Two-Queue Polling Systems with Switching Policy Based on the Queue which is Not Being Served

Efrat Perel[1] and Uri Yechiali[2]

[1] 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.
[efratp@afeka.ac.il] [uriy@post.tau.ac.il]

Abstract

We study a system of two non-identical and separate $M/M/1/\bullet$ queues with capacities (buffers) $C_1 < \infty$ and $C_2 = \infty$, respectively, served by a single server that alternates between the queues. The server's switching policy is threshold-based, and, *in contrast to other threshold models*, is determined by the state of the queue that is *not* being served. That is, when neither queue is empty while the server attends Q_i $(i = 1, 2)$, the server switches to the other queue as soon as the latter reaches its threshold. When a served queue becomes empty we consider two switching scenarios: (i) Work-Conserving, and (ii) Non-Work-Conserving. We analyze the two scenarios using Matrix Geometric methods and obtain explicitly the rate matrix R, where its entries are given in terms of the roots of the determinants of two underlying matrices. Numerical examples are presented and extreme cases are investigated.

Keywords Alternating server · Threshold policy · Polling systems · Matrix Geometric · PGFs

Mathematics Subject Classification (2000) Primary: 60K25

1 Introduction

We study two-queue polling-type systems governed by a threshold-based switching policy where, in contrast to many other works in the literature, the server's switching decisions are determined by the queue that is *not* being served. Specifically, whenever 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 the other queue, Q_j, $j \neq i$ 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 than or equal to Q_i's own threshold level. In the latter situation the server remains in Q_i until the number of customers there is reduced below Q_i's threshold level, and only then does it switch to Q_j. When a served Q_i becomes empty, we consider two switching scenarios: (i) Work-Conserving: If Q_j is not empty, the server switches immediately; otherwise, it remains idle until either one of the queues becomes non empty. (ii) Non-Work-Conserving: The server remains in Q_i (idle or busy) until the first moment when Q_j reaches its threshold level. For each Q_i we assume that the queue's capacity is C_i and that customers arrive according to a Poisson process with rate λ_i. The service time for each individual customer is exponentially distributed with mean $1/\mu_i$. All the arrival and service processes are independent. For Q_1 we let $C_1 < \infty$, while for Q_2 we set $C_2 = \infty$. We note that if both capacities C_1 and C_2 are infinite, the problem will be completely different and will require an entirely different approach than the current one. The threshold levels are $K \leq C_1$ for Q_1, and $N < C_2$ for Q_2. The system is depicted in Figure 1.1.

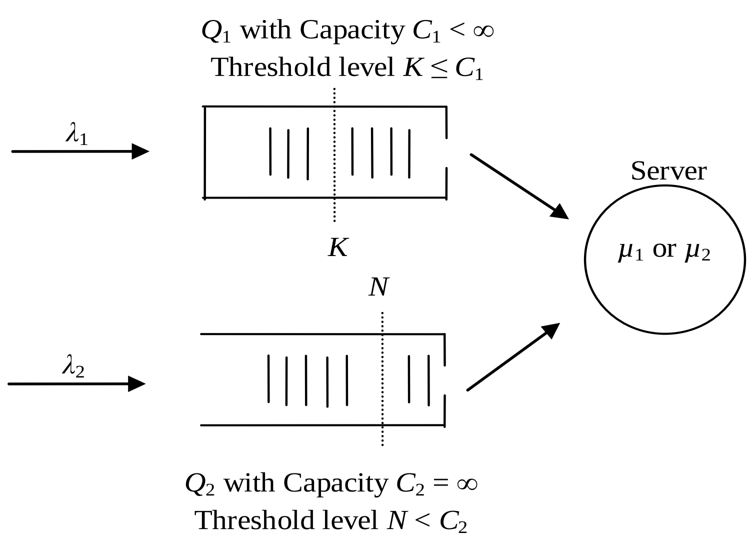

Figure 1.1: Two Queues Served by a Single Alternating Server with Threshold Policy.

A motivation for such a model is, for example, an automated traffic light (or a traffic policeman) 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. Another application arises in data centers, where a rack of discs requires special attention when the amount of recorded data exceeds a certain limit (threshold), causing an inefficient operation that calls for a clean-up action. A more abstract example refers to human beings, who often 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 do they switch their attention to that task.

Single-server polling systems, where the server visits the queues in a cyclic order, mostly under Exhaustive, Gated, Globally-Gated, or k-limited service regimes, have been studied extensively in the queueing literature (see e.g. Takagi [20], Boxma, Levy and Yechiali [5], Yechiali [21], Boon et al. [3], and many references therein). Threshold based polling systems have also been treated (see e.g. Lee [12], Lee and Sengupta [13], Haverkot et al. [10], Boxma et al. [6, 7], Avram and Gómez-Corral [2], Perel [16] and many others). In most of the above-mentioned studies, the switching policy is determined by the state of the queue that is *presently being served*. Recently, Avrachenkov et al. [1] studied a two-queue finite-buffers system with a threshold-based switching policy. Using algebraic methods, they investigated the effects of buffer sizes, arrival rates and service rates on the system's performance.

In this paper we concentrate on the derivation of the joint distribution function of the queue-size process for each of the two scenarios described above. To this end, we formulate each system as a quasi birth-and-death (QBD) process having a three-dimensional state space. We study the system's steady-state behavior by applying Matrix Geometric methods (see e.g., Neuts [14], Latouche and Ramaswami [11]) and obtain explicitly the rate matrix R. A detailed analysis of the Work-Conserving switching scenario is presented, while the Non-Work-Conserving scenario is only briefly discussed (since its analysis is very similar to that of the former). The two scenarios are compared numerically.

The structure of the paper is as follows: In Section 2 the mechanism of the Work-Conserving scenario is characterized. In Section 3 the system is defined as a QBD process and a Matrix Geometric approach is employed to derive the system's steady-state probabilities. Investigating the rate matrix R reveals that its elements are closely related to the roots of two polynomial equations,

$det\,(A(z)) = 0$, and $det\,(B(z)) = 0$, where $A(z)$ and $B(z)$ are two matrices related to the Probability Generating Functions (PGFs) of the phases of the QBD process. We show that the entries of the rate matrix R are explicitly calculated in terms of the roots of the determinants of the above two matrices. The theoretical relationship between the diagonal elements of R and the roots of the matrices $A(z)$ and $B(z)$ has not been analytically investigated, but has already been observed in other studies such as Paz and Yechiali [15], Perel N. and Yechiali [17] and Hanukov et al. [9]. In Section 4 the Non-Work-Conserving switching scenario is briefly treated, while in Section 5 numerical results are presented and the two scenarios are compared. The numerical results are followed by a discussion pointing out various phenomena occurring as a result of changes in parameters and queue capacities. Section 6 deals with extreme cases, while Section 7 concludes the paper.

2 Work-Conserving Scenario: Model Description

Consider a single-server two-queue polling-type system where the server's switching instants between the queues follow a threshold policy based on the queue *that is not being served*. Each queue i (Q_i), $i = 1, 2$, operates as an $M/M/1/C_i$ queue, with a Poisson arrival rate λ_i and exponentially distributed service time having mean $1/\mu_i$. The overall capacity of Q_1 is $1 \leq C_1 < \infty$ and of Q_2 is $C_2 = \infty$. That is, customers arriving at Q_1 and finding C_1 customers present there are blocked and balk from the system. When the server attends a non-empty Q_1 (Q_2, respectively), it continues serving customers there until the first moment thereafter when the number of customers in the other queue, Q_2 (Q_1), reaches its threshold level, N (K). At that instant the server immediately switches to Q_2 (Q_1) and continues serving there until the first moment thereafter when the queue size in Q_1 (Q_2) reaches K (N). At that moment the server switches back to Q_1 (Q_2), and so forth. Denoting by $L_i(t)$ the number of customers in Q_i at time t, then, if at a called-for switching moment from Q_1 (Q_2) to Q_2 (Q_1) the number of customers in Q_1 (Q_2) is still $L_1(t) \geq K$ ($L_2(t) \geq N$), the server remains in Q_1 (Q_2) until the first moment thereafter when $L_1(t)$ ($L_2(t)$) reduces to $K - 1$ ($N - 1$), and only then switches to Q_2 (Q_1). When the server empties Q_1 (Q_2) while $L_2(t) > 0$ ($L_1(t) > 0$), it immediately switches to the other queue. To keep the analysis less cumbersome we analyze the case where $K = C_1$ and $N < C_2$ (noting that the analysis of the case where $K < C_1$ is similar). Let $I(t) = 1$ if at time t the server attends Q_1, and $I(t) = 2$ if the server attends Q_2. The triple $(L_1(t), L_2(t), I(t))$ defines a non reducible continuous-time Markov chain with transition-rate

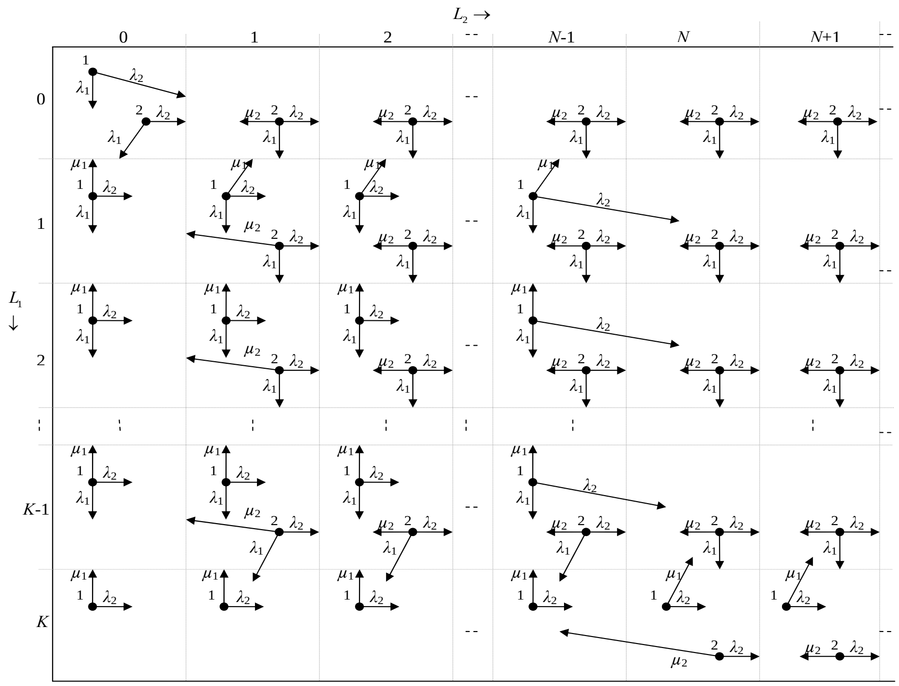

Figure 2.1: Transition-rate diagram of $(L_1(t), L_2(t), I(t))$. Work-Conserving.

diagram depicted in Figure 2.1 (the numbers 1 or 2 appearing next to each node indicate whether $I(t) = 1$, or $I(t) = 2$, respectively). Each box (k, n) depicts both the state where $I(t) = 1$ and the state where $I(t) = 2$. It will be shown that a necessary and sufficient condition for stability is $\lambda_2 < \mu_2$. In such a case, let $L_i = \lim_{t\to\infty} L_i(t)$ and $I = \lim_{t\to\infty} I(t)$. Consequently, for a system in steady state, let $P_{kn}(i) = \mathbb{P}(L_1 = k, L_2 = n, I = i)$, where $0 \le k \le K$; $0 \le n$; $i = 1, 2$.

3 The QBD Process

3.1 Matrix Geometric

The triple $(L_1(t), L_2(t), I(t))$ defines a quasi birth-and-death (QBD) process, where $L_2(t)$ denotes the level and the pair $(L_1(t), I(t))$ indicates the phase of the process. We order the resulting infinite-state space S as follows: We start with column $L_2 = 0$ and go down the boxes from $L_1 = 0$ to $L_1 = K$, where in each box we specify first the state associated with $I = 1$, and then the state associated with $I = 2$ (if any). We proceed similarly with columns $L_2 = 1, 2, 3, ..., N, N + 1, ...$. Thus, the state's space is $S = \{(0, 0, 1), (0, 0, 2), (1, 0, 1), (2, 0, 1), ..., (K, 0, 1)$ **;** $(0, 1, 2), (1, 1, 1), (1, 1, 2),$

$..., (K-1,1,1), (K-1,1,2), (K,1,1)$ **;** $...$ **;** $(0,N-1,2), ..., (K-1,N-1,1), (K-1,N-1,2),$ $(K,N-1,1)$ **;** $(0,N,2), (1,N,2), ..., (K-1,N,2), (K,N,1), (K,N,2)$ **;** $(0,N+1,2), (1,N+1,2), ...,$ $(K-1,N+1,2), (K,N+1,1), (K,N+1,2)$ **;** $...\}$.

The generator matrix Q is given by

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

where $\mathbf{0}$ is a matrix of zeros, and starting from the upper diagonal, B_0^0, B_0, B_0^{N-1}, A_0; B_1^0, B_1, A_1; B_2, B_2^1, A_2^N and A_2 are the following matrices: B_0^0 is of size $(K+2) \times 2K$, B_0 is of size $2K \times 2K$, B_0^{N-1} is of size $2K \times (K+2)$, A_0 is of size $(K+2) \times (K+2)$; B_1^0 is of size $(K+2) \times (K+2)$, B_1 is of size $2K \times 2K$, A_1 is of size $(K+2) \times (K+2)$; B_2 is of size $2K \times 2K$, B_2^1 is of size $2K \times (K+2)$, A_2^N is of size $(K+2) \times 2K$, and A_2 is of size $(K+2) \times (K+2)$. The above matrices are given by

$$B_0^0 = \begin{pmatrix} \lambda_2 & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & 0 \\ \lambda_2 & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & \lambda_2 & 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \lambda_2 & 0 & \cdots & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & \cdots & \cdots & \lambda_2 & 0 & 0 \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 & 0 & \lambda_2 \end{pmatrix},$$

$$B_0 = diag(\lambda_2),$$

$$B_0^{N-1} = \begin{pmatrix} \lambda_2 & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & \lambda_2 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & \lambda_2 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & \cdots & \cdots & \lambda_2 & 0 & 0 \\ 0 & \cdots & \cdots & \cdots & \cdots & \lambda_2 & 0 & 0 \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 & \lambda_2 & 0 \end{pmatrix},$$

$$A_0 = diag(\lambda_2).$$

With $\beta_0 = \lambda_1 + \lambda_2$; $\beta_1 = \lambda_1 + \lambda_2 + \mu_1$; and $\beta_2 = \lambda_1 + \lambda_2 + \mu_2$,

$$B_1^0 = \begin{pmatrix} -\beta_0 & 0 & \lambda_1 & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & -\beta_0 & \lambda_1 & 0 & 0 & \cdots & \cdots & 0 \\ \mu_1 & 0 & -\beta_1 & \lambda_1 & 0 & \ddots & \cdots & 0 \\ 0 & 0 & \mu_1 & -\beta_1 & \lambda_1 & 0 & 0 & \cdots \\ 0 & 0 & 0 & \mu_1 & -\beta_1 & \lambda_1 & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \mu_1 & -\beta_1 & \lambda_1 \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 & \mu_1 & -(\lambda_2 + \mu_1) \end{pmatrix},$$

$$B_1 = \begin{pmatrix} -\beta_2 & 0 & \lambda_1 & \cdots & \cdots & \cdots & \cdots & 0 \\ \mu_1 & -\beta_1 & 0 & \lambda_1 & 0 & \cdots & \cdots & 0 \\ 0 & 0 & -\beta_2 & 0 & \lambda_1 & \ddots & \cdots & 0 \\ 0 & \mu_1 & 0 & -\beta_1 & 0 & \lambda_1 & 0 & \cdots \\ 0 & 0 & 0 & 0 & -\beta_2 & 0 & \lambda_1 & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & 0 & -\beta_2 & \lambda_1 \\ 0 & \cdots & \cdots & \cdots & \cdots & \mu_1 & 0 & -(\lambda_2 + \mu_1) \end{pmatrix},$$

$$A_1 = \begin{pmatrix} -\beta_2 & \lambda_1 & 0 & \cdots & \cdots & 0 \\ 0 & -\beta_2 & \lambda_1 & 0 & \vdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & -\beta_2 & 0 & \lambda_1 \\ \vdots & \ddots & \ddots & \mu_1 & -(\lambda_2+\mu_1) & 0 \\ \vdots & \ddots & \ddots & \ddots & 0 & -(\lambda_2+\mu_2) \end{pmatrix},$$

$$B_2^1 = \begin{pmatrix} 0 & \mu_2 & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & 0 & 0 & \cdots & \cdots & \cdots & \vdots \\ \vdots & \ddots & \mu_2 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \mu_2 & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & 0 \end{pmatrix}, \quad B_2 = \begin{pmatrix} \mu_2 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & 0 & 0 & \cdots & \cdots & \cdots & \vdots \\ \vdots & \ddots & \mu_2 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \mu_2 & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & 0 \end{pmatrix},$$

$$A_2^N = \begin{pmatrix} \mu_2 & 0 & \cdots & \cdots & 0 \\ 0 & 0 & \mu_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & 0 \\ \vdots & \ddots & \ddots & 0 & \mu_2 \end{pmatrix}, \quad A_2 = \begin{pmatrix} \mu_2 & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & \mu_2 & 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mu_2 & \ddots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & 0 & \mu_2 \end{pmatrix}.$$

Define the steady-state probability vector $\vec{\mathcal{P}} = \left(\vec{P}_0, \vec{P}_1, ..., \vec{P}_N, ...\right)$, satisfying $\vec{\mathcal{P}}Q = \vec{0}$, $\vec{\mathcal{P}} \cdot \vec{e} = 1$, where $\vec{0}$ is a vector of 0's and $\vec{e}$ is a vector of 1's. Also, the probability vector

$$\vec{P}_n = \begin{cases} \left(P_{00}(1), P_{00}(2), P_{10}(1), P_{20}(1), ..., P_{K-1,0}(1), P_{Kn}(1)\right), & n = 0, \\ \left(P_{0n}(2), ..., P_{K-1,n}(1), P_{K-1,n}(2), P_{Kn}(1)\right), & 0 < n < N, \\ \left(P_{0n}(2), ..., P_{K-1,n}(2), P_{Kn}(1), P_{Kn}(2)\right), & n \geq N, \end{cases}$$

satisfies

$$\vec{P}_0 B_1^0 + \vec{P}_1 B_2^1 = \vec{0}, \tag{3.1}$$

$$\vec{P}_1 B_0^0 + \vec{P}_1 B_1 + \vec{P}_2 B_2 = \vec{0}, \tag{3.2}$$

$$\vec{P}_{n-1} B_0 + \vec{P}_n B_1 + \vec{P}_{n+1} B_2 = \vec{0}, \qquad 2 \leq n \leq N-2, \tag{3.3}$$

$$\vec{P}_{N-2} B_0 + \vec{P}_{N-1} B_1 + \vec{P}_N A_2^N = \vec{0}, \tag{3.4}$$

$$\vec{P}_{N-1} B_0^{N-1} + \vec{P}_N A_1 + \vec{P}_{N+1} A_2 = \vec{0}, \tag{3.5}$$

$$\vec{P}_{n-1} A_0 + \vec{P}_n A_1 + \vec{P}_{n+1} A_2 = \vec{0}, \qquad n \geq N+1. \tag{3.6}$$

Summing equations (3.1)-(3.6) and rearranging terms, we arrive at

$$\begin{aligned} &\mu_1 \left(P_{1\bullet}(1) - P_{10}(1) + P_{K\bullet}(1) - \sum_{n=0}^{N-1} P_{K,n}(1) \right) + \lambda_2 \left(P_{00}(1) + \sum_{k=1}^{K-1} P_{k,N-1}(1) \right) \\ &\qquad = \mu_2 \left(P_{\bullet 1}(2) - P_{01}(2) + P_{KN}(2) \right) + \lambda_1 \left(P_{00}(2) + \sum_{n=1}^{N-1} P_{K-1,n}(2) \right). \end{aligned} \tag{3.7}$$

Indeed, equation (3.7) states that the mean switching rate from state $I = 1$ to state $I = 2$ (left hand side of (3.7)) is equal to the mean switching rate from state $I = 2$ to state $I = 1$ (right hand side of (3.7)).

Let $A = A_0 + A_1 + A_2$. Then,

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

Let $\vec{\pi} = \left(\pi_0, \pi_1, ..., \pi_{K-1}, \pi_K^{(1)}, \pi_K^{(2)} \right)$ be the stationary probability vector of the matrix A, i.e. $\vec{\pi} A = \vec{0}$ and $\vec{\pi} \cdot \vec{e} = 1$. Then, $\vec{\pi} = (\underbrace{0, 0, \ldots, 0}_{K+1 \text{ times}}, 1)$. Thus, the stability condition $\vec{\pi} A_0 \vec{e} < \vec{\pi} A_2 \vec{e}$ (see [14]) becomes

$$\lambda_2 < \mu_2. \tag{3.8}$$

The probability vectors are given by

$$\vec{P}_n = \vec{P}_N R^{n-N}, \quad n \geq N, \tag{3.9}$$

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

$$A_0 + RA_1 + R^2 A_2 = \mathbf{0}. \tag{3.10}$$

The vectors $\vec{P}_0$, $\vec{P}_1$, ..., $\vec{P}_N$, can be found by solving the set of equations (3.1)-(3.4), together with the normalization equation,

$$\sum_{n=0}^{N-1} \vec{P}_n \vec{e} + \vec{P}_N [\mathcal{I} - R]^{-1} \vec{e} = 1,$$

where $\mathcal{I}$ is the identity matrix. We note that the above set of equations could be solved efficiently using the Level-Dependent QBD approach, see Bright and Taylor [8] and Phung-Duc et al. [19].
The mean total number of customers in Q_i, $\mathbb{E}[L_i]$, $i = 1, 2$ is given by

$$\mathbb{E}[L_1] = \vec{P}_0 \vec{Z}_0 + \sum_{n=1}^{N-1} \vec{P}_n \vec{Z} + \sum_{n=N}^{\infty} \vec{P}_n \vec{Z}_N, \tag{3.11}$$

$$\begin{aligned} \mathbb{E}[L_2] &= \sum_{n=1}^{\infty} n \vec{P}_n \vec{e} = \sum_{n=1}^{N-1} n \vec{P}_n \vec{e} + \sum_{n=N}^{\infty} n \vec{P}_N R^{n-N} \vec{e} \\ &= \sum_{n=1}^{N-1} n \vec{P}_n \vec{e} + (N-1) \vec{P}_N [\mathcal{I} - R]^{-1} \vec{e} + \vec{P}_N [\mathcal{I} - R]^{-2} \vec{e}, \end{aligned} \tag{3.12}$$

where, $\vec{Z}_0 = (0, 0, 1, 2, ..., K-1, K)$, $\vec{Z} = (0, 1, 1, 2, 2, ..., K-1, K-1, K)$ and $\vec{Z}_N = (0, 1, 2, ..., K-1, K, K)$.
We denote the elements of the matrix R by R_{lm}, for $0 \leq l, m \leq K+1$. By using equation (3.10) and explicitly writing the $(K+2)^2$ equations for the $(K+2)^2$ elements of R, we conclude that the matrix R is an upper triangular matrix, with only one non-zero element, $R_{K,K-1}$, beneath the main diagonal. Therefore, solving equation (3.10) yields an analytic closed-form expression for the elements of the rate matrix R. We will show that the elements of R are closely related to the roots of two polynomial equations, $det\,(A(z)) = 0$, and $det\,(B(z)) = 0$, where $A(z)$ and $B(z)$ are two matrices related to the probability generating functions (PGFs) defined in the following section.

3.2 Probability Generating Functions

In this section we briefly describe an alternative approach to solving the QBD process, namely, the PGF approach. It can be argued that given the analysis of Section 3.1, the PGF approach is redundant. Nevertheless, in our case, a brief investigation via the PGF method is useful for gaining further insights into the analysis of the system (see e.g. Phung-Duc [18]).

Splitting the set of equations (3.1)-(3.6) into two separate sets, one for $I = 1$, running from $k = 0$ to $k = K$; the other for $I = 2$, running over all $n \geq 0$, allows us to define two sets of probability generating functions: For $I = 1$, $G_k(z) = \sum_n P_{kn}(1)z^n$, $1 \leq k \leq K$; while for $I = 2$, $F_k(z) = \sum_n P_{kn}(2)z^n$, $0 \leq k \leq K$. After some algebra, one obtains two sets of linear equations, where the unknowns are the sought-for PGFs, as follows:

$$A(z)\vec{G}(z) = \vec{P}(z), \quad B(z)\vec{F}(z) = \vec{\Pi}(z), \tag{3.13}$$

where the column vectors $\vec{G}(z)$ and $\vec{P}(z)$ are of order K, while their counterparts, $\vec{F}(z)$ and $\vec{\Pi}(z)$, are of order $K+1$. The square matrices $A(z)$ and $B(z)$ are of orders K and $K+1$, respectively. Specifically,

$$\begin{aligned}
\vec{G}(z) &= \left(G_1(z), G_2(z), ..., G_K(z)\right)^t, \\
\vec{F}(z) &= \left(F_0(z), F_1(z), ..., F_K(z)\right)^t, \\
\vec{P}(z) &= \left(P_1(z), P_2(z), ..., P_K(z)\right)^t, \\
\vec{\Pi}(z) &= \left(\Pi_0(z), \Pi_1(z), ..., \Pi_K(z)\right)^t,
\end{aligned}$$

with

$$P_k(z) = \begin{cases} \lambda_1\left(P_{00}(1) + P_{00}(2)\right) - \lambda_2 P_{1,N-1}(1)z^N + \mu_2 P_{11}(2), & k = 1 \\ -\lambda_2 P_{k,N-1}(1)z^N + \mu_2 P_{k1}(2), & 2 \leq k \leq K-2 \\ \mu_1 \sum\limits_{n=0}^{N-1} P_{Kn}(1)z^n - \lambda_2 P_{K-1,N-1}(1)z^N + \mu_2 P_{K-1,1}(2), & k = K-1 \\ \lambda_1 \sum\limits_{n=1}^{N-1} P_{K-1,n}(2)z^n + \mu_2 P_{KN}(2)z^{N-1}, & k = K \end{cases}$$

$$\Pi_k(z) = \begin{cases} \mu_1 z G_1(z) - \mu_1 z P_{10}(1) + \lambda_2 z^2 P_{00}(1) - \mu_2(1-z)P_{00}(2), & k = 0 \\ -\lambda_1 z P_{00}(2) + \lambda_2 P_{1,N-1}(1)z^{N+1} - \mu_2 z P_{11}(2), & k = 1 \\ \lambda_2 P_{k,N-1}(1)z^{N+1} - \mu_2 z P_{k1}(2), & 2 \leq k \leq K-2 \\ \mu_1 z G_K(z) + \lambda_2 P_{K-1,N-1}(1)z^{N+1} - \mu_2 z P_{K-1,1}(2) - \mu_1 \sum\limits_{n=0}^{N-1} P_{Kn}(1)z^n, & k = K-1 \\ -\lambda_1 \sum\limits_{n=1}^{N-1} P_{K-1,n}(2)z^{n+1} - \mu_2 P_{KN}(2)z^N, & k = K \end{cases}$$

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

where

$$
\alpha(z) = \lambda_1 + \mu_1 + \lambda_2(1-z),
$$
$$
\alpha_K(z) = \mu_1 + \lambda_2(1-z),
$$

and

$$
B(z) = \begin{pmatrix}
\beta(z) & 0 & 0 & \cdots & \cdots & 0 \\
-\lambda_1 z & \beta(z) & 0 & 0 & \cdots & 0 \\
0 & -\lambda_1 z & \beta(z) & 0 & 0 & \vdots \\
\vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\
0 & \ddots & \ddots & -\lambda_1 z & \beta(z) & 0 \\
0 & \cdots & \cdots & 0 & -\lambda_1 z & \beta_K(z)
\end{pmatrix},
$$

where

$$
\beta(z) = (\lambda_2 z - \mu_2)(1-z) + \lambda_1 z,
$$
$$
\beta_K(z) = (\lambda_2 z - \mu_2)(1-z).
$$

We first explore the roots of $|A(z)| = 0$.

Theorem 3.1. *For any $\lambda_1 > 0$, $\mu_1 > 0$, $\lambda_2 > 0$ and $K \geq 1$, $|A(z)|$ is a polynomial of degree K possessing K distinct roots in the open interval $(1, \infty)$, where one of them is $z_K = 1 + \frac{\mu_1}{\lambda_2}$.*

Proof. The proof is detailed in the Appendix. □

Now, we address the roots of $|B(z)| = 0$.

Theorem 3.2. *For any $\lambda_2 > 0$, $\mu_2 > 0$, $\lambda_1 > 0$ and $K \geq 1$, $|B(z)|$ is a polynomial of degree $2(K+1)$ possessing a single root at $z^* = 1$, another root of multiplicity K, z_1, in the open interval $(0,1)$, and a root z_2 (also of multiplicity K), in the open interval $(1,\infty)$. Another root, $z_3 = \frac{\mu_2}{\lambda_2}$, exists in the open interval $(0,1)$ iff $\lambda_2 > \mu_2$.*

Proof. The matrix $B(z)$ possesses nonzero elements on the main diagonal and on the lower main diagonal. All other entries are 0. Therefore,

$$|B(z)| = \prod_{k=0}^{K} B_{kk}(z) = (\beta(z))^K \beta_K(z), \tag{3.14}$$

where $B_{kk}(z)$ is the k-th element of the diagonal of $B(z)$. The polynomial $\beta(z)$ has only two roots: $z_1 = \frac{\lambda_2+\mu_2+\lambda_1-\sqrt{(\lambda_2+\mu_2+\lambda_1)^2-4\lambda_2\mu_2}}{2\lambda_2} < 1$, and $z_2 = \frac{\lambda_2+\mu_2+\lambda_1+\sqrt{(\lambda_2+\mu_2+\lambda_1)^2-4\lambda_2\mu_2}}{2\lambda_2} > 1$. Therefore, z_1 and z_2 are roots of $|B(z)|$, each of multiplicity K. The polynomial $\beta_K(z)$ has only two roots: $z^* = 1$, and $z_3 = \frac{\mu_2}{\lambda_2}$. Clearly, $z_3 < 1$ if and only if $\lambda_2 > \mu_2$ (in which case the system is unstable). This completes the proof of Theorem 3.2. □

Note 3.1. *The root z_1 above is the Laplace-Stieltjes Transform (evaluated at λ_1) of the busy period in an $M/M/1$ queue with arrival rate λ_2 and service rate μ_2. The mean duration of such a busy period is $\frac{1}{\mu_2-\lambda_2}$, which is finite (stable system) if and only if $\lambda_2 < \mu_2$.*

3.3 The Structure of R

By explicitly writing equation (3.10) it is observed that R is an (almost fully) upper diagonal matrix with only a single non-zero element in the diagonal below the main. This is illustrated in the example below for $K = 8$.

$$R = \begin{pmatrix}
R_{00} & R_{01} & R_{02} & R_{03} & R_{04} & R_{05} & R_{06} & R_{07} & 0 & R_{09} \\
0 & R_{11} & R_{12} & R_{13} & R_{14} & R_{15} & R_{16} & R_{17} & 0 & R_{19} \\
0 & 0 & R_{22} & R_{23} & R_{24} & R_{25} & R_{26} & R_{27} & 0 & R_{29} \\
0 & 0 & 0 & R_{33} & R_{34} & R_{35} & R_{36} & R_{37} & 0 & R_{39} \\
0 & 0 & 0 & 0 & R_{44} & R_{45} & R_{46} & R_{47} & 0 & R_{49} \\
0 & 0 & 0 & 0 & 0 & R_{55} & R_{56} & R_{57} & 0 & R_{59} \\
0 & 0 & 0 & 0 & 0 & 0 & R_{66} & R_{67} & 0 & R_{69} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & R_{77} & 0 & R_{79} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & R_{87} & R_{88} & R_{89} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & R_{99}
\end{pmatrix},$$

Solving for the elements on the main diagonal, it follows that

$$
\begin{aligned}
&R_{kk} = \frac{1}{z_2}, \text{ for all } 0 \le k \le K-1,\\
&R_{KK} = \frac{1}{z_K},\\
&R_{K+1,K+1} = \frac{1}{z_3}.
\end{aligned}
$$

With $\beta_2 = \lambda_1 + \lambda_2 + \mu_2$ and $\sum_{i=1}^{0} (\cdot) \triangleq 0$, one can calculate successively the other elements of R, which results in

$$
\begin{aligned}
&R_{0k} = \frac{\lambda_1 R_{0,k-1} + \mu_2 \sum_{i=1}^{k-1} R_{0i} R_{0,k-i}}{\beta_2 - 2\mu_2/z_2}, \text{ for all } 1 \le k \le K-1,\\
&R_{l,l+k} = R_{0k}, \text{ for all } 1 \le l \le K-2, \quad 1 \le k \le K-(l+1),\\
&R_{K,K-1} = \frac{\mu_1 \frac{1}{z_K}}{\beta_2 - \mu_2 \left(\frac{1}{z_2} + \frac{1}{z_K}\right)},\\
&R_{k,K} = 0, \text{ for all } 0 \le k \le K+1, k \ne K,\\
&R_{K,K+1} = \frac{\mu_2 R_{K,K-1} R_{K-1,K+1} + \lambda_1 R_{K,K-1}}{\lambda_2 + \mu_2 - \mu_2 \left(\frac{1}{z_2} + \frac{1}{z_3}\right)}\\
&R_{K-1,K+1} = \frac{\lambda_1 \frac{1}{z_3}}{\lambda_2 + \mu_2 - \mu_2 \left(\frac{1}{z_2} + \frac{1}{z_3}\right)},\\
&R_{l,K+1} = \frac{\mu_2 \sum_{k=1}^{K-(l+1)} R_{l,l+k} R_{l+k,K+1} + \lambda_1 R_{l,K-1}}{\lambda_2 + \mu_2 - \mu_2 \left(\frac{1}{z_2} + \frac{1}{z_3}\right)}, \text{ for all } 0 \le l \le K-2.
\end{aligned}
\tag{3.15}
$$

We indicate that all the elements on the main diagonal of R are the inverse of the roots of $|A(z)| = 0$ and $|B(z)| = 0$ in the open interval $(1, \infty)$ (see Subsection 3.2), while all other elements are expressed in terms of the inverse of those roots along with parameters of the system. Furthermore, in any diagonal, starting with the main and above, all the elements along the diagonal are equal to each other, except for the last two.

4 Non-Work-Conserving Scenario

We now briefly present a Non-Work-Conserving switching scenario: if a served Q_i becomes empty, the server remains in Q_i until the number of customers in Q_j reaches its threshold. The transition-rate diagram of the triple $(L_1(t), L_2(t), I(t))$ for this scenario is depicted in Figure 4.1.

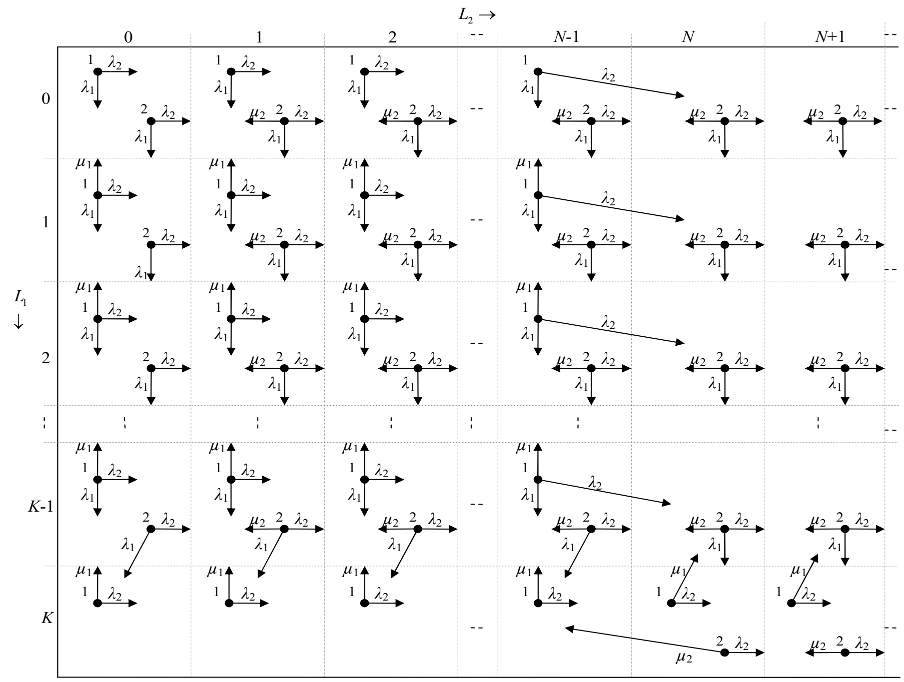

Figure 4.1: Transition-rate diagram of $(L_1(t), L_2(t), I(t))$. Non-Work-Conserving.

4.1 The QBD Process

Constructing the QBD process that represents this scenario leads to a generator matrix Q with exactly the same structure as the generator matrix in Section 3. However, some matrices appearing in Q are slightly different, namely the matrices B_0^0 and B_2^1 simply become B_0 and B_2, respectively, and B_1^0 is given by

$$B_1^0 = \begin{pmatrix} -\beta_0 & 0 & \lambda_1 & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & -\beta_0 & 0 & \lambda_1 & 0 & \cdots & \cdots & 0 \\ \mu_1 & 0 & -\beta_1 & 0 & \lambda_1 & \ddots & \cdots & 0 \\ 0 & 0 & 0 & -\beta_0 & 0 & \lambda_1 & 0 & \cdots \\ 0 & 0 & \mu_1 & 0 & -\beta_1 & 0 & \lambda_1 & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & 0 & -\beta_0 & \lambda_1 \\ 0 & \cdots & \cdots & \cdots & \cdots & \mu_1 & 0 & -(\lambda_2+\mu_1) \end{pmatrix}_{(2K+1)\times(2K+1)}.$$

All other matrices remain the same. Therefore, the stability condition for this case is $\lambda_2 < \mu_2$, as in the Work-Conserving scenario (see equation (3.8)).

Equation (3.7) above, equating the switching rates between the queues, transforms in the Non-Work-Conserving scenario into

$$\mu_1\left(P_{K\bullet}(1)-\sum_{n=0}^{N-1}P_{K,n}(1)\right)+\lambda_2\sum_{k=0}^{K-1}P_{k,N-1}(1)=\mu_2P_{KN}(2)+\lambda_1\sum_{n=0}^{N-1}P_{K-1,n}(2). \tag{4.1}$$

Notice that in this case a switch occurs only when the non served queue reaches its threshold and the served queue is beneath its threshold level.
With $\vec{P}_0=(P_{00}(1),P_{00}(2),P_{1,0}(1),P_{1,0}(2),P_{2,0}(1),P_{2,0}(2),...,P_{K-1,0}(1),P_{K-1,0}(2),P_{Kn}(1))$ and the same rate matrix, R, we have

$$\vec{P}_n=\vec{P}_NR^{n-N},\quad n\geq N, \tag{4.2}$$

As before, the vectors $\vec{P}_0$, $\vec{P}_1$, ..., $\vec{P}_{N-1}$, are obtained by solving the following set of linear equations:

$$\begin{aligned}
&\vec{P}_0B_1^0+\vec{P}_1B_2=\vec{0},\\
&\vec{P}_{n-1}B_0+\vec{P}_nB_1+\vec{P}_{n+1}B_2=\vec{0},\quad 1\leq n\leq N-2,\\
&\vec{P}_{N-2}B_0+\vec{P}_{N-1}B_1+\vec{P}_NA_2^N=\vec{0},\\
&\sum_{n=0}^{N-1}\vec{P}_n\vec{e}+\vec{P}_N[\mathcal{I}-R]^{-1}\vec{e}=1.
\end{aligned}$$

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

$$\mathbb{E}[L_1]=\sum_{n=0}^{N-1}\vec{P}_n\vec{Z}+\sum_{n=N}^{\infty}\vec{P}_n\vec{Z}_N=\sum_{n=0}^{N-1}\vec{P}_n\vec{Z}+\vec{P}_N[\mathcal{I}-R]^{-1}\vec{Z}_N, \tag{4.3}$$

$$\begin{aligned}
\mathbb{E}[L_2]&=\sum_{n=1}^{\infty}n\vec{P}_n\vec{e}=\sum_{n=1}^{N-1}n\vec{P}_n\vec{e}+\sum_{n=N}^{\infty}n\vec{P}_NR^{n-N}\vec{e}\\
&=\sum_{n=1}^{N-1}n\vec{P}_n\vec{e}+(N-1)\vec{P}_N[\mathcal{I}-R]^{-1}\vec{e}+\vec{P}_N[\mathcal{I}-R]^{-2}\vec{e},
\end{aligned} \tag{4.4}$$

where $\vec{Z}=(0,0,1,1,2,2,...,K-1,K-1,K)$ and $\vec{Z}_N=(0,1,2,...,K-1,K,K)$.

5 Numerical Examples and Comparison Between the Scenarios

This section presents several numerical results, followed by a discussion. Define SR to be the average switching rate between the queues and W_i to be the the time a customer resides in Q_i. Tables 5.1 –

5.4 exhibit results for both scenarios for the performance measures $\mathbb{E}[L_i]$, $\mathbb{E}[W_i]$, $i = 1, 2$, and SR, where $K = 10$ and $N = 3$, for different values of λ_1, λ_2, μ_1, and μ_2. In each table we investigate the impact of one of the parameters, while all other parameters remain unchanged. Specifically, Tables 5.1, 5.2, 5.3 and 5.4 present, respectively, the impact of λ_1, λ_2, μ_1, and μ_2.

Table 5.1: The impact of λ_1, when $\lambda_2 = 3$, $\mu_1 = 3$, $\mu_2 = 4$, $K = 10$ and $N = 3$

	Work-Conserving Scenario					Non-Work-Conserving Scenario				
Values of λ_1	$\mathbb{E}[L_1]$	$\mathbb{E}[L_2]$	$\mathbb{E}[W_1]$	$\mathbb{E}[W_2]$	SR	$\mathbb{E}[L_1]$	$\mathbb{E}[L_2]$	$\mathbb{E}[W_1]$	$\mathbb{E}[W_2]$	SR
0.001	0.0038	3.0008	3.7679	1.0002	0.0011	7.7611	3.0025	7768.78	1.0008	0.0005
0.01	0.0381	3.0077	3.8105	1.0026	0.0111	7.7841	3.0251	786.066	1.0084	0.0049
0.1	0.4337	3.0875	4.3382	1.0292	0.0906	7.9875	3.2348	87.3689	1.0782	0.0451
0.5	3.3355	3.5958	7.2424	1.1986	0.2570	8.5919	3.9322	24.9671	1.3107	0.1748
1	6.8243	4.2937	10.0213	1.4312	0.3198	9.0279	4.4828	17.4707	1.4943	0.2701
2	9.1070	5.0466	12.2017	1.6822	0.3768	9.4572	5.0683	14.2979	1.6894	0.3587
4	9.7206	5.5190	12.9615	1.8397	0.4103	9.7470	5.5139	13.3463	1.8380	0.4061
10	9.9126	5.8187	13.2168	1.9396	0.4097	9.9131	5.8181	13.2404	1.9394	0.4095
100	9.9924	5.9845	13.3232	1.9948	0.3804	9.9924	5.9845	13.3232	1.9948	0.3804
100000	10	6	13.3333	2	0.3750	10	6	13.3333	2	0.3750

Table 5.2: The impact of λ_2, when $\lambda_1 = 2$, $\mu_1 = 3$, $\mu_2 = 4$, $K = 10$ and $N = 3$

	Work-Conserving Scenario					Non-Work-Conserving Scenario				
Values of λ_2	$\mathbb{E}[L_1]$	$\mathbb{E}[L_2]$	$\mathbb{E}[W_1]$	$\mathbb{E}[W_2]$	SR	$\mathbb{E}[L_1]$	$\mathbb{E}[L_2]$	$\mathbb{E}[W_1]$	$\mathbb{E}[W_2]$	SR
0.001	1.8717	0.0020	0.9413	2.0281	0.0002	1.8861	1.0086	0.9489	1008.65	0.0003
0.01	1.8749	0.0203	0.9430	2.0277	0.0098	2.0171	1.0017	1.0175	100.17	0.0033
0.1	1.9281	0.1810	0.9703	1.8104	0.0828	3.1637	0.9486	1.6344	9.4862	0.0303
0.5	2.5291	0.6191	1.2863	1.2383	0.2697	5.8629	0.9968	3.2727	1.9936	0.1302
1	3.8554	1.0983	2.0402	1.0983	0.3786	7.2984	1.3298	4.4786	1.3298	0.2609
2	7.1581	2.4053	4.9487	1.2026	0.4811	8.6921	2.4900	7.1384	1.2500	0.4437
2.5	8.3307	3.4013	7.5063	1.3605	0.4655	9.1169	3.4441	9.5361	1.3776	0.4393
3	9.1070	5.0466	12.2017	1.6822	0.3768	9.4572	5.0683	14.2979	1.6894	0.3587
3.5	9.6263	9.3307	25.6908	2.6659	0.2175	9.7449	9.3449	28.5923	2.6700	0.2084
3.75	9.8203	16.9278	50.5541	4.5141	0.1188	9.8713	16.9409	55.2392	4.5176	0.1142

Table 5.3: The impact of μ_1, when $\lambda_1 = 2$, $\lambda_2 = 3$, $\mu_2 = 4$, $K = 10$ and $N = 3$

	Work-Conserving Scenario					Non-Work-Conserving Scenario				
Values of μ_1	$\mathbb{E}[L_1]$	$\mathbb{E}[L_2]$	$\mathbb{E}[W_1]$	$\mathbb{E}[W_2]$	SR	$\mathbb{E}[L_1]$	$\mathbb{E}[L_2]$	$\mathbb{E}[W_1]$	$\mathbb{E}[W_2]$	SR
0.001	9.9991	2997.74	5323.26	999.247	0.0010	9.9991	2997.74	5323.39	999.246	0.0010
0.01	9.9939	300.667	821.572	100.222	0.0068	9.9939	300.665	821.753	100.222	0.0066
0.1	9.9833	34.5552	300.373	11.5184	0.0284	9.9834	34.5548	300.677	11.5183	0.0281
0.5	9.9354	10.7279	79.4401	3.5760	0.1160	9.9368	10.7272	80.0763	3.5757	0.1143
1	9.8580	7.5220	39.4318	2.5073	0.2091	9.8684	7.5218	40.4425	2.5073	0.2036
2	9.6044	5.7383	19.213	1.9128	0.3272	9.6948	5.7441	20.7735	1.9147	0.3137
4	8.3313	4.6364	8.5645	1.5455	0.3913	9.1639	4.6948	11.107	1.5649	0.3665
10	4.8172	3.5796	3.1374	1.1932	0.4523	7.4667	3.9222	5.7012	1.3074	0.2731
100	3.0412	3.0335	1.7714	1.0112	0.5349	5.3816	3.5102	3.3611	1.1700	0.1561
100000	2.9057	3	1.6812	1	0.5387	5.1592	3.4778	3.1780	1.1592	0.1456

Table 5.4: The impact of μ_2, when $\lambda_1 = 2$, $\lambda_2 = 3$, $\mu_1 = 3$, $K = 10$ and $N = 3$

	Work-Conserving Scenario					Non-Work-Conserving Scenario				
Values of μ_2	$\mathbb{E}[L_1]$	$\mathbb{E}[L_2]$	$\mathbb{E}[W_1]$	$\mathbb{E}[W_2]$	SR	$\mathbb{E}[L_1]$	$\mathbb{E}[L_2]$	$\mathbb{E}[W_1]$	$\mathbb{E}[W_2]$	SR
3.25	9.7635	14.0703	41.7325	4.6901	0.1266	9.8288	14.0774	45.4663	4.6925	0.1229
3.5	9.5435	8.1925	22.3185	2.7308	0.2264	9.6876	8.2028	24.8562	2.7343	0.2184
3.75	9.3248	6.1210	15.5966	2.0403	0.3091	9.5642	6.1361	17.7984	2.0454	0.2963
4	9.1070	5.0466	12.2017	1.6822	0.3768	9.4572	5.0683	14.2979	1.6894	0.3587
10	4.5628	1.3895	2.4663	0.4632	0.7896	8.6360	1.5888	6.1304	0.5296	0.5712
100	2.0034	0.6175	1.0091	0.2058	1.0283	8.3652	0.7077	5.4933	0.2359	0.4774
1000	1.8836	0.5691	0.9475	0.1897	1.0454	8.3468	0.6454	5.4634	0.2151	0.4681
100000	1.8715	0.5640	0.9412	0.1880	1.0471	8.3448	0.6388	5.4603	0.2129	0.4671

Tables 5.5 – 5.8 present numerical results for the Work-Conserving Scenario, where $K = 10$. In each table, results for $N = 5$ vs. $N = 10$ are compared. Tables 5.5, 5.6, 5.7 and 5.8 show, respectively, the impact of λ_1, λ_2, μ_1 and μ_2.

Table 5.5: The impact of λ_1, when $\lambda_2 = 3$, $\mu_1 = 3$, $\mu_2 = 4$ and $K = 10$, for $N = 5$ vs. $N = 10$

	$N = 5$					$N = 10$				
Values of λ_1	$\mathbb{E}[L_1]$	$\mathbb{E}[L_2]$	$\mathbb{E}[W_1]$	$\mathbb{E}[W_2]$	SR	$\mathbb{E}[L_1]$	$\mathbb{E}[L_2]$	$\mathbb{E}[W_1]$	$\mathbb{E}[W_2]$	SR
0.001	0.0035	3.0011	3.4991	1.0004	0.0010	0.0033	3.0013	3.3457	1.0004	0.0010
0.01	0.0353	3.0115	3.5321	1.0038	0.0100	0.0337	3.0137	3.3669	1.0046	0.0097
0.1	0.3967	3.1370	3.9676	1.0457	0.0800	0.3652	3.1793	3.6526	1.0598	0.0766
0.5	3.1723	4.0857	6.8097	1.3619	0.2108	2.9212	5.0758	6.1318	1.6919	0.1835
1	6.8590	5.4946	9.9274	1.8315	0.2663	6.9622	8.8247	9.8080	2.9416	0.2280
2	9.1678	6.8062	12.2579	2.2687	0.3523	9.2348	11.5552	12.3213	3.8517	0.3418
4	9.7298	7.4634	12.9733	2.4878	0.4044	9.7337	12.4387	12.9783	4.1462	0.4031
10	9.9128	7.8144	13.2171	2.6048	0.4093	9.9127	12.8139	13.2171	4.2713	0.4093
100	9.9924	7.9845	13.3232	2.6615	0.3804	9.9924	12.9845	13.3232	4.3282	0.3804
100000	10	8	13.3333	2.6667	0.3750	10	13	13.3333	4.3333	0.3750

Table 5.6: The impact of λ_2, when $\lambda_1 = 2$, $\mu_1 = 3$, $\mu_2 = 4$ and $K = 10$, for $N = 5$ vs. $N = 10$

	$N = 5$					$N = 10$				
Values of λ_2	$\mathbb{E}[L_1]$	$\mathbb{E}[L_2]$	$\mathbb{E}[W_1]$	$\mathbb{E}[W_2]$	SR	$\mathbb{E}[L_1]$	$\mathbb{E}[L_2]$	$\mathbb{E}[W_1]$	$\mathbb{E}[W_2]$	SR
0.001	1.8717	0.0020	0.9413	2.0282	0.0010	1.8717	0.0020	0.9413	2.0282	0.0010
0.01	1.8748	0.0203	0.9429	2.0346	0.0098	1.8748	0.0203	0.9429	2.0347	0.0098
0.1	1.9152	0.2065	0.9635	2.0653	0.0820	1.9134	0.2105	0.9625	2.1051	0.0819
0.5	2.4010	0.9070	1.2173	1.8140	0.2478	2.2904	1.2223	1.1568	2.4445	0.2358
1	3.6890	1.7258	1.9385	1.7258	0.3165	3.4624	2.9291	1.7970	2.9291	0.2657
2	7.2094	3.7460	4.9463	1.8730	0.4043	7.3061	7.4733	4.9495	3.7366	0.3395
2.5	8.4136	4.9939	7.5479	1.9976	0.4163	8.5297	9.4168	7.6077	3.7667	0.3851
3	9.1678	6.8062	12.2579	2.2687	0.3523	9.2348	11.5552	12.3213	3.8517	0.3418
3.5	9.6540	11.1979	25.7421	3.1994	0.2089	9.6786	16.0952	25.7896	4.5986	0.2068
3.75	9.8336	18.8324	50.6023	5.0202	0.1151	9.8444	23.7698	50.6418	6.3386	0.1145

Table 5.7: The impact of μ_1, when $\lambda_1 = 2$, $\lambda_2 = 3$, $\mu_2 = 4$ and $K = 10$, for $N = 5$ vs. $N = 10$

	$N = 5$					$N = 10$				
Values of μ_1	$\mathbb{E}[L_1]$	$\mathbb{E}[L_2]$	$\mathbb{E}[W_1]$	$\mathbb{E}[W_2]$	SR	$\mathbb{E}[L_1]$	$\mathbb{E}[L_2]$	$\mathbb{E}[W_1]$	$\mathbb{E}[W_2]$	SR
0.001	9.9991	2999.66	5323.38	999.886	0.0009	9.9991	3004.59	5323.48	1001.53	0.0009
0.01	9.9939	302.617	821.678	100.872	0.0066	9.9939	307.577	821.763	102.526	0.0066
0.1	9.9834	36.5466	300.388	12.1822	0.0282	9.9834	41.5401	300.4	13.8467	0.0281
0.5	9.9359	12.7071	79.4439	4.2357	0.1145	9.9361	17.6921	79.4459	5.8974	0.1142
1	9.8611	9.4703	39.4444	3.1568	0.2041	9.8630	14.4321	39.4518	4.8107	0.2028
2	9.6257	7.6086	19.2534	2.5362	0.3129	9.6427	12.4986	19.2857	4.1662	0.3083
4	8.4234	6.2305	8.5797	2.0768	0.3571	8.5580	10.6908	8.6138	3.5636	0.3353
10	4.6743	4.1513	2.9030	1.3838	0.4273	4.4013	5.7058	2.5209	1.9019	0.3831
100	2.7595	3.0671	1.5244	1.0224	0.5422	2.3724	3.1326	1.2205	1.0442	0.5493
100000	2.6128	3	1.4336	1	0.5477	2.2490	3	1.1521	1	0.5592

Table 5.8: The impact of μ_2, when $\lambda_1 = 2$, $\lambda_2 = 3$, $\mu_1 = 3$ and $K = 10$, for $N = 5$ vs. $N = 10$

	$N = 5$					$N = 10$				
Values of μ_2	$\mathbb{E}[L_1]$	$\mathbb{E}[L_2]$	$\mathbb{E}[W_1]$	$\mathbb{E}[W_2]$	SR	$\mathbb{E}[L_1]$	$\mathbb{E}[L_2]$	$\mathbb{E}[W_1]$	$\mathbb{E}[W_2]$	SR
3.25	9.7811	15.9701	41.7695	5.3234	0.1233	9.7945	20.9096	41.8002	6.9699	0.1228
3.5	9.5774	10.051	22.3682	3.3503	0.2180	9.6069	14.9445	22.4138	4.9815	0.2157
3.75	9.3734	7.9324	15.6516	2.6441	0.2936	9.4213	12.7628	15.7077	4.2543	0.2880
4	9.1678	6.8062	12.2579	2.2687	0.3523	9.2348	11.5552	12.3213	3.8517	0.3418
10	4.4701	2.2132	2.4084	0.7377	0.5783	4.2789	4.1358	2.2890	1.3786	0.4236
100	1.9898	1.1127	1.0022	0.3709	0.7879	1.9633	2.0981	0.9887	0.6994	0.6333
1000	1.8823	1.0462	0.9468	0.3487	0.8025	1.8797	1.9929	0.9455	0.6643	0.6445
100000	1.8714	1.0392	0.9412	0.3464	0.8040	1.8714	1.9819	0.9412	0.6606	0.6456

Discussion

1. When comparing the two scenarios (Tables 5.1 – 5.4), the average switching rate between the queues, SR, is *always* smaller in the Non-Work-Conserving scenario than in the Work-Conserving scenario, while the opposite statement holds for $\mathbb{E}[L_i]$ and $\mathbb{E}[W_i]$, $i = 1, 2$. This occurs since, in the Non-Work-Conserving scenario, the server may remain idle in an empty queue even if there are waiting customers in the other queue, causing a decrease in the switching rate on the one hand, and an increase in mean queue sizes and mean waiting times,

on the other hand.

2. When $\lambda_1 \to \infty$ or $\mu_1 \to 0$ the performance measures of the two scenarios approach the same values, independently of all other parameters. See Tables 5.1 and 5.3, for $\lambda_1 \geq 10$, and $\mu_1 \leq 0.5$, respectively.

3. When the arrival rate into one of the queues, say λ_i, is relatively small, the corresponding measures $\mathbb{E}[L_i]$ and $\mathbb{E}[W_i]$ in the Non-Work-Conserving scenario are significantly greater than the comparable values in the Work-Conserving scenario (see Tables 5.1 and 5.2). This follows since the server remains idle in an empty queue as long as the threshold level in the opposite queue has not been reached. Hence, the customers of Q_i wait for a long time until Q_i's threshold is reached, upon which the server is called for service there.

4. In the Non-Work-Conserving scenario, initially, as λ_1 increases, $\mathbb{E}[W_1]$ decreases. However for large values of λ_1 ($\lambda_1 \geq 10$), $\mathbb{E}[W_1]$ increases as λ_1 increases. This occurs since increasing values of λ_1 cause L_1 to ascend at a faster rate, which increases the switching rate and decreases the time intervals between switches. In the Work-Conserving scenario, $\mathbb{E}[W_1]$ increases when λ_1 increases (see Table 5.1). Notice that $\mathbb{E}[L_1]$ increases in both scenarios.

5. Table 5.2 exhibits an apparently counter-intuitive phenomenon for both scenarios, namely, as λ_2 increases, $E[W_2]$ first decreases and then increases. A similar phenomenon occurs in the Non-Work-Conserving scenario for the values of $E[L_2]$.

6. In Tables 5.1, 5.2, 5.5 and 5.6, SR first increases and then decreases when λ_1 (λ_2) increases, the exact point of change in direction (increase or decrease) depends on the entire set of parameters. This occurs in both scenarios. In contrast, for the Work-Conserving scenario, when λ_1 and λ_2 are fixed but μ_1 or μ_2 increase (Tables 5.3, 5.4, 5.7 and 5.8), SR always increases.

7. The rate of service μ_2 of the unbounded Q_2 has a more profound effect then μ_1 (the service rate of the bounded Q_1) on the values of $E[L_1]$ and $E[L_2]$ (and consequently on $E[W_1]$ and $E[W_2]$). See Tables 5.3, 5.4, 5.7 and 5.8.

6 Extreme Cases

We investigate the influence of extreme values of λ_1, λ_2, μ_1 and μ_2 (as they reach 0 or ∞) on the system's performance measures in the two different switching scenarios. Some of the cases (e.g. $\lambda_2 \to \infty$ or $\mu_2 \to 0$, $\lambda_1 \to 0$, $\lambda_2 \to 0$) follow directly from basic queueing principles. Other cases require more intricate analysis.

We first address extreme cases that lead to identical system structure in the two policies, and then address extreme cases that lead to different system structures.

$\underline{\lambda_2 \to \infty \text{ or } \mu_2 \to 0}$

These two cases are not stable, since the stability condition, $\lambda_2 < \mu_2$, is not satisfied.

$\underline{\mu_1 \to 0}$

The system is unstable. Once the server attends Q_1 and the number of customers there is at its threshold level, meaning that $L_1 = K$, the number of customers there will not reduce below the threshold level and the server will never switch back to Q_2 even when the number of customers in Q_2 reaches its threshold, N. Therefore, the number of customers in Q_2 will increase to ∞.

$\underline{\lambda_1 \to 0}$

It is clear that in both scenarios, $\mathbb{P}(I = 1) = 0$ and $\mathbb{P}(I = 2) = 1$, meaning that Q_2 operates as an $M(\lambda_2)/M(\mu_2)/1$ system. Therefore, $\mathbb{P}(L_1 = 0) = 1$ and $P_{\text{loss}}(1) \equiv \mathbb{P}(L_1 = K) = 0$. Clearly then, $\mathbb{E}[L_2] = \frac{\rho_2}{1-\rho_2}$, where $\rho_i = \frac{\lambda_i}{\mu_i}$, $i = 1, 2$.

$\underline{\lambda_2 \to 0}$

It is straightforward that $\mathbb{P}(I = 1) = 1$, and $\mathbb{P}(I = 2) = 0$. Therefore, Q_1 operates as an $M(\lambda_1)/M(\mu_1)/1/K$ system for which $P_{\text{loss}}(1) = \frac{\rho_1^K(1-\rho_1)}{1-\rho_1^{K+1}}$, and $\mathbb{E}[L_1] = \frac{\rho_1}{1-\rho_1} - \frac{(K+1)\rho_1^{K+1}}{1-\rho_1^{K+1}}$.

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

When $\lambda_1 \to \infty$, Q_1 is always at its maximum capacity, meaning $L_1 \equiv K$ 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 its maximum value, N. Then, at the next instant when the server completes a service of a customer in Q_1, it immediately switches to Q_2. Then, before a service completion in Q_2, an arrival at Q_1 will occur, causing a switch back to Q_1 as soon as the number of customers at Q_2 reduces below N. Hence, the only possible states with nonzero probabilities are $(K, n, 1)$, for $n \geq N-1$, and $(K, n, 2)$, for $n \geq N$. Therefore, $\mathbb{P}(I = 1) = \sum_{n=N-1}^{\infty} P_{Kn}(1) = P_{K\bullet}(1)$, and $\mathbb{P}(I = 2) = \sum_{n=N}^{\infty} P_{Kn}(2) = P_{K\bullet}(2)$. As a

consequence, $\mathbb{P}(I=1) = 1-\rho_2$, $\mathbb{P}(I=2) = \rho_2$ and $\mathbb{E}[L_2] = \frac{\rho_2}{1-\rho_2} + N - (1-\rho_2) + \frac{\lambda_2}{\mu_1}$.
Note that the parameter K does not appear in any of the results above.

The next two extreme cases lead to a different system structure in each of the switching scenarios.

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

In the Work-Conserving switching scenario, if $\mu_1 \to \infty$ then, whenever the server is at Q_1, he immediately reduces the number of customers there to 0, and will remain at Q_1 until the first moment thereafter that a customer arrives at Q_2. 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 its own threshold, N. If Q_1 reaches its threshold and Q_2 is *not* below its threshold, the server stays at Q_2 until the number of customers there is reduced below Q_2's threshold, upon which the server switches to Q_1, and immediately empties the queue and returns to Q_2. Note that in this case $P_{\text{loss}}(1) = P_{K\bullet}(2)$.
In the Non-Work-Conserving switching scenario, $\mathbb{P}(I=1) = P_{0\bullet}(1) = \sum_{n=0}^{N-1} P_{0n}(1)$. The server will remain at Q_1 until the first moment when the number of customers in Q_2 reaches the value N and will remain there until the number of customers in Q_1 reaches the value K. Then, given that Q_2 is below the threshold N, the server will switch to Q_1 and immediately reduce the occupancy there to 0. Note that in this case $P_{\text{loss}}(1) = P_{K\bullet}(2)$ as well.

$\underline{\mu_2 \to \infty}$

In the Work-Conserving switching scenario, when $\mu_2 \to \infty$, the server immediately empties Q_2 upon entering it and resides there until a customer arrives at Q_1. Therefore, $\mathbb{P}(I=2) = P_{00}(2)$. The server remains in Q_1 until Q_2 reaches its threshold and Q_1 is below the threshold K. If Q_2 reaches its threshold and Q_1 is *not* below its threshold, the server stays at Q_1 until the number of customers there reduces below K, upon which the server switches to Q_2, empties it instantaneously, and returns to Q_1. In this case $P_{\text{loss}}(1) = P_{K\bullet}(1)$.
In the Non-Work-Conserving case, $\mathbb{P}(I=2) = P_{\bullet 0}(2) = \sum_{k=0}^{K-1} P_{k0}(2)$. The server will remain at Q_2 until the first moment when the number of customers in Q_1 reaches the value K and will remain there until the number of customers in Q_2 reaches the value N. Then, if Q_1 is below the threshold K, the server will switch to Q_2 and immediately reduce the occupancy there to 0. In this case, too, $P_{\text{loss}}(1) = P_{K\bullet}(1)$.

7 Concluding Remarks and Future Investigations

This paper studies a two-queue polling-type system with a non-orthodox threshold-based switching policy, which depends on the queue that is not being served. Employing the Matrix Geometric method, we derive the joint steady-state probabilities of the system's state and its performance measures. We reveal that the entries of the main diagonal of the rate matrix R of the Matrix Geometric are the reciprocal of the roots of matrices defining the Probability Generating Functions associated with the phases of the QBD process. We remark that this phenomenon appears in other studies such as Paz and Yechiali [15], Perel N. and Yechiali [17], Phung-Duc [18], and Hanukov et al. [9]. This relationship has not been shown analytically as a general property and it calls for further investigation. Furthermore, unlike many cases in which the rate matrix is calculated numerically, we are able to derive closed-form expressions for all the elements of R. Another direction of research is to study the non-preemptive version of the model. A third direction, which is much more involved, is to assume that the switch-over times are non-zero. Finally, the analysis of the case when both capacities are infinite seems to be a challenging task.

Acknowledgement We thank Nir Perel for instructive discussions during the course of this research.

References

[1] Avrachenkov, K.; Perel, E.; Yechiali, U. Finite-buffer polling systems with threshold-based switching policy. TOP **2016**, 24, 541–571.

[2] Avram, F.; Gómez-Corral, A. On the optimal control of a two-queue polling model. Operations Research Letters **2006**, 34, 339–348.

[3] Boon, M.M.A.; van der Mei, R.D.; Winands, E.M.M. Applications of polling systems. Surveys in Operations Research and Management Science **2011**, 16, 67–82.

[4] Boxma, O.J.; Down, D.G. Dynamic server assignment in a two-queue model. European Journal of Operational Research **1997**, 103, 595–609.

[5] Boxma, O.J.; Levy, H.; Yechiali, U. Cyclic reservation schemes for efficient operation of multiple-queue single-server systems. Annals of Operations Research **1992**, 35, 187–208.

[6] Boxma, O.J.; Koole, G.; Mitrani, I. A two-queue polling model with a threshold service policy. In Dowd, P. and Gelenbe, E., eds., Proceedings MASCOTS '95, IEEE Computer Society Press, Los Alamitos, CA, **1995**, 84–89.

[7] Boxma, O.J.; Koole, G.; Mitrani, I. Polling models with threshold switching. In Baccelli, F., Jean-Marie, A., and Mitrani, I., eds., Quantitative Methods in Parallel Systems '95, Springer Verlag, Berlin, **1995**, 129–140.

[8] Bright, L.W.; Taylor, P.G. Calculating the equilibrium distribution in level dependent quasi-birth-and-death processes. Stochastic Models **1995**, 11, 497–526.

[9] Hanukov, G.; Avinadav, T.; Chernonog, T.; Spiegel, U.; Yechiali, A queueing system with decomposed service and inventoried preliminary services. Accepted for publication in Applied Mathematical Modelling, **2017**, http://dx.doi.org/10.1016/j.apm.2017.03.008.

[10] Haverkot, B.; Idzenga, H.P.; Kim, B.G. Performance evaluation of threshold-based ATM cell scheduling policies under Markov modulated Poisson traffic using stochastic Petri nets. In Proceedings IFIP '95, Performance Modelling and Evaluation of ATM Networks, Chapman & Hall, **1995**, 553–572.

[11] Latouche, G.; Ramaswami, V. *Introduction to Matrix Analytic Methods in Stochastic Modeling*; SIAM and ASA: Philadelphia, 1999.

[12] Lee, D.-S. A two-queue model with exhaustive and limited service disciplines. Communications in Statistics. Stochastic Models **1996**, 12, 285–305.

[13] Lee, D.-S.; Sengupta, B. Queueing analysis of a threshold based priority scheme for ATM networks. IEEE/ACM Transactions on Networking **1993**, 1, 709–717.

[14] Neuts, M.F. *Matrix Geometric Solutions in Stochastic Models - an Algorithmic Approach.* The Johns Hopkins University Press: Baltimore and London, 1981.

[15] Paz, N.; Yechiali, U. An $M/M/1$ queue in random environment with disasters. Asia-Pacific Journal of Operational Research **2014**, 31(3), 1450016 (12 pages). DOI: 10.1142/S021759591450016X.

[16] Perel, E. Queues with Customers Acting as Servers & Polling Systems with a Threshold-Based Switching Policy. Ph.D. Dissertation, Department of Statistics & Operations Research, School of Mathematical Sciences, Tel-Aviv University, 2014.

[17] Perel N.; Yechiali, U. The Israeli Queue with priorities. Stochastic Models **2013**, 29, 353–379.

[18] Phung-Duc, T. Exact solutions for M/M/c/setup queues. Telecommunication Systems, **2016**, DOI: 10.1007/s11235-016-0177-z.

[19] Phung-Duc, T.; Masuyama, H.; Kasahara, S.; Takahashi, Y. A simple algorithm for the rate matrices of level-dependent QBD processes. In Proceedings of the 5th international conference on queueing theory and network applications, ACM. **2010**, 46–52.

[20] Takagi, H. *Analysis of Polling Systems.* The MIT Press, 1986.

[21] Yechiali, U. Analysis and control of polling systems. In Donatiello, L. and Nelson, R., editors, *Performance Evaluation of Computer and Communication Systems, Joint Tutorial Papers of Performance '93 and Sigmetrics '93*, London, UK. Springer Berlin, **1993**, 729, 630–650.

A Appendix

Proof of Theorem 3.1

Proof. Let $q_0(z) = 1$. Define the minors of the diagonal of $A(z)$, starting from the upper left-hand corner, as follows:

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

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

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

From (A.1) and (A.2) we conclude that

1. By definition, $q_0(z) = 1$ and therefore has no roots.

2. For every $1 \leq k \leq K-1$, $q_k(z)$ 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 $q_{k-2}(z)$, $q_{k-3}(z)$, ..., $q_0(z)$, which contradicts the above conclusion.

3. $\text{Sign}(q_k(\infty)) = (-1)^k$, for all k.

4. $q_k(1) = \sum_{i=0}^{k} \lambda_1^i \mu_1^{k-i} > 0$, for all $0 \leq i \leq K$.

5. $q_K(1) = \mu_1 \sum_{i=0}^{K-1} \lambda_1^i \mu_1^{K-1-i} > 0$.

6. Given $\tilde{z}$, a root of $q_k(z)$, then $\text{sign}(q_{k-1}(\tilde{z}) q_{k+1}(\tilde{z})) = -1$, for every $1 \leq k \leq K-2$.

7. $q_k(z)$ is a polynomial of degree k for all $0 \leq k \leq K$.

From the above conclusions it follows that $q_1(z)$ has only one root, $z_{1,1} = 1 + \frac{\lambda_1 + \mu_1}{\lambda_2} > 1$. $q_2(1) = \sum_{i=0}^{2} \lambda_1^i \mu_1^{2-i} > 0$, $q_2(z_{1,1}) < 0$, $q_2(\infty) > 0$. 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 $q_3(1) = \sum_{i=0}^{3} \lambda_1^i \mu_1^{3-i} > 0$, $q_3(z_{2,1}) < 0$, $q_3(z_{2,2}) > 0$, $q_3(\infty) < 0$. 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: $z_{k,1} \in (1, z_{k-1,1})$, $z_{k,2} \in (z_{k-1,1}, z_{k-1,2})$, ..., $z_{k,k} \in (z_{k-1,k-1}, \infty)$.
$q_K(z) = \alpha_K(z) q_{K-1}(z)$ has K roots, where $K-1$ of them are the $K-1$ distinct roots of $q_{K-1}(z)$ and another root (which appears in the matrix R whose structure is discussed in Subsection 3.3) is

$$
z_K = 1 + \frac{\mu_1}{\lambda_2}. \tag{A.3}
$$

This completes the proof of Theorem 3.1. $\square$