2 P_{K_1K_2}(1)(1-z) z^{K_2} + \mu_2 P_{K_1K_2}(2) z^{K_2-1}, & k = K_1 \end{cases}$$

$$\Pi_n(w) = \begin{cases} -\lambda_1 P_{K_1-1,n}(2) w^{K_1}, & 0 \le n \le K_2 - 2 \\ -\lambda_1 P_{K_1-1,K_2-1}(2) w^{K_1} - \mu_2 P_{K_1K_2}(2) w^{K_1}, & n = K_2 - 1 \\ \lambda_2 \sum_{k=0}^{K_1-1} P_{k,K_2-1}(1) w^k + \lambda_1 P_{K_1K_2}(2)(1-w) w^{K_1} + \mu_1 P_{K_1K_2}(1) w^{K_1-1}, & n = K_2 \end{cases}$$

$$A(z) = \begin{pmatrix} \alpha_0(z) & -\mu_1 & 0 & \cdots & \cdots & 0 \\ -\lambda_1 & \alpha(z) & -\mu_1 & 0 & \cdots & 0 \\ 0 & -\lambda_1 & \alpha(z) & -\mu_1 & 0 & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & -\lambda_1 & \alpha(z) & -\mu_1 \\ 0 & \cdots & \cdots & 0 & -\lambda_1 & \alpha_{K_1}(z) \end{pmatrix},$$

where $\alpha_0(z) = \lambda_1 + \lambda_2(1-z)$; $\alpha(z) = \lambda_1 + \mu_1 + \lambda_2(1-z)$; $\alpha_{K_1}(z) = \mu_1 + \lambda_2(1-z)$.

$$B(w) = \begin{pmatrix} \beta_0(w) & -\mu_2 & 0 & \cdots & \cdots & 0 \\ -\lambda_2 & \beta(w) & -\mu_2 & 0 & \cdots & 0 \\ 0 & -\lambda_2 & \beta(w) & -\mu_2 & 0 & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & -\lambda_2 & \beta(w) & -\mu_2 \\ 0 & \cdots & \cdots & 0 & -\lambda_2 & \beta_{K_2}(w) \end{pmatrix},$$

where $\beta_0(w) = \lambda_2 + \lambda_1(1-w)$; $\beta(w) = \lambda_2 + \mu_2 + \lambda_1(1-w)$; $\beta_{K_2}(w) = \mu_2 + \lambda_1(1-w)$.

To obtain $G_k(z)$ and $F_n(w)$ we use Cramer's rule. I.e., for every $0 \le k \le K_1$, $G_k(z) = \frac{|A_k(z)|}{|A(z)|}$, where $|A|$ is the determinant of the matrix $A$ and $A_k(z)$ is the matrix obtained from $A(z)$ by replacing its $k$-th column by $\vec{P}(z)$. Similarly, $F_n(w) = \frac{|B_n(w)|}{|B(w)|}$, for every $0 \le n \le K_2$. Thus, each PGF $G_k(z)$ and $F_n(w)$ is expressed in terms of the $K_1 + K_2 + 2$ unknown probabilities, $P_{0,K_2-1}(1)$; $P_{1,K_2-1}(1)$, ..., $P_{K_1-1,K_2-1}(1)$, $P_{K_1K_2}(1)$; $P_{K_1-1,0}(2)$, $P_{K_1-1,1}(2)$, ..., $P_{K_1-1,K_2-1}(2)$ and $P_{K_1K_2}(2)$ appearing in $\vec{P}(z)$ and $\vec{\Pi}(w)$. In order to find $\vec{P}(z)$ and $\vec{\Pi}(w)$ we need to find $K_1 + K_2 + 2$ equations relating those $K_1 + K_2 + 2$ unknowns. We do that in the next subsection by characterizing and using the roots of $|A(z)|$ and $|B(w)|$. Since $G_k(z)$ and $F_n(w)$ are probability generating functions, defined in our finite state space model for all reals $z$ and $w$, each root of $|A(z)|$ (respectively of $|B(w)|$) is a root of $|A_k(z)|$ (respectively of $|B_n(w)|$), for every $0 \le k \le K_1$ (respectively for every $0 \le n \le K_2$).

*Remark* 2.1. It should be noted that, in our specific model, the matrix $A(z)$ can be presented as $A(z) = A - \lambda_A I$, where $\lambda_A = \lambda_2 z$ and $A$ is a tridiagonal Jacobi matrix (see e.g., Da Fonseca (2006)) with a constant element along the upper diagonal and another constant along the lower diagonal. $A$ is given by

$$A = \begin{pmatrix} \lambda_1 + \lambda_2 & -\mu_1 & 0 & \cdots & \cdots & 0 \\ -\lambda_1 & \lambda_1 + \mu_1 + \lambda_2 & -\mu_1 & 0 & \cdots & 0 \\ 0 & -\lambda_1 & \lambda_1 + \mu_1 + \lambda_2 & -\mu_1 & 0 & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & -\lambda_1 & \lambda_1 + \mu_1 + \lambda_2 & -\mu_1 \\ 0 & \cdots & \cdots & 0 & -\lambda_1 & \mu_1 + \lambda_2 \end{pmatrix}.$$

Clearly, the roots of the equation $|A(z)| = 0$ coincide with the roots of $|A - \lambda_A I| = 0$. Thus, the various eigenvalues of the matrix $A$ (divided by $\lambda_2$) are the roots of the polynomial $|A(z)| = 0$. The same holds for the matrix $B(w)$. Note that this relation to eigenvalues does not occur in other studies where roots of the polynomial $|A(z)| = 0$ are sought (see e.g., Litvak and Yechiali (2003); Perel E. and Yechiali (2008, 2013)).

## Derivation of the Boundary Probabilities

**Theorem 2.1.** *For any $\lambda_1 > 0$, $\mu_1 > 0$, $\lambda_2 > 0$ and $K_1 \geq 1$, $|A(z)|$ is a polynomial of degree $K_1 + 1$ possessing a single root $z^* = 1$ and $K_1$ distinct roots in the open interval $(1, \infty)$.*

*Proof.* Let $q_0(z) = 1$. Define the minors of the diagonal of $A(z)$, starting from the higher left side corner, as follows (see also Usmani (1994)):

$$q_1(z) = \alpha_0(z), \quad q_2(z) = \begin{vmatrix} \alpha_0(z) & -\mu_1 \\ -\lambda_1 & \alpha(z) \end{vmatrix}, ..., \quad q_{K_1+1}(z) = |A(z)|. \tag{2.14}$$

The polynomials $q_k(z)$, $1 \leq k \leq K_1 + 1$, satisfy the following recursions:

$$\begin{aligned} q_1(z) &= \alpha_0(z) q_0(z), \\ q_k(z) &= \alpha(z) q_{k-1}(z) - \lambda_1 \mu_1 q_{k-2}(z) \ , \quad \text{for} \quad 2 \leq k \leq K_1, \\ q_{K_1+1}(z) &= \alpha_K(z) q_{K_1}(z) - \lambda_1 \mu_1 q_{K_1-1}(z). \end{aligned} \tag{2.15}$$

From (2.14) and (2.15) we conclude that

1. By definition, $q_0(z) = 1$ and therefore has no roots.
2. $q_k(z)$ and $q_{k+1}(z)$ have no joint roots in $(0, \infty)$. Otherwise, suppose they have a joint root, then it would also be a root for $q_{k-1}(z)$, $q_{k-2}(z)$, ..., $q_0(z)$ which contradicts the above conclusion.
3. $\text{Sign}(q_k(\infty)) = (-1)^k$, for all $k$.
4. $q_k(1) = \lambda_1^k > 0$, for all $0 \leq k \leq K_1$.
5. $q_{K_1+1}(1) = 0$.
6. Given $\tilde{z}$ a root of $q_k(z)$, then $\text{sign}(q_{k-1}(\tilde{z}) q_{k+1}(\tilde{z})) = -1$.
7. $q_k(z)$ is a polynomial of degree $k$ for all $0 \leq k \leq K_1 + 1$.

From the above we conclude that $q_1(z)$ has only one root, $z_{1,1} = 1 + \frac{\lambda_1}{\lambda_2} > 1$. $q_2(1) = \lambda_1^2 > 0$, $q_2(z_{1,1}) < 0$, $q_2(\infty) > 0$. Therefore, the 2 roots of $q_2(z)$ satisfy: $z_{2,1} \in (1, z_{1,1})$, $z_{2,2} \in (z_{1,1}, \infty)$. Similarly, $q_3(z)$ is of degree 3 and therefore can have no more than 3 distinct roots. Also $q_3(1) = \lambda_1^3 > 0$, $q_3(z_{2,1}) < 0$, $q_3(z_{2,2}) > 0$, $q_3(\infty) < 0$. This implies that $q_3(z)$ has exactly 3 distinct roots satisfying: $z_{3,1} \in (1, z_{2,1})$, $z_{3,2} \in (z_{2,1}, z_{2,2})$, $z_{3,3} \in (z_{2,2}, \infty)$.
In general, for $2 \leq k \leq K_1$, given $k - 1$ distinct roots of $q_{k-1}(z)$, the roots of $q_k(z)$ satisfy: $z_{k,1} \in (1, z_{k-1,1})$, $z_{k,2} \in (z_{k-1,1}, z_{k-1,2})$, ..., $z_{k,k} \in (z_{k-1,k-1}, \infty)$.
$q_{K_1+1}(z)$ has $K_1 + 1$ roots, where $z_{K_1+1,1} = z^* = 1$ is one of them. From the above we have another $K_1$ distinct roots in the open interval $(1, \infty)$, satisfying $z_{K_1+1,2} \in (z_{K_1,1}, z_{K_1,2})$, $z_{K_1+1,3} \in (z_{K_1,2}, z_{K_1,3})$, ..., $z_{K_1+1,K_1} \in (z_{K_1,K_1-1}, z_{K_1,K_1})$, $z_{K_1+1,K_1+1} \in (z_{K_1,K_1}, \infty)$.
This completes the proof of Theorem 2.1. □

**Theorem 2.2.** *For any $\lambda_2 > 0$, $\mu_2 > 0$, $\lambda_1 > 0$ and $K_2 \geq 1$, $|B(w)|$ is a polynomial of degree $K_2 + 1$ possessing a single root $w^* = 1$ and $K_2$ distinct roots in the open interval $(1, \infty)$.*

*Proof.* The proof is identical to the proof of Theorem 2.1. □

To find the $K_1 + K_2 + 2$ unknown probabilities appearing in $\vec{P}(z)$ and $\vec{\Pi}(w)$, we use the $K_1$ distinct roots in the open interval $(1, \infty)$ of the polynomial $|A(z)|$ and the $K_2$ distinct roots in the open interval $(1, \infty)$ of the polynomial $|B(w)|$, which provide us with $K_1 + K_2$ equations for those probabilities. Another equation is (2.4), and the last one is provided by using the fact that $\sum_{k=0}^{K_1} G_k(1) + \sum_{n=0}^{K_2} F_n(1) = 1$. Thus, we have a set of $K_1 + K_2 + 2$ linearly independent equations in the $K_1 + K_2 + 2$ unknown probabilities (see e.g. Litvak and Yechiali (2003); Perel E. and Yechiali (2008, 2013); Perel N. and Yechiali (2013)). Once the above $K_1 + K_2 + 2$ probabilities are calculated, the PGFs are completely determined.

*Remark* 2.2. We note that the polynomials $q_k(z)$, $1 \leq k \leq K_1 + 1$, are usually used to show that the roots of the determinant of the matrix $A(z)$ exist, and are inside or outside the $(0, 1)$ interval. We are not aware of a case where the polynomials were explicitly calculated. In our case we are able to derive explicit closed-form expressions for these polynomials.

**Theorem 2.3.** *For every $1 \leq k \leq K_1$, $q_k(z)$ is of the form*

$$q_k(z) = \sum_{l=0}^{\lfloor \frac{k}{2} \rfloor} (-\lambda_1 \mu_1)^l \left( \binom{k-l-1}{l} \alpha^{k-2l-1}(z) \alpha_0(z) + \binom{k-l-1}{l-1} \alpha^{k-2l}(z) \right), \tag{2.16}$$

*where, for all $k \geq 0$, $\binom{k}{l} = 0$ for every $l < 0$, and $l > k$.*

*Proof.* The proof is given in the Appendix. □

**Corollary 2.1.** *The polynomial $q_{K_1+1}(z) = |A(z)|$ is given by*

$$\begin{aligned} |A(z)| = & \sum_{l=0}^{\lfloor \frac{K_1+1}{2} \rfloor} (-\lambda_1 \mu_1)^l \left( \binom{K_1 - l}{l} \alpha^{K_1-2l}(z) \alpha_0(z) + \binom{K_1 - l}{l-1} \alpha^{K_1+1-2l}(z) \right) \\ & - \lambda_1 \sum_{l=0}^{\lfloor \frac{K_1}{2} \rfloor} (-\lambda_1 \mu_1)^l \left( \binom{K_1 - l - 1}{l} \alpha^{K_1-2l-1}(z) \alpha_0(z) + \binom{K_1 - 1 - l - 1}{l-1} \alpha^{K_1-2l}(z) \right). \end{aligned} \tag{2.17}$$

*Proof.* By substituting equation (2.16) in equation (2.15) once for $k = K_1$, and then for $k = K_1 - 1$. □

By symmetry we have the following:

**Corollary 2.2.**

$$\begin{aligned} |B(w)| = & \sum_{l=0}^{\lfloor \frac{K_2+1}{2} \rfloor} (-\lambda_2 \mu_2)^l \left( \binom{K_2 - l}{l} \beta^{K_2-2l}(w) \beta_0(w) + \binom{K_2 - l}{l-1} \beta^{K_2+1-2l}(w) \right) \\ & - \lambda_2 \sum_{l=0}^{\lfloor \frac{K_2}{2} \rfloor} (-\lambda_2 \mu_2)^l \left( \binom{K_2 - l - 1}{l} \beta^{K_2-2l-1}(w) \beta_0(w) + \binom{K_2 - l - 1}{l-1} \beta^{K_2-2l}(w) \right). \end{aligned} \tag{2.18}$$

The mean total number of customers in $Q_i$, $i = 1, 2$ is obtained by

$$\mathbb{E}[L_1] = \sum_{k=0}^{K_1} kP_{k\bullet}(1) + \sum_{k=0}^{K_1} kP_{k\bullet}(2) = \sum_{k=0}^{K_1} kG_k(1) + \sum_{n=0}^{K_2} F_n'(1), \tag{2.19}$$

$$\mathbb{E}[L_2] = \sum_{n=0}^{K_2} nF_n(1) + \sum_{k=0}^{K_1} G_k'(1). \tag{2.20}$$

Clearly, by Little's Law, $\mathbb{E}[W_i] = \mathbb{E}[L_i]/\lambda_i^{eff}$, where

$$\lambda_i^{eff} = \lambda_i \left(1 - \mathbb{P}\left(L_i = K_i\right)\right) = \lambda_i \left(1 - \mathbb{P}\left(\text{arriving customer is lost in } Q_i\right)\right).$$

$\mathbb{P}(I = 1)$ and $\mathbb{P}(I = 2)$ are obtained by

$$\mathbb{P}(I = 1) = \sum_{k=0}^{K_1} G_k(1), \tag{2.21}$$

$$\mathbb{P}(I = 2) = \sum_{n=0}^{K_2} F_n(1). \tag{2.22}$$

The above performance measures, $\mathbb{E}[L_i]$, $\mathbb{E}[W_i]$, $\mathbb{P}\left(\text{arriving customer is lost in } Q_i\right)$, and $\mathbb{P}(I = i)$, $i = 1, 2$, are calculated numerically in Section 4.

## 2.4 Matrix Analysis

We define a non-reducible Markov chain $(I(t), L_1(t), L_2(t))$ with a finite state space $S = \{(i, k, n)|\ i = 1, 2;\ 0 \le k \le K_1;\ 0 \le n \le K_2\}$ under the order $S = \{(1, 0, 0),\ (1, 0, 1),\ ...,(1, 0, K_2 - 1);\ ...;\ (1, K_1 - 1, 0),\ ...,\ (1, K_1 - 1, K_2 - 1);\ (1, K_1, 0),\ ...,\ (1, K_1, K_2 - 1),\ (1, K_1, K_2);\ (2, 0, 0),\ (2, 1, 0),...,\ (2, K_1 - 1, 0);\ ...;\ (2, 0, K_2 - 1),\ ...,\ (2, K_1 - 1, K_2 - 1);\ (2, 0, K_2),\ ...,\ (2, K_1 - 1, K_2),\ (2, K_1, K_2)\}$ and a generator matrix $Q$ given by

$$Q = \begin{pmatrix} A_1^0 & A_0 & \mathbf{0} & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & A_3^0 \\ A_2 & A_1 & A_0 & \mathbf{0} & \cdots & \cdots & \cdots & \cdots & \cdots & A_3^1 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \mathbf{0} & A_2 & A_1 & A_0 & \mathbf{0} & \cdots & \cdots & \cdots & A_3^{K_1-1} \\ \vdots & \vdots & \mathbf{0} & A_2^{K_1} & A_1^{K_1} & \mathbf{0} & \cdots & \cdots & \cdots & A_3^{K_1} \\ \vdots & \ddots & \ddots & \mathbf{0} & B_3^0 & B_1^0 & B_0 & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \ddots & \ddots & \mathbf{0} & B_3^1 & B_2 & B_1 & B_0 & \mathbf{0} & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{0} & B_3^{K_2-1} & \mathbf{0} & \cdots & B_2 & B_1 & B_0 \\ \vdots & \ddots & \ddots & \mathbf{0} & B_3^{K_2} & \mathbf{0} & \cdots & \cdots & B_2^{K_2} & B_1^{K_2} \end{pmatrix} = \begin{pmatrix} \mathcal{A} & \mathbf{0}_1 & \vec{\boldsymbol{A}}_3 \\ \mathbf{0}_2 & \vec{\boldsymbol{B}}_3 & \mathcal{B} \end{pmatrix}, \tag{2.23}$$

where, $\mathbf{0}$ is a matrix of zeros, and the sub-matrices in $Q$ are the following: $A_1^0$, $A_0$, $A_2$ and $A_1$ are each of

size $K_2 \times K_2$; $A_1^{K_1}$ is of size $(K_2+1)\times(K_2+1)$; $A_2^{K_1}$ is of size $(K_2+1)\times K_2$; $A_3^0$, $A_3^1$, ..., $A_3^{K_1-1}$ are each of size $K_2 \times (K_1+1)$; $A_3^{K_1}$ is of size $(K_2+1)\times(K_1+1)$; $B_1^0$ , $B_0$, $B_2$ and $B_1$ are each of size $K_1 \times K_1$; $B_1^{K_2}$ is of size $(K_1+1)\times(K_1+1)$; $B_2^{K_2}$ is of size $(K_1+1)\times K_1$; $B_3^0$, $B_3^1$, ..., $B_3^{K_2-1}$ are each of size $K_1 \times (K_2+1)$ and $B_3^{K_2}$ is of size $(K_1+1)\times(K_2+1)$. They are given by

$$A_0 = diag(\lambda_1), \quad A_2 = diag(\mu_1),$$

$$A_1^0 = \begin{pmatrix} -\beta_0 & \lambda_2 & 0 & \cdots & 0 \\ 0 & -\beta_0 & \lambda_2 & 0 & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & -\beta_0 & \lambda_2 \\ \vdots & \ddots & \ddots & 0 & -\beta_0 \end{pmatrix}, \quad A_1 = \begin{pmatrix} -\beta_1 & \lambda_2 & 0 & \cdots & 0 \\ 0 & -\beta_1 & \lambda_2 & 0 & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & -\beta_1 & \lambda_2 \\ \vdots & \ddots & \ddots & 0 & -\beta_1 \end{pmatrix},$$

where, $\beta_0 = \lambda_1 + \lambda_2$, $\beta_1 = \lambda_1 + \lambda_2 + \mu_1$ and $\beta_2 = \lambda_1 + \lambda_2 + \mu_2$.

$$A_1^{K_1} = \begin{pmatrix} -(\mu_1+\lambda_2) & \lambda_2 & 0 & \cdots & 0 \\ 0 & -(\mu_1+\lambda_2) & \lambda_2 & 0 & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & -(\mu_1+\lambda_2) & \lambda_2 \\ \vdots & \ddots & \ddots & 0 & -\mu_1 \end{pmatrix}, \quad A_2^{K_1} = \begin{pmatrix} \mu_1 & 0 & \cdots & \cdots & 0 \\ 0 & \mu_1 & 0 & \cdots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \mu_1 \\ 0 & \cdots & \cdots & \cdots & 0 \end{pmatrix}.$$

For all $0 \le k \le K_1 - 1$,

$$\left(A_3^k\right)_{(st),(lm)} = \begin{cases} \lambda_2 & (st) = (k, K_2-1), (lm) = (kK_2), \\ 0 & otherwise \end{cases}$$

$$\left(A_3^{K_1}\right)_{(st),(lm)} = \begin{cases} \mu_1 & (st) = (K_1, K_2), (lm) = (K_1-1, K_2), \\ 0 & otherwise \end{cases}$$

The matrices $B_0$, $B_1^0$, $B_1$, $B_1^{K_2}$, $B_2$, and $B_2^{K_2}$ are similar to $A_0$, $A_1^0$, $A_1$, $A_1^{K_1}$, $A_2$, and $A_2^{K_1}$, respectively, where $\lambda_2$, $\mu_2$, $\beta_2$, and $K_2$ change role with $\lambda_1$, $\mu_1$, $\beta_1$, and $K_1$, respectively. Finally, for all $0 \le n \le K_2 - 1$,

$$\left(B_3^n\right)_{(st),(lm)} = \begin{cases} \lambda_1 & (st) = (K_1-1, n), (lm) = (K_1 n), \\ 0 & otherwise \end{cases}$$

$$\left(B_3^{K_2}\right)_{(st),(lm)} = \begin{cases} \mu_2 & (st) = (K_1, K_2), (lm) = (K_1, K_2-1), \\ 0 & otherwise \end{cases}$$

We emphasize that the structure of the matrix $Q$ is *different* from the structure of a classical QBD process generator matrix. Usually the generator matrix of a QBD process has at most 3 non zero block-diagonals and all other entries are zero. In this case, the matrix $Q$ has 3 non zero block-diagonals (appearing in the

matrices $\mathcal{A}$ and $\mathcal{B}$) and in addition two non zero columns, $\vec{\boldsymbol{A}}_3$ and $\vec{\boldsymbol{B}}_3$ as in (2.23).
Define the steady state probability vector $\vec{P} = \left(\vec{P}_0^1, \vec{P}_1^1, ..., \vec{P}_{K_1}^1, \vec{P}_0^2, \vec{P}_1^2, ..., \vec{P}_{K_2}^2\right)$, where

$$\begin{aligned}
&\vec{P}_k^1 = (P_{k0}(1), P_{k1}(1), ..., P_{k,K_2-1}(1)), && 0 \le k \le K_1 - 1,\\
&\vec{P}_{K_1}^1 = (P_{K_1 0}(1), ..., P_{K_1,K_2-1}(1), P_{K_1 K_2}(1)),\\
&\vec{P}_n^2 = (P_{0n}(2), P_{1n}(2), ..., P_{K_1-1,n}(2)), && 0 \le n \le K_2 - 1,\\
&\vec{P}_{K_2}^2 = (P_{0K_2}(2), ..., P_{K_1-1,K_2}(2), P_{K_1 K_2}(2)).
\end{aligned}$$

Then the steady state probability vector satisfies $\vec{P}Q = \vec{0}$. Specifically, we have

$$\begin{aligned}
&\vec{P}_0^1 A_1^0 + \vec{P}_1^1 A_2 = \vec{0},\\
&\vec{P}_0^1 A_0 + \vec{P}_1^1 A_1 + \vec{P}_2^1 A_2 = \vec{0},\\
&\vdots\\
&\vec{P}_{K_1-2}^1 A_0 + \vec{P}_{K_1-1}^1 A_1 + \vec{P}_{K_1}^1 A_2^{K_1} = \vec{0},\\
&\vec{P}_{K_1-1}^1 A_0 + \vec{P}_{K_1-1}^1 A_1^{K_1} + \sum_{n=0}^{K_2} \vec{P}_n^2 B_3^n = \vec{0},\\
&\vec{P}_0^2 B_1^0 + \vec{P}_1^2 B_2 = \vec{0},\\
&\vec{P}_0^2 B_0 + \vec{P}_1^2 B_1 + \vec{P}_2^2 B_2 = \vec{0},\\
&\vdots\\
&\vec{P}_{K_2-2}^2 B_0 + \vec{P}_{K_2-1}^2 B_1 + \vec{P}_{K_2}^2 B_2^{K_2} = \vec{0},\\
&\sum_{k=0}^{K_1} \vec{P}_k^1 A_3^k + \vec{P}_{K_2-1}^2 B_0 + \vec{P}_{K_2}^2 B_1^{K_2} = \vec{0}.
\end{aligned} \tag{2.24}$$

Clearly, one can solve directly (numerically) the set (2.24) (including the normalization equation, $\sum_{k=0}^{K_1} \vec{P}_k^1 \cdot \vec{e} + \sum_{n=0}^{K_2} \vec{P}_n^2 \cdot \vec{e} = 1$, where $\vec{e}$ is a vector of 1's). This requires some computational effort. We indicate again that in contrast to most analytic methods in Queueing Theory that treat cases where at least one of the process dimension is infinite, our case is finite in all its dimensions. Thus traditional Matrix Geometric methods are not directly applicable here. In addition, truncation methods seem unnecessary here. Therefore, we present an *alternative* algorithmic-type method to ease the required computational effort. For this aim we borrow from the ideas presented by De Nitto Personè and Grassi (1996) and modify them to our purposes due to the special structure of the matrix $Q$. We claim:

**Theorem 2.4.** *The following equation holds:*

$$\vec{P}_k^1 = \vec{P}_0^1 C_k, \qquad 1 \le k \le K_1 - 1, \tag{2.25}$$

*where for all $1 \le k \le K_1 - 1$, $C_k$ is a matrix of size $K_2 \times K_2$ satisfying $C_0 = I_{K_2}$, $C_1 = -\frac{1}{\mu_1} A_1^0$ and for all $2 \le k \le K_1 - 1$, $C_k = -\frac{1}{\mu_1}\left(\lambda_1 C_{k-2} + C_{k-1} A_1\right)$.*

*Proof.* The proof is given in the Appendix. □

**Theorem 2.5.** *The following equation holds:*

$$\vec{P}_n^2 = \vec{P}_0^2 D_n, \qquad 1 \le n \le K_2 - 1, \tag{2.26}$$

*where for all* $1 \le n \le K_2 - 1$, $D_n$ *is a matrix of size* $K_1 \times K_1$ *satisfying* $D_0 = I_{K_2}$, $D_1 = -\frac{1}{\mu_2} B_1^0$ *and for all* $2 \le n \le K_2 - 1$, $D_n = -\frac{1}{\mu_2}\left(\lambda_2 D_{n-2} + D_{n-1} B_1\right)$.

*Proof.* The proof is identical to the proof of Theorem 2.4. □

From Theorem 2.4 we have that $\vec{P}_k^1$, $1 \le k \le K_1 - 1$, are expressed in terms of the probability vector $\vec{P}_0^1$, and from Theorem 2.5 we have that $\vec{P}_n^2$, $1 \le n \le K_2 - 1$, are expressed in terms of the probability vector $\vec{P}_0^2$. Therefore, the solution of (2.24) can be calculated by solving only the following reduced linear system:

$$\begin{aligned}
&\vec{P}_0^1\left(\lambda_1 C_{K_1-2} + C_{K-1} A_1\right) + \vec{P}_{K_1}^1 A_2^{K_1} = \vec{0},\\
&\vec{P}_0^1 \lambda_1 C_{K_1-1} + \vec{P}_{K_1}^1 A_1^{K_1} + \vec{P}_0^2 \sum_{n=0}^{K_2-1} D_n B_3^n + \vec{P}_{K_2}^2 B_3^{K_2} = \vec{0},\\
&\vec{P}_0^2\left(\lambda_2 D_{K_2-2} + D_{K_2-1} B_1\right) + \vec{P}_{K_2}^2 B_2^{K_2} = \vec{0},\\
&\vec{P}_0^1 \sum_{k=0}^{K_1-1} C_k A_3^k + \vec{P}_{K_1}^1 A_3^{K_1} + \vec{P}_0^2 \lambda_2 D_{K_2-1} + \vec{P}_{K_2}^2 B_1^{K_2} = \vec{0}.
\end{aligned} \tag{2.27}$$

Therefore, instead of solving tediously the set of linear equations (2.24), it is enough to calculate the matrices $C_k$, $1 \le k \le K_1 - 1$ and $D_n$, $1 \le n \le K_2 - 1$, and solve the set of linear equations (2.27) with the normalization equation, $\vec{P}_0^1 \sum_{k=0}^{K_1-1} C_k \vec{e} + \vec{P}_{K_1}^1 \vec{e} + \vec{P}_0^2 \sum_{n=0}^{K_2-1} D_n \vec{e} + \vec{P}_{K_2}^2 \vec{e} = 1$, which yields the set of sought-for probability vectors $\vec{P}_0^1$, ..., $\vec{P}_{K_1-1}^1$, $\vec{P}_{K_1}^1$, $\vec{P}_0^2$, ..., $\vec{P}_{K_2-1}^2$, $\vec{P}_{K_2}^2$.

**Theorem 2.6.** *For every* $0 \le k \le K_1 - 1$, $C_k$ *is of the form*

$$C_k = \left(\frac{-1}{\mu_1}\right)^k \sum_{l=0}^{\lfloor \frac{k}{2} \rfloor} \left(-\lambda_1 \mu_1\right)^l \left(\binom{k-1-l}{l} A_1^0 \left(A_1\right)^{k-2l-1} + \binom{k-1-l}{l-1} \left(A_1\right)^{k-2l}\right). \tag{2.28}$$

*where, for all* $k \ge 0$, $\binom{k}{l} = 0$ *for every* $l < 0$, *and* $l > k$.

*Proof.* The proof is given in the Appendix. □

*Remark* 2.3. Notice the similarity in structure of equations (2.16) and (2.28).

**Theorem 2.7.** *For every* $0 \le n \le K_2 - 1$, $D_n$ *is of the form*

$$D_n = \left(\frac{-1}{\mu_2}\right)^n \sum_{l=0}^{\lfloor \frac{n}{2} \rfloor} \left(-\lambda_2 \mu_2\right)^l \left(\binom{n-1-l}{l} B_1^0 \left(B_1\right)^{n-2l-1} + \binom{n-1-l}{l-1} \left(B_1\right)^{n-2l}\right). \tag{2.29}$$

*Proof.* The proof is similar to the proof of Theorem 2.6 □

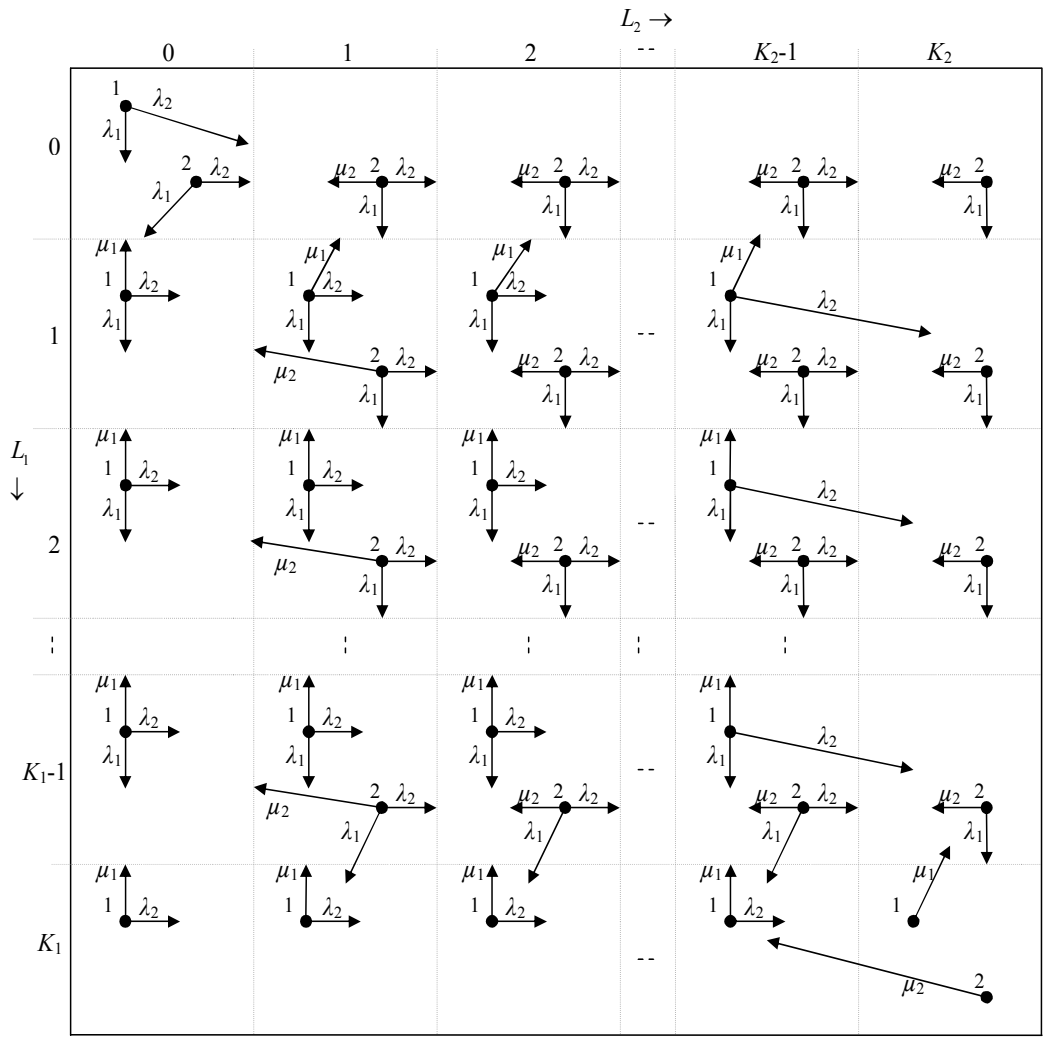


Figure 3.1: Transition rate diagram of $(L_1, L_2, I)$ for Scenario 2.

# 3 Scenario 2: Work-Conserving Policy

We now briefly present the work-conserving switching scenario: if a served $Q_i$ becomes empty, the server immediately switches to $Q_j$ if the latter is not empty. The transition-rate diagram of the triple $(L_1, L_2, I)$ for this scenario is depicted in Figure 3.1. The balance equations here are similar to those of Scenario 1, and therefore the details are omitted from the presentation. Equation (2.4), equating the switching rates between the queues, becomes

$$\mu_1 \left(P_{1\bullet}(1) - P_{10}(1) + P_{K_1K_2}(1)\right) + \lambda_2 \left(P_{00}(1) + P_{\bullet K_2-1}(1) - P_{K_1,K_2-1}(1)\right) = \\ \mu_2 \left(P_{\bullet 1}(2) - P_{01}(2) + P_{K_1K_2}(2)\right) + \lambda_1 \left(P_{00}(2) + P_{K_1-1\bullet}(2) - P_{K_1-1,K_2}(2)\right). \tag{3.1}$$

Notice that in this case a switch occurs also when the served queue becomes empty and the non served queue is not empty.

Repeating the algebra in Subsection 2.2, we arrive at

$$P_{00}(1) + P_{00}(2) = \left(\mathbb{P}\left(I = 1\right) - \rho_1^{eff}\right) + \left(\mathbb{P}\left(I = 2\right) - \rho_2^{eff}\right) = 1 - \left(\rho_1^{eff} + \rho_2^{eff}\right), \tag{3.2}$$

### 3.1 Matrix Analysis for Scenario 2

We define a non-reducible Markov chain $(L_1(t), L_2(t), I(t))$ with a finite state space $S = \{(k, n, i)|\ 0 \leq k \leq K_1;\ 0 \leq n \leq K_2;\ i = 1, 2\}$ under a *different order* $S = \{(0,0,1),(0,0,2),(1,0,1),\ (2,0,1),\ ...,\ (K_1,0,1)\ ;\ (0,1,2),\ (1,1,1),\ (1,1,2),...,\ (K_1-1,1,1),\ (K_1-1,1,2),\ (K_1,1,1);\ ...;\ (0,K_2-1,2),...,\ (K_1-1,K_2-1,1),\ (K_1-1,K_2-1,2),\ (K_1,K_2-1,1);\ (0,K_2,2),\ (1,K_2,2),...,\ (K_1-1,K_2,2),\ (K_1,K_2,1),\ (K_1,K_2,2)\}$. This lexicographic order leads to a different generator matrix $Q$ from which all steady state probability vectors can be calculated (details are omitted). The two scenarios are compared numerically in the next sections.

## 4 Numerical Examples

Define $P_{\text{loss}}(i) = \mathbb{P}(\text{arriving customer is lost in } Q_i)$, $i = 1, 2$, and $SR \equiv$ average switching rate between the queues. Tables 4.1 – 4.4 exhibit numerical results when $K_1 = 10$ for different values of $\lambda_1$, $\lambda_2$, $\mu_1$, and $\mu_2$. In Tables 4.1 – 4.2, $K_2 = 3$, while in Tables 4.3 – 4.4, $K_2 = 8$. Tables 4.1 and 4.3 relate to Scenario 1, while Tables 4.2 and 4.4 relate to Scenario 2. Tables 4.1 – 4.4 are constructed as follows: the first row gives a set of basic values of the four parameters $\lambda_1 = 1$, $\lambda_2 = 1$, $\mu_1 = 5$ and $\mu_2 = 5$, and the calculations of the resulting performance measures. The second, third, fourth, and fifth rows give, respectively, the values of the measures when in each row only one of the parameters is changed. Row six gives another set of basic parameters, $\lambda_1 = 4$, $\lambda_2 = 4$, $\mu_1 = 2$ and $\mu_2 = 2$, and the following rows give the corresponding results.

Table 4.1: Numerical Results for $K_1 = 10$ and $K_2 = 3$. Non-Work-Conserving

| | $\mathbb{E}[L_1]$ | $\mathbb{E}[L_2]$ | $\mathbb{P}(I=1)$ | $\mathbb{P}(I=2)$ | $P_{\text{loss}}(1)$ | $P_{\text{loss}}(2)$ | $\mathbb{E}[W_1]$ | $\mathbb{E}[W_2]$ | $SR$ |
|---|---|---|---|---|---|---|---|---|---|
| **Basic parameters 1** | 4.92 | 0.5853 | 0.2678 | 0.7322 | 0.0247 | 0.0289 | 5.0447 | 0.6028 | 0.0573 |
| $\lambda_1 = 10$ | 9.2883 | 2.0230 | 0.8354 | 0.1646 | 0.5824 | 0.2033 | 2.2241 | 2.5393 | 0.6366 |
| $\lambda_2 = 10$ | 8.9856 | 2.4148 | 0.1590 | 0.8410 | 0.2050 | 0.5996 | 11.3025 | 0.6032 | 0.6369 |
| $\mu_1 = 10$ | 4.1234 | 0.5350 | 0.2307 | 0.7693 | 0.0095 | 0.0252 | 4.1629 | 0.5488 | 0.0206 |
| $\mu_2 = 10$ | 4.8464 | 0.4200 | 0.2713 | 0.7287 | 0.0234 | 0.0114 | 4.9626 | 0.4248 | 0.0520 |
| **Basic parameters 2** | 9.7206 | 2.7247 | 0.4995 | 0.5005 | 0.7502 | 0.7508 | 9.7300 | 2.7339 | 0.7654 |
| $\lambda_1 = 40$ | 9.9699 | 2.7952 | 0.5912 | 0.4088 | 0.9704 | 0.7956 | 8.4312 | 3.4192 | 0.8123 |
| $\lambda_2 = 40$ | 9.7952 | 2.9700 | 0.4087 | 0.5913 | 0.7956 | 0.9704 | 11.9820 | 2.5117 | 0.8123 |
| $\mu_1 = 20$ | 6.8523 | 2.4698 | 0.1776 | 0.8224 | 0.2271 | 0.6001 | 2.2163 | 1.5442 | 0.5168 |
| $\mu_2 = 20$ | 9.5333 | 1.7123 | 0.7499 | 0.2501 | 0.6251 | 0.3415 | 6.3566 | 0.6501 | 0.8539 |

Table 4.2: Numerical Results for $K_1 = 10$ and $K_2 = 3$. Work-Conserving

| | $\mathbb{E}[L_1]$ | $\mathbb{E}[L_2]$ | $\mathbb{P}(I=1)$ | $\mathbb{P}(I=2)$ | $P_{\text{loss}}(1)$ | $P_{\text{loss}}(2)$ | $\mathbb{E}[W_1]$ | $\mathbb{E}[W_2]$ | $SR$ |
|---|---|---|---|---|---|---|---|---|---|
| **Basic parameters 1** | 0.3495 | 0.3110 | 0.5025 | 0.4975 | 0.0001 | 0.0096 | 0.3495 | 0.3140 | 0.4566 |
| $\lambda_1 = 10$ | 9.2747 | 2.0258 | 0.8407 | 0.1593 | 0.5797 | 0.2036 | 2.2065 | 2.5435 | 0.6383 |
| $\lambda_2 = 10$ | 7.3468 | 2.4167 | 0.1777 | 0.8223 | 0.1136 | 0.5895 | 8.2886 | 0.5887 | 0.6589 |
| $\mu_1 = 10$ | 0.1795 | 0.2587 | 0.4834 | 0.5266 | $2.2{\cdot}10^{-7}$ | 0.0070 | 0.1795 | 0.2605 | 0.4721 |
| $\mu_2 = 10$ | 0.2717 | 0.1690 | 0.5197 | 0.4803 | $2.3{\cdot}10^{-7}$ | 0.0021 | 0.2717 | 0.1694 | 0.4756 |
| **Basic parameters 2** | 9.7161 | 2.7259 | 0.5021 | 0.4979 | 0.7489 | 0.7511 | 9.6749 | 2.7376 | 0.7676 |
| $\lambda_1 = 40$ | 9.9699 | 2.7952 | 0.5913 | 0.4087 | 0.9704 | 0.7956 | 8.4312 | 3.4192 | 0.8123 |
| $\lambda_2 = 40$ | 9.7951 | 2.9700 | 0.4088 | 0.5912 | 0.7956 | 0.9704 | 11.9809 | 2.5117 | 0.8129 |
| $\mu_1 = 20$ | 6.4101 | 2.4296 | 0.1700 | 0.8300 | 0.2078 | 0.5867 | 2.0230 | 1.4695 | 0.6386 |
| $\mu_2 = 20$ | 9.2975 | 1.8312 | 0.8692 | 0.1308 | 0.5654 | 0.3460 | 5.3487 | 0.7000 | 1.0043 |

Table 4.3: Numerical Results for $K_1 = 10$ and $K_2 = 8$. Non-Work-Conserving

| | $\mathbb{E}[L_1]$ | $\mathbb{E}[L_2]$ | $\mathbb{P}(I=1)$ | $\mathbb{P}(I=2)$ | $P_{\text{loss}}(1)$ | $P_{\text{loss}}(2)$ | $\mathbb{E}[W_1]$ | $\mathbb{E}[W_2]$ | $SR$ |
|---|---|---|---|---|---|---|---|---|---|
| **Basic parameters 1** | 3.3738 | 2.2050 | 0.4431 | 0.5569 | 0.0131 | 0.0132 | 3.4186 | 2.2346 | 0.0567 |
| $\lambda_1 = 10$ | 9.3124 | 6.9699 | 0.8279 | 0.1721 | 0.5897 | 0.2106 | 2.2698 | 8.8290 | 0.6361 |
| $\lambda_2 = 10$ | 8.9427 | 7.2870 | 0.1608 | 0.8391 | 0.1998 | 0.5815 | 11.1756 | 1.7412 | 0.6222 |
| $\mu_1 = 10$ | 2.9203 | 2.1956 | 0.4395 | 0.5605 | 0.0060 | 0.0138 | 2.9380 | 2.2262 | 0.0566 |
| $\mu_2 = 10$ | 3.3729 | 1.8686 | 0.4457 | 0.5543 | 0.0134 | 0.0060 | 3.4188 | 1.8798 | 0.0570 |
| **Basic parameters 2** | 9.7225 | 7.7217 | 0.4994 | 0.5006 | 0.7509 | 0.7509 | 9.7580 | 7.7498 | 0.7638 |
| $\lambda_1 = 40$ | 9.9700 | 7.7952 | 0.5912 | 0.4088 | 0.9704 | 0.7956 | 8.4316 | 9.5359 | 0.8123 |
| $\lambda_2 = 40$ | 9.7952 | 7.9700 | 0.4087 | 0.5913 | 0.7956 | 0.9704 | 11.9820 | 6.7399 | 0.8123 |
| $\mu_1 = 20$ | 6.9507 | 7.4388 | 0.1770 | 0.8230 | 0.2387 | 0.6037 | 2.2826 | 4.6922 | 0.5951 |
| $\mu_2 = 20$ | 9.4382 | 5.2031 | 0.8076 | 0.1924 | 0.5986 | 0.2364 | 5.8778 | 1.7035 | 0.6185 |

## Remarks on the Numerical Results

- In all cases presented, $\mathbb{E}[W_1]$ and $P_{\text{loss}}(1)$ under Non-Work-Conserving scenario are each larger than their corresponding values under Work-Conserving scenario. This is clearly a consequence of the difference between the scenarios
- The above observation is true for $\mathbb{E}[W_2]$ and $P_{\text{loss}}(2)$, for almost all cases, unless $\mu_2$ becomes relatively large.

In Section 6 we will further discuss the above numerical results together with those presented in the following Section 5 that analyzes extreme cases. In addition, in Section 6 we exhibit graphically the oscillation phenomenon occurring in this two-dimensional stochastic process.

Table 4.4: Numerical Results for $K_1 = 10$ and $K_2 = 8$. Work-Conserving

| | $\mathbb{E}[L_1]$ | $\mathbb{E}[L_2]$ | $\mathbb{P}(I=1)$ | $\mathbb{P}(I=2)$ | $P_{\text{loss}}(1)$ | $P_{\text{loss}}(2)$ | $\mathbb{E}[W_1]$ | $\mathbb{E}[W_2]$ | $SR$ |
|---|---|---|---|---|---|---|---|---|---|
| **Basic parameters 1** | 0.3334 | 0.3332 | 0.5000 | 0.5000 | $1.7{\cdot}10^{-6}$ | $1.6{\cdot}10^{-5}$ | 0.3334 | 0.3332 | 0.4518 |
| $\lambda_1 = 10$ | 9.2822 | 6.9192 | 0.8395 | 0.1605 | 0.5802 | 0.1977 | 2.2113 | 8.6239 | 0.6197 |
| $\lambda_2 = 10$ | 8.8860 | 7.2884 | 0.1608 | 0.8392 | 0.1959 | 0.5804 | 11.0515 | 1.7370 | 0.6168 |
| $\mu_1 = 10$ | 0.1814 | 0.2664 | 0.4825 | 0.5175 | $3.6{\cdot}10^{-7}$ | $2.4{\cdot}10^{-6}$ | 0.1814 | 0.2664 | 0.4711 |
| $\mu_2 = 10$ | 0.2665 | 0.1814 | 0.5177 | 0.4825 | $1.1{\cdot}10^{-7}$ | $3.2{\cdot}10^{-6}$ | 0.2665 | 0.1814 | 0.4711 |
| **Basic parameters 2** | 9.7201 | 7.7201 | 0.5000 | 0.5000 | 0.7500 | 0.7509 | 9.7201 | 7.7201 | 0.7648 |
| $\lambda_1 = 40$ | 9.9699 | 7.7952 | 0.5913 | 0.4087 | 0.9704 | 0.7956 | 8.4312 | 9.5354 | 0.8123 |
| $\lambda_2 = 40$ | 9.7952 | 7.9700 | 0.4087 | 0.5913 | 0.7956 | 0.9704 | 11.9819 | 6.7399 | 0.8123 |
| $\mu_1 = 20$ | 6.9360 | 7.3484 | 0.1550 | 0.8450 | 0.2258 | 0.5775 | 2.2398 | 4.3482 | 0.6012 |
| $\mu_2 = 20$ | 9.3398 | 5.3813 | 0.8482 | 0.1518 | 0.5759 | 0.2411 | 5.5057 | 1.7728 | 0.6434 |

# 5 Extreme Cases

Due to the symmetry between the queues, we investigate only the influence of extreme values of $\lambda_1$ and $\mu_1$, as they reach 0 or $\infty$, on the system's performance measures $\mathbb{E}[L_i]$, $\mathbb{E}[W_i]$, $\mathbb{P}(I = i)$, $P_{\text{loss}}(i)$, $i = 1, 2$, and $SR$. We first address extreme cases which lead to identical system structure in the two scenarios, and then we address extreme cases which lead to different system structures.

Numerical results for the case where $K_1 = 10$ are presented in Tables 5.1, 5.2, 5.5 and 5.6 (change in $\lambda_1$), and in Tables 5.3, 5.4, 5.7 and 5.8 (change in $\mu_1$). In Tables 5.1– 5.4 we set $K_2 = 3$, while in Tables 5.5 – 5.8, $K_2 = 8$.

Table 5.1: The impact of $\lambda_1$ (when $\lambda_2 = 4$, $\mu_1 = 7$, $\mu_2 = 7$, $K_1 = 10$ and $K_2 = 3$). Non-Work-Conserving

| | $\mathbb{E}[L_1]$ | $\mathbb{E}[L_2]$ | $\mathbb{P}(I=1)$ | $\mathbb{P}(I=2)$ | $P_{\text{loss}}(1)$ | $P_{\text{loss}}(2)$ | $\mathbb{E}[W_1]$ | $\mathbb{E}[W_2]$ | $SR$ |
|---|---|---|---|---|---|---|---|---|---|
| $\lambda_1 = 0$ | 0 | 0.8559 | 0 | 1 | 0 | 0.0895 | 0 | 0.2350 | 0 |
| $\lambda_1 = 0.1$ | 6.5305 | 0.8797 | 0.0149 | 0.9851 | 0.0041 | 0.0937 | 65.5725 | 0.2426 | 0.0239 |
| $\lambda_1 = 0.5$ | 6.6867 | 0.9733 | 0.0727 | 0.9273 | 0.0214 | 0.1098 | 13.6666 | 0.2733 | 0.1206 |
| $\lambda_1 = 5$ | 8.4174 | 1.8150 | 0.5011 | 0.4989 | 0.3008 | 0.2849 | 2.4078 | 0.6345 | 1.1267 |
| $\lambda_1 = 50$ | 9.8891 | 2.5030 | 0.7209 | 0.2791 | 0.8991 | 0.5117 | 1.9596 | 1.2816 | 1.9054 |
| $\lambda_1 = 500$ | 9.9897 | 2.5314 | 0.7323 | 0.2677 | 0.9897 | 0.5315 | 1.9488 | 1.3501 | 1.8733 |
| $\lambda_1 = 5000000$ | 10 | 2.5333 | 0.7333 | 0.2667 | 0.999999 | 0.5333 | 1.9480 | 1.3571 | 1.8667 |

Table 5.2: The impact of $\lambda_1$ (when $\lambda_2 = 4$, $\mu_1 = 7$, $\mu_2 = 7$, $K_1 = 10$ and $K_2 = 3$). Work-Conserving

| | $\mathbb{E}[L_1]$ | $\mathbb{E}[L_2]$ | $\mathbb{P}(I=1)$ | $\mathbb{P}(I=2)$ | $P_{\text{loss}}(1)$ | $P_{\text{loss}}(2)$ | $\mathbb{E}[W_1]$ | $\mathbb{E}[W_2]$ | $SR$ |
|---|---|---|---|---|---|---|---|---|---|
| $\lambda_1 = 0$ | 0 | 0.8559 | 0 | 1 | 0 | 0.08951 | 0 | 0.2350 | 0 |
| $\lambda_1 = 0.1$ | 0.0394 | 0.8644 | 0.0298 | 0.9702 | $0.2.8{\cdot}10^{-14}$ | 0.0906 | 0.3937 | 0.2376 | 0.1000 |
| $\lambda_1 = 0.5$ | 0.2190 | 0.9027 | 0.1305 | 0.8695 | $0.7.8{\cdot}10^{-8}$ | 0.0959 | 0.4380 | 0.2496 | 0.4155 |
| $\lambda_1 = 5$ | 7.0230 | 1.7853 | 0.5745 | 0.4255 | 0.2069 | 0.2669 | 1.7711 | 0.6089 | 1.1676 |
| $\lambda_1 = 50$ | 9.8891 | 2.5030 | 0.7210 | 0.2790 | 0.8991 | 0.5117 | 1.9594 | 1.2816 | 1.9054 |
| $\lambda_1 = 500$ | 9.9897 | 2.5314 | 0.7323 | 0.2677 | 0.9897 | 0.5315 | 1.9488 | 1.3509 | 1.8733 |
| $\lambda_1 = 5000000$ | 10 | 2.5333 | 0.7333 | 0.2667 | 0.999999 | 0.5333 | 1.9480 | 1.3571 | 1.8667 |

Table 5.3: The impact of $\mu_1$ (when $\lambda_1 = 4$, $\lambda_2 = 4$, $\mu_2 = 7$, $K_1 = 10$ and $K_2 = 3$). Non-Work-Conserving

| | $\mathbb{E}[L_1]$ | $\mathbb{E}[L_2]$ | $\mathbb{P}(I=1)$ | $\mathbb{P}(I=2)$ | $P_{\text{loss}}(1)$ | $P_{\text{loss}}(2)$ | $\mathbb{E}[W_1]$ | $\mathbb{E}[W_2]$ | $SR$ |
|---|---|---|---|---|---|---|---|---|---|
| $\mu_1 = 0$ | 10 | 3 | 1 | 0 | 1 | 1 | $\infty$ | $\infty$ | 0 |
| $\mu_1 = 0.1$ | 9.9757 | 2.9298 | 0.9700 | 0.0300 | 0.99757 | 0.9532 | 101.837 | 15.6572 | 0.0950 |
| $\mu_1 = 0.5$ | 9.8893 | 2.7068 | 0.8750 | 0.1250 | 0.8906 | 0.8048 | 22.6050 | 3.4674 | 0.3952 |
| $\mu_1 = 5$ | 8.7921 | 1.8486 | 0.5163 | 0.4837 | 0.3551 | 0.3101 | 3.4083 | 0.6699 | 1.0642 |
| $\mu_1 = 50$ | 4.2667 | 1.1740 | 0.2000 | 0.8000 | 0.0141 | 0.1398 | 1.0820 | 0.3412 | 0.1123 |
| $\mu_1 = 500$ | 3.7556 | 1.1431 | 0.1830 | 0.8170 | 0.0050 | 0.1346 | 0.9436 | 0.3302 | 0.0093 |
| $\mu_1 = 5000000$ | 3.7060 | 1.1406 | 0.1816 | 0.8184 | 0.0042 | 0.1342 | 0.9304 | 0.3293 | $9.1{\cdot}10^{-7}$ |

Table 5.4: The impact of $\mu_1$ (when $\lambda_1 = 4$, $\lambda_2 = 4$, $\mu_2 = 7$, $K_1 = 10$ and $K_2 = 3$). Work-Conserving

| | $\mathbb{E}[L_1]$ | $\mathbb{E}[L_2]$ | $\mathbb{P}(I=1)$ | $\mathbb{P}(I=2)$ | $P_{\text{loss}}(1)$ | $P_{\text{loss}}(2)$ | $\mathbb{E}[W_1]$ | $\mathbb{E}[W_2]$ | $SR$ |
|---|---|---|---|---|---|---|---|---|---|
| $\mu_1 = 0$ | 10 | 3 | 1 | 0 | 1 | 1 | $\infty$ | $\infty$ | 0 |
| $\mu_1 = 0.1$ | 9.9755 | 2.9348 | 0.9741 | 0.0259 | 0.9756 | 0.9547 | 102.409 | 16.1805 | 0.0976 |
| $\mu_1 = 0.5$ | 9.8847 | 2.7254 | 0.8911 | 0.1089 | 0.8886 | 0.8094 | 22.1859 | 3.5745 | 0.4085 |
| $\mu_1 = 5$ | 7.7191 | 1.8548 | 0.5941 | 0.4059 | 0.2612 | 0.2959 | 2.6119 | 0.6586 | 1.1080 |
| $\mu_1 = 50$ | 0.9603 | 0.8777 | 0.3536 | 0.6464 | 0.0011 | 0.0924 | 0.2404 | 0.2418 | 1.3796 |
| $\mu_1 = 500$ | 0.7955 | 0.8566 | 0.3476 | 0.6524 | 0.0006 | 0.0896 | 0.1990 | 0.2352 | 1.4022 |
| $\mu_1 = 5000000$ | 0.7797 | 0.8559 | 0.3474 | 0.6526 | 0.0006 | 0.0895 | 0.1950 | 0.2350 | 1.4027 |

Table 5.5: The impact of $\lambda_1$ (when $\lambda_2 = 4$, $\mu_1 = 7$, $\mu_2 = 7$, $K_1 = 10$ and $K_2 = 8$). Non-Work-Conserving

| | $\mathbb{E}[L_1]$ | $\mathbb{E}[L_2]$ | $\mathbb{P}(I=1)$ | $\mathbb{P}(I=2)$ | $P_{\text{loss}}(1)$ | $P_{\text{loss}}(2)$ | $\mathbb{E}[W_1]$ | $\mathbb{E}[W_2]$ | $SR$ |
|---|---|---|---|---|---|---|---|---|---|
| $\lambda_1 = 0$ | 0 | 1.2745 | 0 | 1 | 0 | 0.0049 | 0 | 0.3202 | 0 |
| $\lambda_1 = 0.1$ | 4.9287 | 1.4117 | 0.0196 | 0.9704 | 0.0017 | 0.0084 | 49.3692 | 0.3559 | 0.0119 |
| $\lambda_1 = 0.5$ | 4.9678 | 1.9274 | 0.0931 | 0.9069 | 0.0084 | 0.0214 | 10.0196 | 0.4924 | 0.0566 |
| $\lambda_1 = 5$ | 8.0553 | 6.0533 | 0.5363 | 0.4637 | 0.2716 | 0.2418 | 2.2117 | 1.9960 | 0.9348 |
| $\lambda_1 = 50$ | 9.8894 | 7.5042 | 0.7203 | 0.2797 | 0.8992 | 0.5124 | 1.9626 | 3.8475 | 1.9058 |
| $\lambda_1 = 500$ | 9.9897 | 7.5314 | 0.7323 | 0.2677 | 0.9897 | 0.5315 | 1.9488 | 4.0192 | 1.8733 |
| $\lambda_1 = 5000000$ | 10 | 7.5333 | 0.7333 | 0.2667 | 0.999999 | 0.5333 | 1.9480 | 4.0357 | 1.8667 |

Table 5.6: The impact of $\lambda_1$ (when $\lambda_2 = 4$, $\mu_1 = 7$, $\mu_2 = 7$, $K_1 = 10$ and $K_2 = 8$). Work-Conserving

| | $\mathbb{E}[L_1]$ | $\mathbb{E}[L_2]$ | $\mathbb{P}(I=1)$ | $\mathbb{P}(I=2)$ | $P_{\text{loss}}(1)$ | $P_{\text{loss}}(2)$ | $\mathbb{E}[W_1]$ | $\mathbb{E}[W_2]$ | $SR$ |
|---|---|---|---|---|---|---|---|---|---|
| $\lambda_1 = 0$ | 0 | 1.2745 | 0 | 1 | 0 | 0.0049 | 0 | 0.3202 | 0 |
| $\lambda_1 = 0.1$ | 0.0565 | 1.2932 | 0.0288 | 0.9712 | $8.9 \cdot 10^{-12}$ | 0.0050 | 0.5654 | 0.3249 | 0.0934 |
| $\lambda_1 = 0.5$ | 0.3109 | 1.3937 | 0.1229 | 0.8771 | $3.2 \cdot 10^{-6}$ | 0.0059 | 0.6218 | 0.3505 | 0.3672 |
| $\lambda_1 = 5$ | 7.5747 | 5.8061 | 0.5490 | 0.4510 | 0.2343 | 0.2139 | 1.9786 | 1.8465 | 0.9147 |
| $\lambda_1 = 50$ | 9.8891 | 7.5026 | 0.7210 | 0.2790 | 0.8910 | 0.5117 | 1.9595 | 3.8411 | 1.9051 |
| $\lambda_1 = 500$ | 9.9897 | 7.5314 | 0.7323 | 0.2677 | 0.9897 | 0.5315 | 1.9488 | 4.0192 | 1.8733 |
| $\lambda_1 = 5000000$ | 10 | 7.5333 | 0.7333 | 0.2667 | 0.999999 | 0.5333 | 1.9480 | 4.0357 | 1.8667 |

Table 5.7: The impact of $\mu_1$ (when $\lambda_1 = 4$, $\lambda_2 = 4$, $\mu_2 = 7$, $K_1 = 10$ and $K_2 = 8$). Non-Work-Conserving

| | $\mathbb{E}[L_1]$ | $\mathbb{E}[L_2]$ | $\mathbb{P}(I=1)$ | $\mathbb{P}(I=2)$ | $P_{\text{loss}}(1)$ | $P_{\text{loss}}(2)$ | $\mathbb{E}[W_1]$ | $\mathbb{E}[W_2]$ | $SR$ |
|---|---|---|---|---|---|---|---|---|---|
| $\mu_1 = 0$ | 10 | 8 | 1 | 0 | 1 | 1 | $\infty$ | $\infty$ | 0 |
| $\mu_1 = 0.1$ | 9.9757 | 7.9010 | 0.9702 | 0.0298 | 0.9757 | 0.9482 | 102.8220 | 38.1134 | 0.0945 |
| $\mu_1 = 0.5$ | 9.8888 | 7.5920 | 0.8775 | 0.1225 | 0.8904 | 0.7870 | 22.5556 | 8.9111 | 0.3871 |
| $\mu_1 = 5$ | 8.5254 | 6.0910 | 0.5514 | 0.4486 | 0.3255 | 0.2618 | 3.1599 | 2.0628 | 0.8900 |
| $\mu_1 = 50$ | 3.1548 | 4.0234 | 0.3785 | 0.6215 | 0.0056 | 0.0768 | 0.7932 | 1.0895 | 0.2568 |
| $\mu_1 = 500$ | 2.8346 | 3.9967 | 0.3753 | 0.6247 | 0.0007 | 0.0759 | 0.7091 | 1.0812 | 0.2513 |
| $\mu_1 = 5000000$ | 2.8020 | 3.9945 | 0.3750 | 0.6250 | 0.0002 | 0.0758 | 0.7006 | 1.0805 | 0.2509 |

Table 5.8: The impact of $\mu_1$ (when $\lambda_1 = 4$, $\lambda_2 = 4$, $\mu_2 = 7$, $K_1 = 10$ and $K_2 = 8$). Work-Conserving

| | $\mathbb{E}[L_1]$ | $\mathbb{E}[L_2]$ | $\mathbb{P}(I=1)$ | $\mathbb{P}(I=2)$ | $P_{\text{loss}}(1)$ | $P_{\text{loss}}(2)$ | $\mathbb{E}[W_1]$ | $\mathbb{E}[W_2]$ | $SR$ |
|---|---|---|---|---|---|---|---|---|---|
| $\mu_1 = 0$ | 10 | 8 | 1 | 0 | 1 | 1 | $\infty$ | $\infty$ | 0 |
| $\mu_1 = 0.1$ | 9.9757 | 7.9019 | 0.9704 | 0.0296 | 0.9757 | 0.9483 | 102.7950 | 38.2017 | 0.0946 |
| $\mu_1 = 0.5$ | 9.8884 | 7.5955 | 0.8785 | 0.1215 | 0.8902 | 0.7873 | 22.5132 | 8.9272 | 0.3875 |
| $\mu_1 = 5$ | 8.1645 | 5.9921 | 0.5667 | 0.4333 | 0.2926 | 0.2432 | 2.8853 | 1.9794 | 0.8749 |
| $\mu_1 = 50$ | 1.3769 | 1.3715 | 0.3200 | 0.6800 | 0.0014 | 0.0066 | 0.3447 | 0.3452 | 1.2496 |
| $\mu_1 = 500$ | 1.1681 | 1.2806 | 0.3129 | 0.6871 | 0.0003 | 0.0050 | 0.2921 | 0.3218 | 1.2912 |
| $\mu_1 = 5000000$ | 1.1483 | 1.2745 | 0.3125 | 0.6875 | 0.0002 | 0.0049 | 0.2872 | 0.3202 | 1.2933 |

$\underline{\lambda_1 \to 0}$

In this case it is clear that, in both scenarios, $\mathbb{P}(L_1 = 0) = 1$; Hence, $\mathbb{P}(I = 1) = 0$ and $\mathbb{P}(I = 2) = 1$. Therefore, $Q_2$ operates as an $M(\lambda_2)/M(\mu_2)/1/K_2$ system for which $P_{\text{loss}}(2) = \frac{\rho_2^{K_2}(1-\rho_2)}{1-\rho_2^{K_2+1}}$, and $\mathbb{E}[L_2] = \frac{\rho_2}{1-\rho_2} - \frac{(K_2+1)\rho_2^{K_2+1}}{1-\rho_2^{K_2+1}}$, where $\rho_2 = \frac{\lambda_2}{\mu_2}$.

$\underline{\mu_1 \to 0}$

This case is also straightforward and identical in both scenarios: $\mathbb{P}(I = 1) = 1$ and $\mathbb{P}(I = 2) = 0$. Therefore, $\mathbb{P}(L_1 = K_1) = 1$ and $\mathbb{P}(L_2 = K_2) = 1$, $P_{\text{loss}}(1) = 1$, and $P_{\text{loss}}(2) = 1$.

$\underline{\lambda_1 \to \infty}$

In both scenarios, whenever $\lambda_1 \to \infty$, $Q_1$ is always at it's maximum capacity, meaning $L_1 \equiv K_1$ and $P_{\text{loss}}(1) = 1$. In such a case the server serves the customers of $Q_1$ until the number of customers in $Q_2$ reaches it's maximum value, $K_2$. Then, once the server completes the service of a customer at $Q_1$, it immediately switches to $Q_2$. Before service completion there, an arrival to $Q_1$ will occur, causing a switch back to $Q_1$, once service at $Q_2$ ends. Hence, the only possible states with nonzero probabilities are $(K_1, K_2, 1)$, $(K_1, K_2 - 1, 1)$, and $(K_1, K_2, 2)$. Therefore, $\mathbb{P}(I = 1) = P_{K_1K_2}(1) + P_{K_1,K_2-1}(1)$, $\mathbb{P}(I = 2) = P_{K_1K_2}(2)$, and $P_{\text{loss}}(2) = P_{K_1K_2}(1) + P_{K_1K_2}(2)$. The sets of balance equations are reduced to

$$\begin{aligned}
\lambda_2 P_{K_1,K_2-1}(1) &= \mu_2 P_{K_1K_2}(2) \\
\mu_1 P_{K_1K_2}(1) &= \lambda_2 P_{K_1,K_2-1}(1) \\
\mu_2 P_{K_1K_2}(2) &= \mu_1 P_{K_1K_2}(1).
\end{aligned} \tag{5.1}$$

Solving (5.1) we arrive at

$$P_{K_1,K_2-1}(1) = \frac{\mu_1\mu_2}{\lambda_2\mu_2 + \mu_1\lambda_2 + \mu_1\mu_2}, \quad P_{K_1K_2}(1) = \frac{\lambda_2\mu_2}{\lambda_2\mu_2 + \mu_1\lambda_2 + \mu_1\mu_2}, \quad P_{K_1K_2}(2) = \frac{\mu_1\lambda_2}{\lambda_2\mu_2 + \mu_1\lambda_2 + \mu_1\mu_2}.$$

Hence, $\mathbb{P}(I = 1) = \frac{\mu_1\mu_2+\lambda_2\mu_2}{\lambda_2\mu_2+\mu_1\lambda_2+\mu_1\mu_2}$, $\mathbb{P}(I = 2) = \frac{\mu_1\lambda_2}{\lambda_2\mu_2+\mu_1\lambda_2+\mu_1\mu_2}$, $\mathbb{E}[L_2] = K_2 - \frac{\mu_1\mu_2}{\lambda_2\mu_2+\mu_1\lambda_2+\mu_1\mu_2}$, and $P_{\text{loss}}(2) = \frac{\lambda_2\mu_2+\lambda_2\mu_1}{\lambda_2\mu_2+\mu_1\lambda_2+\mu_1\mu_2}$. Note that the capacity of $Q_1$, $K_1$, does not affect the results.

The next extreme case leads to a different system structure in each of the switching scenarios.

$\underline{\mu_1 \to \infty}$

In Scenario 1, if $\mu_1 \to \infty$ then, whenever the server is at $Q_1$, it immediately reduces the number of customers there to 0, and will remain at $Q_1$ until the first moment when the number of customers in $Q_2$ reaches the threshold $K_2$. The server stays in the latter queue until the number of customers in $Q_1$ reaches the value $K_1$. Then, when $Q_2$ reduces below its threshold, $K_2$, the server will switch to $Q_1$ and immediately reduce the occupancy there to 0. Hence, $\mathbb{P}(I = 1) = P_{0\bullet}(1) = \sum_{n=0}^{K_2-1} P_{0n}(1)$.

In Scenario 2, if $\mu_1 \to \infty$ the server empties $Q_1$ instantaneously and switches to $Q_2$, given that the latter is not empty. Therefore, $\mathbb{P}(I = 1) = P_{00}(1)$. The server stays in $Q_2$ until $Q_1$ reaches its threshold and $Q_2$ is below the threshold $K_2$. Still, in both switching scenarios the proportion of time the server resides in $Q_1$ is not 0 (see Tables 5.3, 5.4, 5.7 and 5.8), and in both scenarios $P_{\text{loss}}(1) = P_{K_1K_2}(2)$, and $P_{\text{loss}}(2) = P_{\bullet K_2}(2)$.

The two-queue process is subject to an oscillation phenomenon: when the occupancy of $Q_i$ decreases, the occupancy of the other queue increases. This feature is illustrated in Figure 5.1, were we simulate a rather typical case presented in Table 5.7 with $\mu_1 = 50$. The periodicity pattern revealed in Figure 5.1 is further discussed in Section 6.

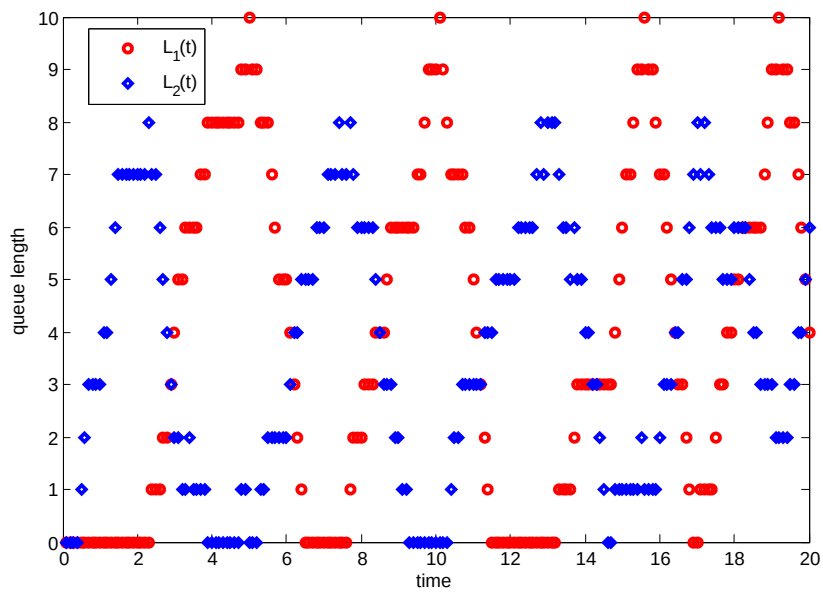


Figure 5.1: The case $\mu_1 = 50$ from Table 5.7.

# 6 Oscillations and Comparison Between the Scenarios

## 6.1 Oscillations

In Figures 6.1 and 6.2, we plot simulation results for the set of parameters of Figure 5.1, but with larger threshold values. It is seen that with increasing threshold levels, the system exhibits periodicity. We

conjecture that the two-queue process converges to a periodic process as the threshold values increase. There is another important observation in the case of large thresholds. When the thresholds are large, there is practically no difference between the cases $K_i = C_i$ and $K_i < C_i$. Yet, the limiting case $C_i \to \infty$, resulting in an unbounded two-dimensional process, is further being investigated by the authors.

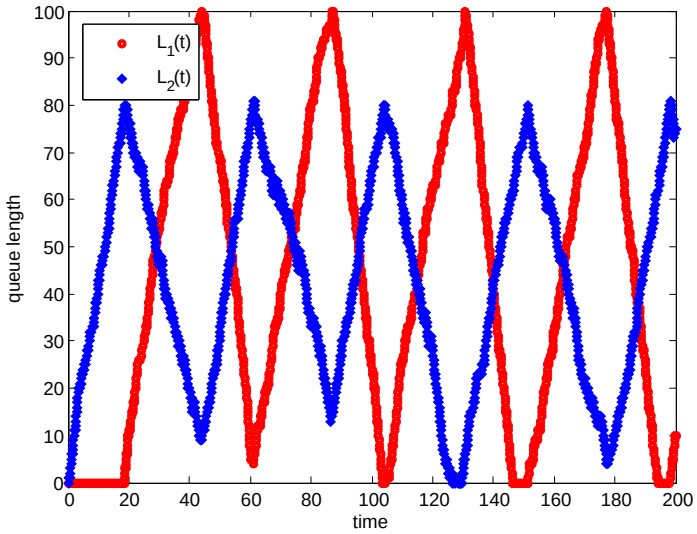


Figure 6.1: The case $\mu_1 = 50$ from Table 5.7 with $K_1 = 100$ and $K_2 = 80$.

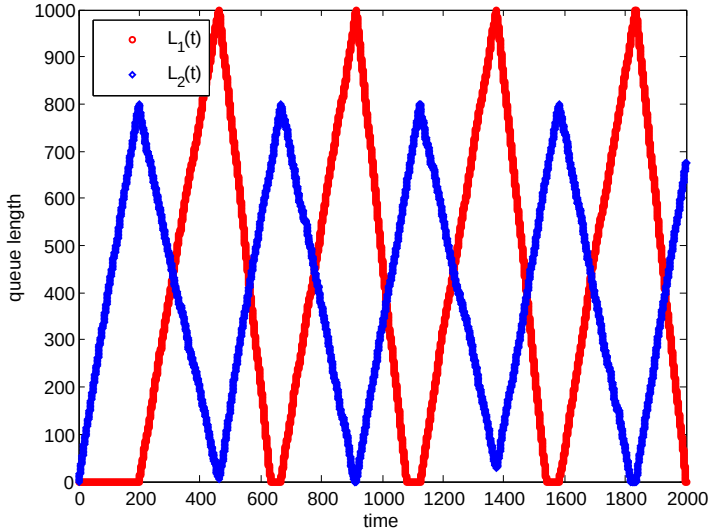


Figure 6.2: The case $\mu_1 = 50$ from Table 5.7 with $K_1 = 1000$ and $K_2 = 800$.

## 6.2 Comparison Between the Scenarios

Assume that every time unit a customer resides in the system (whether in $Q_1$ or in $Q_2$) costs $C_{RES}$ units of money, and that every server switch-over from $Q_i$ to $Q_j$, $i \neq j$, costs $C_{SR}$ units of money. Then, we define the mean cost per unit of time

$$\mathbb{E}[C] \equiv C_{RES}\left(\mathbb{E}[L_1] + \mathbb{E}[L_2]\right) + C_{SR} \cdot SR. \tag{6.1}$$

We now exhibit cases where Scenario 1 is economically preferable than Scenario 2. Figures 6.3 and 6.4 depict two graphs presenting the change in $\mathbb{E}[C]$ as a function of $\lambda_1$ and as a function of $\mu_1$, respectively, for both scenarios. In Figure 6.3 all other parameters assume the following values: $K_1 = 4$, $K_2 = 3$, $C_{RES} = 1$, $C_{SR} = 5$, $\lambda_2 = 4$, $\mu_1 = 7$, $\mu_2 = 7$, while in Figure 6.4 all other parameters assume the following values: $K_1 = 4$, $K_2 = 3$, $C_{RES} = 1$, $C_{SR} = 5$, $\lambda_1 = 4$, $\lambda_2 = 4$, $\mu_2 = 7$. It is seen that, for the above set of parameters, the non-work-conserving switching policy is more economical than the work-conserving switching policy.

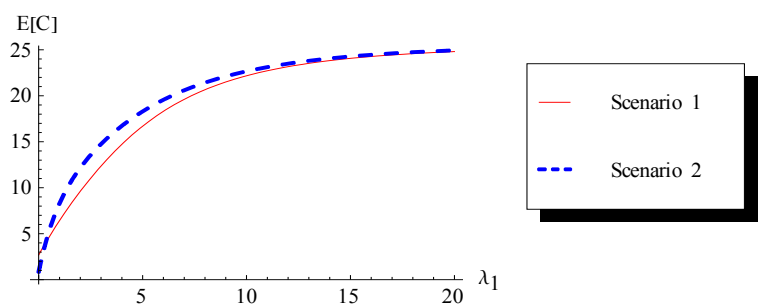


Figure 6.3: $\mathbb{E}[C]$ as a function of $\lambda_1$ for both scenarios.

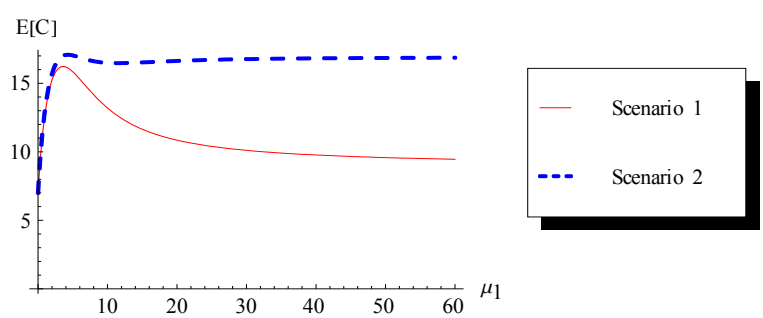


Figure 6.4: $\mathbb{E}[C]$ as a function of $\mu_1$ for both scenarios.

## Discussion

1. In all numerical calculations presented in Sections 4 and 5, the average switching rate between the queues (the parameter $SR$) is *always* smaller in the Non-Work-Conserving scenario than in the Work-

Conserving scenario, and when $\mu_1 \to \infty$ this phenomenon becomes highly *significant*. See Tables 5.3 and 5.7, as opposed to Tables 5.4 and 5.8, respectively.

2. When $\lambda_i$ is larger than $\mu_i$ the performance measures of both scenarios coincide, independently of all other parameters. See the 6-th row in Tables 4.1, 4.2, 4.3, and 4.4. This result is also exhibited in Tables 5.1, 5.2, 5.5, and 5.6, for $\lambda_1 \geq 50$.

3. In most of the presented calculations the performance measures related to the Work-Conserving scenario are better than those of the Non-Work-Conserving scenario, i.e., shorter queue lengths, smaller waiting times, and smaller loss probabilities. However, when high switching costs are involved (see Figures 6.3 and 6.4) the Non-Work-Conserving scenario is more economical than the Work-Conserving scenario.

4. When the service rates are sufficiently greater than the input rates, the system tends to a periodic behavior when the threshold levels are large.

# 7 Conclusions

In this paper we studied a two-queue finite-buffer polling-type system with a threshold switching policy. In contrast to other threshold-policy studies, the server determines its switching instants according to the size of the queue which is *not being served*. Employing both PGFs and matrix analytic approach we derived the joint and marginal steady state probabilities of the system's state. The solution of the PGFs was obtained by solving two finite sets of linear systems of the form $A(z)\vec{G}(z) = \vec{P}(z)$ and $B(w)\vec{F}(w) = \vec{\Pi}(w)$, respectively, where $\vec{G}(z)$ and $\vec{F}(w)$ are each a vector whose entries are the sought-for PGFs. $A(z)$ and $B(w)$ are finite square matrices with entries constructed from the parameters of the system. $\vec{P}(z)$ and $\vec{\Pi}(w)$ are finite-dimensional vectors consisting of unknown boundary probabilities. We constructed a procedure to calculate the boundary probabilities determining the PGFs by deriving explicit closed-form combinatorial expressions for the determinants of $A(z)$ and $B(w)$. We are not aware of any previous relevant study that obtained such explicit expressions. Using the matrix analytic approach revealed that the generator matrix defining the process differs from the classical generators of QBD processes, and therefore the analysis required the calculation of certain matrices defined by combinatorial expressions resembling the combinatorial expressions derived for the determinants of $A(z)$ and $B(w)$, appearing in the analysis via PGFs. In addition, we considered two switching scenarios: work-conserving and non-work-conserving. For each scenario we calculated the mean number of customers present in each queue and the mean sojourn time; the proportion of time the server spends in each queue; the proportion of lost customers; and the server's average switching rate between the queues. A comparison between the two scenarios was presented, and it was shown that the non-work-conserving scenario may be economically better when high switching costs are involved. We also observed that the dynamics of the system converges to a periodic behavior when the threshold levels are large. A formal justification of this observed phenomenon is possibly a future research direction. Other

future research directions are the limiting cases when the buffer capacities tend to infinity and the presence of non-zero switching times. It is expected that non-zero switching times can have similar effect as switching costs and provide one more raison d'être for the non-work-conserving policy.

**Acknowledgement** We thank the referees for their constructive comments that improved the presentation of the paper and enriched the discussion regarding the roots of the polynomials $|A(z)| = 0$ and $|B(w)| = 0$.

# References

Arazi A, Ben-Jacob E, Yechiali U (2005) Controlling an oscillating Jackson-type network having state-dependant service rates. Mathematical Methods of Operations Research 62: 453–466.

Avram F, Gómez-Corral A (2006) On the optimal control of a two-queue polling model. Operations Research Letters 34: 339–348.

Boon MAA, van der Mei RD, Winands EMM (2011) Applications of polling systems. Surveys in Operations Research and Management Science 16 (2): 67–82.

Boxma OJ, Down DG (1997) Dynamic server assignment in a two-queue model. European Journal of Operational Research 103: 595–609.

Boxma OJ, Koole GM, Mitrani I (1995a) A two-queue polling model with a threshold service policy. In Dowd PW and Gelenbe E, eds., Proceedings MASCOTS '95, IEEE Computer Society Press, Los Alamitos, CA, 84–89.

Boxma OJ, Koole GM, Mitrani I (1995b) Polling models with threshold switching. In Baccelli F, Jean-Marie A, and Mitrani I, eds., Quantitative Methods in Parallel Systems, Springer Verlag, Berlin, 129–140.

Bright LW, Taylor PG (1995) Calculating the equilibrium distribution in level dependent quasi-birth-and-death processes. Stochastic Models 11 (3): 497–526.

Coffman Jr. EG, Puhalskii AA and Reiman MI (1995) Polling systems with zero switchover times: A heavy-traffic averaging principle. Annals of Applied Probability, 5(3): 681–719.

Da Fonseca CM (2006) On the location of the eigenvalues of Jacobi matrices. Applied Mathematics Letters 19 (11) 11681174.

De Nitto Personè V, Grassi V (1996) Solution of finite QBD processes. Journal of Applied Probability 33 (4): 1003–1010.

Haverkort B, Idzenga HP, Kim BG (1994) Performance evaluation of threshold-based ATM cell scheduling policies under Markov modulated Poisson traffic using stochastic Petri nets. Modelling and Evaluation of

ATM Networks, volume 17 of IFIP Conference Proceedings, 553–572. Chapman & Hall.

Latouche G, Ramaswami V (1999) Introduction to Matrix Analytic Methods in Stochastic Modeling. SIAM and ASA, Philadelphia.

Lee D-S (1996) A two-queue model with exhaustive and limited service disciplines. Communications in Statistics. Stochastic Models 12 (2): 285–305.

Lee D-S, Sengupta B (1993) Queueing analysis of a threshold based priority scheme for ATM networks. IEEE/ACM Transactions on Networking 1 (6): 709–717.

Litvak N, Yechiali U (2003) Routing in queues with delayed information. Queueing Systems 43: 147–165.

Neuts MF (1981) Matrix Geometric Solutions in Stochastic Models - an Algorithmic Approach. The Johns Hopkins University Press, Baltimore and London.

Perel E, Yechiali U (2008) Queues where customers of one queue act as servers of the other queue. Qeueueing Systems 60: 271–288.

Perel E, Yechiali U (2013) On customers acting as servers. Asia-Pacific Journal of Operational Research 30 (5), DOI: 10.1142/S021759591350019X.

Perel N, Yechiali U (2013) The Israeli Queue with priorities. Stochastic Models 29: 353–379.

Takagi H (1986) Analysis of Polling Systems. The MIT Press.

Usmani RA, (1994) Inversion of a tridiagonal Jacobi matrix. Linear Algebra and its Applications 212/213: 413–414.

Yechiali U (1993) Analysis and control of polling systems. In Donatiello L and Nelson R, eds, Performance Evaluation of Computer and Communication Systems, Joint Tutorial Papers of Performance '93 and Sigmetrics '93, volume 729, 630–650, London, UK. Springer Berlin.

# A Appendix

## Proof of Theorem 2.3

*Proof.* By induction over $k$.
For $k = 1$,

$$q_1(z) = \alpha_0(z) = (-\lambda_1\mu_1)^0 \left( \binom{0}{0} \alpha^0(z)\alpha_0(z) + \binom{0}{-1} \alpha^1(z) \right).$$

For $k = 2$,

$$q_2(z) = \alpha(z)\alpha_0(z) - \lambda_1\mu_1 = \sum_{l=0}^{1} (-\lambda_1\mu_1)^l \left( \binom{1-l}{l} \alpha^{1-2l}(z)\alpha_0(z) + \binom{1-l}{l-1} \alpha^{2-2l}(z) \right).$$

We now show that the proposition is valid for any $k$. Suppose $k = 2i$ (the case where $k = 2i+1$ is similar and hence omitted from the presentation), and notice that for all $k \geq 0$, $\binom{k}{0} = \binom{k}{k} = 1$, and $\binom{k}{l} = \binom{k-1}{l} + \binom{k-1}{l-1}$, for every $0 \leq l \leq k$,

$$
\begin{aligned}
q_k(z) =& \alpha(z) q_{k-1}(z) - \lambda_1 \mu_1 q_{k-2}(z) \\
=& \alpha(z) \sum_{l=0}^{i-1} (-\lambda_1\mu_1)^l \left( \binom{2i-1-l-1}{l} \alpha^{2i-1-2l-1}(z)\alpha_0(z) + \binom{2i-1-l-1}{l-1} \alpha^{2i-1-2l}(z) \right) \\
& - \lambda_1\mu_1 \sum_{l=0}^{i-1} (-\lambda_1\mu_1)^l \left( \binom{2i-2-l-1}{l} \alpha^{2i-2-2l-1}(z)\alpha_0(z) + \binom{2i-2-l-1}{l-1} \alpha^{2i-2-2l}(z) \right) \\
=& \sum_{l=0}^{i-1} (-\lambda_1\mu_1)^l \left( \binom{2i-1-l-1}{l} \alpha^{2i-1-2l}(z)\alpha_0(z) + \binom{2i-1-l-1}{l-1} \alpha^{2i-2l}(z) \right) \\
& + \sum_{l=0}^{i-1} (-\lambda_1\mu_1)^{l+1} \left( \binom{2i-2-l-1}{l} \alpha^{2i-2-2l-1}(z)\alpha_0(z) + \binom{2i-2-l-1}{l-1} \alpha^{2i-2-2l}(z) \right) \\
=& \sum_{l=0}^{i-1} (-\lambda_1\mu_1)^l \left( \binom{2i-1-l-1}{l} \alpha^{2i-1-2l}(z)\alpha_0(z) + \binom{2i-1-l-1}{l-1} \alpha^{2i-2l}(z) \right) \\
& + \sum_{l=1}^{i} (-\lambda_1\mu_1)^l \left( \binom{2i-2-l}{l-1} \alpha^{2i-2l-1}(z)\alpha_0(z) + \binom{2i-2-l}{l-2} \alpha^{2i-2l}(z) \right) \\
=& (-\lambda_1\mu_1)^0 \left( \binom{2i-2}{0} \alpha^{2i-1}(z)\alpha_0(z) + \binom{2i-2}{-1} \alpha^{2i}(z) \right) \\
& + \sum_{l=1}^{i-1} (-\lambda_1\mu_1)^l \left( \binom{2i-2-l}{l} + \binom{2i-2-l}{l-1} \right) \alpha^{2i-2l-1}(z)\alpha_0(z) \\
& + \sum_{l=1}^{i-1} (-\lambda_1\mu_1)^l \left( \binom{2i-2-l}{l-1} + \binom{2i-2-l}{l-2} \right) \alpha^{2i-2l}(z) \\
& + (-\lambda_1\mu_1)^i \left( \binom{2i-2-i}{i-1} \alpha^{2i-2i-1}(z)\alpha_0(z) + \binom{2i-2-i}{i-2} \alpha^{2i-2i}(z) \right) \\
=& \sum_{l=0}^{i} (-\lambda_1\mu_1)^l \left( \binom{2i-l-1}{l} \alpha^{2i-2l-1}(z)\alpha_0(z) + \binom{2i-l-1}{l-1} \alpha^{2i-2l}(z) \right) \\
=& \sum_{l=0}^{\lfloor \frac{k}{2} \rfloor} (-\lambda_1\mu_1)^l \left( \binom{k-l-1}{l} \alpha^{k-2l-1}(z)\alpha_0(z) + \binom{k-l-1}{l-1} \alpha^{k-2l}(z) \right).
\end{aligned}
$$

This completes the proof. □

## Proof of Theorem 2.4

*Proof.* We will proceed by induction over $k$. First we note that $A_2 = diag\,(\mu_1)$, so that, $A_2^{-1} = diag\left(\frac{1}{\mu_1}\right)$. In addition, $A_0 = diag\,(\lambda_1) = \lambda_1 I_{K_2}$. Now, from (2.24) we have

$$\vec{P}_1^1 = -\frac{1}{\mu_1} \vec{P}_0^1 A_1^0 = \vec{P}_0^1 C_1.$$

Suppose that the proposition holds for all values up to some $k-1$, where $1 \leq k-1 \leq K_1 - 2$. We will show that it holds for $k \leq K_1 - 1$. From (2.24) we have

$$\vec{P}_k^1 = -\left(\vec{P}_{k-2}^1 A_0 + \vec{P}_{k-1}^1 A_1\right) A_2^{-1}.$$

Using the induction assumption with regard to the values of $\vec{P}_{k-2}^1$ and $\vec{P}_{k-1}^1$ we get

$$\begin{aligned}\vec{P}_k^1 &= -\left(\vec{P}_0^1 C_{k-2} A_0 + \vec{P}_0^1 C_{k-1} A_1\right) A_2^{-1} \\ &= -\vec{P}_0^1 \left(C_{k-2} A_0 + C_{k-1} A_1\right) A_2^{-1} \\ &= -\frac{1}{\mu_1} \vec{P}_0^1 \left(\lambda_1 C_{k-2} + C_{k-1} A_1\right).\end{aligned}$$

Therefore $\vec{P}_k^1 = \vec{P}_0^1 C_k$, where $C_k = -\frac{1}{\mu_1}\left(\lambda_1 C_{k-2} + C_{k-1} A_1\right)$.
This completes the proof. □

## Proof of Theorem 2.6

*Proof.* Similarly to the proof of Theorem 2.3, we proceed by induction over $k$.
For $k = 1$,

$$C_1 = \frac{-1}{\mu_1} A_1^0 = \left(\frac{-1}{\mu_1}\right)^1 (-\lambda_1 \mu_1)^0 \left(\begin{pmatrix}0\\0\end{pmatrix} A_1^0 (A_1)^0 + \begin{pmatrix}0\\-1\end{pmatrix} (A_1)^1\right).$$

For $k = 2$,

$$C_2 = \frac{-1}{\mu_1}\left(\lambda_1 I_{K_2} + \left(\frac{-1}{\mu_1}\right) A_1^0 A_1\right) = \left(\frac{-1}{\mu_1}\right)^2 \sum_{l=0}^{1} (-\lambda_1 \mu_1)^l \left(\begin{pmatrix}1-l\\l\end{pmatrix} A_1^0 (A_1)^{1-2l} + \begin{pmatrix}1-l\\l-1\end{pmatrix} (A_1)^{2-2l}\right).$$

We now prove that the proposition holds for any $k$. Suppose $k = 2i+1$ (the case for even values, $k = 2i$, is presented in the proof of Theorem 2.3),

$$\begin{aligned}C_k =& \frac{-1}{\mu_1}\left(C_{k-1} A_1 + \lambda_1 C_{k-2}\right) \\ =& \frac{-1}{\mu_1}\Bigg(\left(\frac{-1}{\mu_1}\right)^{2i} \sum_{l=0}^{i} (-\lambda_1 \mu_1)^l \left(\begin{pmatrix}2i-1-l\\l\end{pmatrix} A_1^0 (A_1)^{2i-2l-1} + \begin{pmatrix}2i-1-l\\l-1\end{pmatrix} (A_1)^{2i-2l}\right) A_1 \\ &+ \lambda_1 \left(\frac{-1}{\mu_1}\right)^{2i-1} \sum_{l=0}^{i-1} (-\lambda_1 \mu_1)^l \left(\begin{pmatrix}2i-1-1-l\\l\end{pmatrix} A_1^0 (A_1)^{2i-1-2l-1} + \begin{pmatrix}2i-1-1-l\\l-1\end{pmatrix} (A_1)^{2i-1-2l}\right)\Bigg) \\ =& \left(\frac{-1}{\mu_1}\right)^{2i+1} \Bigg(\sum_{l=0}^{i} (-\lambda_1 \mu_1)^l \left(\begin{pmatrix}2i-1-l\\l\end{pmatrix} A_1^0 (A_1)^{2i-2l} + \begin{pmatrix}2i-1-l\\l-1\end{pmatrix} (A_1)^{2i+1-2l}\right) \\ &+ \sum_{l=0}^{i-1} (-\lambda_1 \mu_1)^{l+1} \left(\begin{pmatrix}2i-2-l\\l\end{pmatrix} A_1^0 (A_1)^{2i-2-2l} + \begin{pmatrix}2i-2-l\\l-1\end{pmatrix} (A_1)^{2i-1-2l}\right)\Bigg) \\ =& \left(\frac{-1}{\mu_1}\right)^{2i+1} \Bigg(\sum_{l=0}^{k} (-\lambda_1 \mu_1)^l \left(\begin{pmatrix}2i-1-l\\l\end{pmatrix} A_1^0 (A_1)^{2i-2l} + \begin{pmatrix}2i-1-l\\l-1\end{pmatrix} (A_1)^{2i+1-2l}\right) \\ &+ \sum_{l=1}^{i} (-\lambda_1 \mu_1)^l \left(\begin{pmatrix}2i-1-l\\l-1\end{pmatrix} A_1^0 (A_1)^{2i-2l} + \begin{pmatrix}2i-1-l\\l-2\end{pmatrix} (A_1)^{2i+1-2l}\right)\Bigg)\end{aligned}$$

$$
\begin{aligned}
&= \left(\frac{-1}{\mu_1}\right)^{2i+1} \Bigg( (-\lambda_1\mu_1)^0 \left( \binom{2i-1}{0} A_1^0 \left(A_1\right)^{2i} + \binom{2i-1}{-1} \left(A_1\right)^{2i+1} \right) \\
&\quad + \sum_{l=1}^{i} (-\lambda_1\mu_1)^l \left( \binom{2i-1-l}{l} + \binom{2i-1-l}{l-1} \right) A_1^0 \left(A_1\right)^{2i-2l} \\
&\quad + \sum_{l=1}^{i} (-\lambda_1\mu_1)^l \left( \binom{2i-1-l}{l-1} + \binom{2i-1-l}{l-2} \right) \left(A_1\right)^{2i+1-2l} \Bigg) \\
&= \left(\frac{-1}{\mu_1}\right)^{2i+1} \sum_{l=0}^{i} (-\lambda_1\mu_1)^l \left( \binom{2i-l}{l} A_1^0 \left(A_1\right)^{2i-2l} + \binom{2i-l}{l-1} \left(A_1\right)^{2i+1-2l} \right) \\
&= \left(\frac{-1}{\mu_1}\right)^{k} \sum_{l=0}^{\lfloor \frac{k}{2} \rfloor} (-\lambda_1\mu_1)^l \left( \binom{k-l-1}{l} A_1^0 \left(A_1\right)^{k-2l-1} + \binom{k-l-1}{l-1} \left(A_1\right)^{k-2l} \right).
\end{aligned}
$$

This completes the proof. □