A polling model with threshold switching

Onno Boxma* David Perry† Rachel Ravid‡ Uri Yechiali§

April 14, 2025

Abstract

We consider a single-server two-queue Markovian polling system with the following special feature. If the server is serving the infinite-buffer queue Q_2 and the single-buffer queue Q_1 is empty, then it stays at Q_2 until it has become empty; but if a customer joins an empty Q_1, then the server only stays at Q_2 as long as that queue has at least N customers (the threshold). If that customer joins Q_1 while Q_2 has less than N customers, then service at Q_2 is preempted and the server instantaneously switches to Q_1. Arrivals to Q_1 when it is occupied are blocked and lost. This threshold discipline contrasts with the classical multi-queue polling model, where switching instants are typically determined by the length of the queue being served. We (i) derive explicit expressions for the joint queue length distribution; (ii) analyze the busy period distribution by employing an original approach that uses taboo states; and (iii) determine the sojourn time distribution for customers in both queues.

Keywords: polling model; threshold switching policy; busy period; sojourn time.

1 Introduction

The typical polling model is a single-server, multi-queue system in which the server cyclically visits the queues. The system can thus be interpreted as a priority system with cyclically changing priority. At each queue, the stay of

*Department of Mathematics and Computer Science, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands; o.j.boxma@tue.nl

†Holon Institute of Technology, P.O. Box 305, Holon 5810201, Israel; davidper@hit.ac.il

‡ORT Braude College of Engineering Karmiel, Department of Industrial Engineering and Management, P.O. Box 78, Karmiel 2161002, Israel; rachel3@braude.ac.il

§Department of Statistics and Operations Research, School of Mathematical Sciences, Tel-Aviv University, Tel-Aviv 6997801, Israel; uriy@tauex.tau.ac.il

the server is determined by some switching policy, that usually just depends on the queue length at that queue. For example, under the exhaustive service policy the server serves the queue until it is empty, and under the k-limited policy it serves k customers or leaves when the queue has become empty – whichever event occurs first. However, there are also everyday-life situations which are more naturally modelled by a polling model in which the switching policy not only depends on the queue presently being served, but also on the state of other queues.

In the present paper, we consider a 2-queue instance of such a system. Q_1 has a single buffer and Q_2 has an infinite buffer. The arrival processes at the queues are independent non-identical Poisson processes, and the service requirements at both queues are non-identically exponentially distributed. Type-1 customers arriving while Q_1 is occupied are blocked and lost. If the server is serving Q_2 and Q_1 is empty, then the server stays at Q_2 until it has become empty. But if the server is serving Q_2 and a type-1 customer arrives at an empty Q_1, then the server only stays at Q_2 as long as it has at least N customers. If it has less than N customers when the type-1 customer arrives, then the service at Q_2 is preempted and the server instantaneously switches to Q_1.

Motivation. Our main motivation is two-fold. Firstly, we believe it is interesting to study a polling model in which the switching policy is not mainly determined by the queue length of the queue that is presently being served, but depends on a threshold value for the queue length at *the other queue*. Moreover, this gives rise to challenging issues regarding the sojourn time distribution at Q_2, because the sojourn time of customers at Q_2 is influenced by *later arrivals* at this queue.

Secondly, the polling model under consideration naturally represents a production system with maintenance. Arrival of a type-1 customer corresponds to the need to do preventive maintenance. That maintenance will start immediately, unless the number (queue) of items to be produced is equal to or larger than a threshold N; then maintenance is postponed until that number drops below the threshold. Rejection of type-1 arrivals if there is already a type-1 customer corresponds to the situation in which, when maintenance has already been called for, there is no need to repeat the warning. Alternatively (as suggested by a referee), one might view type-1 customers as status updates. A new type-1 arrival then simply replaces an old status update. This would lead to exactly the same queue length and busy period distributions as in our model, but for the sojourn time analysis of type-1 customers one would need a minor adaptation (because some type-1 customers are replaced by another type-1 customer).

Another example is that of a server who serves both premium customers who require a distinguished (typically long) service, and standard customers. There may be various rewards and costs involved, and in particular the holding/sojourn costs of a (typically rare) premium customer may be very significant. In such a situation one might wish to determine a threshold level N for switching from the queue of standard customers to a premium customer. While threshold optimization is outside the scope of the present paper, our work does lie the foundation for such an optimization study.

Related literature. While there is a huge literature on priority queues (see, e.g., the books of Jaiswal [11] and Takagi [17]) and on the analysis of polling models (see, e.g., the methodology-oriented survey [4] and the application-oriented survey [3]) and their control (see, e.g., [8, 19]), there are only few papers devoted to such systems in which the priority changes are based on a threshold value. Some examples are the papers of Lee [13], Lee and Sengupta [14], Haverkort et al. [10], Boxma et al. [5, 6, 7], Avram and Gómez-Corral [2], and the thesis of Perel [15]. In most of those studies, the switching policy is determined by the state of the queue that is presently being served. Exceptions are Avrachenkov et al. [1] and two papers that formed the main motivation for the present study: Perel and Yechiali [16] and Jolles et al. [12]. Our model is a special case of the 2-queue models of [16] (where Q_1 has a finite buffer of size at least one) and [12] (where, moreover, a switch between queues requires a positive switchover time). However, the relative simplicity of our model allows a detailed analysis of not only the joint queue length distribution, but also of both the busy period distribution and the distribution of the sojourn times at both queues. The sojourn time distribution at Q_2 is particularly hard to determine, because the sojourn time of a customer at Q_2 depends not only on the customers in front of her/him, but also on customers arriving after her/him: when there are at least $N-1$ customers behind her, then during the remainder of her stay (s)he is "immune" against a possible preemption by an arrival at Q_1.

Main results. The main contributions of the paper are:

- A detailed queue length analysis, leading to explicit expressions for the joint queue length probabilities.
- Analysis of the busy period distribution; for this we introduce an original approach, using taboo states.
- Analysis of the sojourn time distributions of both types of customers. For type 2 this requires keeping track of the number of customers of type 2 in front of *and behind* the tagged type-2 customer. The queueing

literature does not offer that many examples where a detailed sojourn time analysis is possible even though both earlier and later arrivals of the same priority level affect a customer's sojourn time; one example is the sojourn time analysis for the $M/G/1$ processor sharing queue [18].

- We briefly outline how the non-preemptive case can be analyzed.

Organization of the paper. Section 2 presents a detailed model description and summarizes a few known results for busy periods that play a key role in the sequel. Section 3 is devoted to the joint queue length distribution at both queues. The busy period distribution is treated in Section 4, and the sojourn time distribution for customers of both queues is determined in Section 5. Numerical results are presented and discussed in Section 6. In Section 7 we briefly indicate how the analysis has to be adapted to treat the non-preemptive case.

2 Model description and preliminaries

The main model under consideration is a single-server two-queue Markovian system. Customers arrive at the two queues according to independent Poisson arrival processes, with rate λ_i at Q_i, requiring independent exponentially distributed service times with rate μ_i at Q_i, $i = 1, 2$. We assume that Q_1 has room for just one customer; a type-1 customer who finds Q_1 non-empty is blocked and lost. Q_2, on the other hand, has an infinite buffer.

The single server switches between queues (with zero switchover times) in the following way. If the server is at Q_1, it continues serving the (single) customer present until completion and then moves to Q_2 if that queue is non-empty. If the server is at Q_2, it continues serving customers until either Q_2 becomes empty or until the number at Q_2 drops below threshold N while Q_1 is non-empty. In the latter case, the server switches to Q_1 to serve the single customer who is there present. If a service at Q_2 was interrupted, that service is resumed (or started anew; that probabilistically makes no difference in this case of exponentially distributed service times) once the service at Q_1 is completed.

In Figure 1 we display the state space with its state-transition diagram for $N = 2, 3, 4, 5$. We characterize the states by a 3-tuple, where the first index denotes the number of type-1 customers, the second index the number of type-2 customers, and the third index the type of customer in service ($i = 1, 2$); an empty system is denoted by $(0, 0)$. Considering, e.g., the

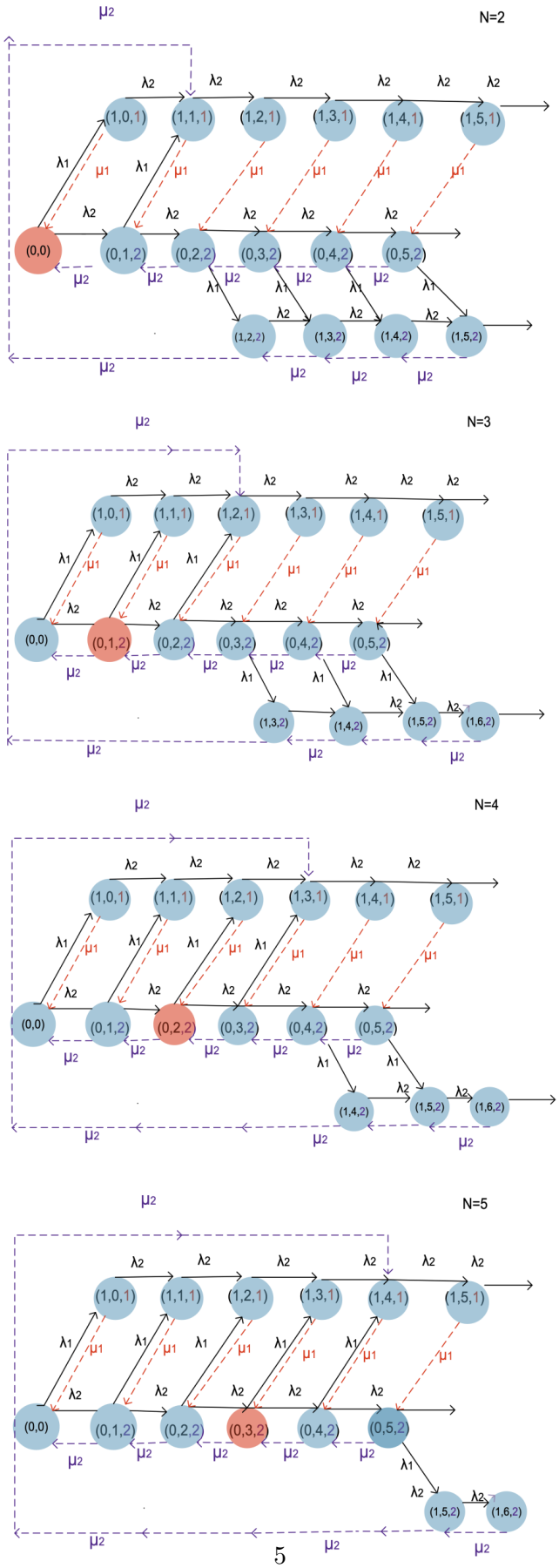

Figure 1: State-transition diagram, for $N = 2, 3, 4, 5$ (with buffer size $K = 1$ in Q_1).

case $N = 4$, a transition from state $(1, 4, 2)$ to state $(1, 3, 1)$ occurs with rate μ_2, corresponding to a departure from Q_2 that causes the number of customers at Q_2 to drop below the threshold level $N = 4$: the server switches instantaneously to Q_1. Furthermore, a transition from $(0, n, 2)$ to $(1, n, 1)$ can occur for $n < N$, as seen in the Figure: an arrival (at rate λ_1) at Q_1 causes the server to switch preemptively, when Q_2 contains less than N customers.

Remark 1 *One could also consider the non-preemptive case. Switches to another queue then only occur at service completion epochs. We briefly discuss this variant in Section 8.*

The ergodicity condition. Let $\rho_2 := \lambda_2/\mu_2$ denote the load of Q_2 in isolation (without interference of Q_1). Clearly, $\rho_2 < 1$ is a necessary ergodicity condition. Some thought will make clear that this condition is also sufficient; for a proof we refer to [16]. Q_2 basically behaves as an $M/M/1$ queue with vacations/interruptions (e.g., maintenance). When it has become empty, the server starts a vacation that may include the service of a type-1 customer; when Q_2 is being served and its queue length drops below N, it may also become interrupted.

The busy period distribution for Q_2 in isolation. Intuitively it is clear that busy periods of Q_2 feature in the queue length behavior of Q_1. We shall see that the busy period of the ordinary $M/M/1$ queue with arrival rate λ_2 and service rate μ_2 plays a key role in the analysis of queue lengths, busy periods and sojournm times of the 2-queue model under consideration. Hence we end this section with a brief recapitulation of the results concerning this busy period (cf. Sections II.2.2 and II.4.4 of [9]).

Let $\Gamma(\alpha)$ denote the Laplace-Stieltjes transform (LST) of a busy period length B_2 in this $M/M/1$ queue. It is given by the unique zero of the equation $x = \beta(\alpha + \lambda_2(1 - x))$ in $|x| \leq 1$, with $\beta(\alpha) = \mu_2/(\mu_2 + \alpha)$ the service time LST. Hence $\Gamma(\alpha)$ is the zero of

$$\lambda_2 x^2 - (\lambda_2 + \mu_2 + \alpha)x + \mu_2 = 0 \tag{1}$$

with absolute value $|x| \leq 1$. Hence we have

$$\Gamma(\alpha) \;=\; \frac{\lambda_2 + \mu_2 + \alpha - \sqrt{(\lambda_2 + \mu_2 + \alpha)^2 - 4\lambda_2\mu_2}}{2\lambda_2}. \tag{2}$$

The mean busy period length is given by

$$\mathbb{E}[B_2] = \frac{1}{\mu_2 - \lambda_2}. \tag{3}$$

3 The steady-state queue length probabilities

In this section we determine the steady-state joint queue length distribution, for general $N \geq 2$. Our treatment will be rather brief, because determining the queue length distribution is not our main goal in this paper, while moreover the queue lengths have been obtained for a slightly more general model in [16], in matrix form (via a QBD approach). However, the fact that Q_1 can contain at most one customer allows us to obtain a more explicit expression for the queue length probabilities than in [16].

The balance equations are as follows (with $I(\cdot)$ an indicator function):

$$\begin{aligned}
(\lambda_1+\lambda_2)P(0,0) &= \mu_1 P(1,0,1)+\mu_2 P(0,1,2), && (4)\\
(\lambda_2+\mu_1)P(1,0,1) &= \lambda_1 P(0,0), && (5)\\
(\lambda_2+\mu_1)P(1,n,1) &= \lambda_1 P(0,n,2)+\lambda_2 P(1,n-1,1) && (6)\\
&\quad +\mu_2 P(1,n+1,2)I(n=N-1), \quad n=1,\ldots,N-1,\\
(\lambda_2+\mu_1)P(1,n,1) &= \lambda_2 P(1,n-1,1), \quad n\geq N, && (7)\\
(\lambda_1+\lambda_2+\mu_2)P(0,n,2) &= \lambda_2 P(0,0)I(n=1)+\lambda_2 P(0,n-1,2)I(n\geq 2) && (8)\\
&\quad +\mu_1 P(1,n,1)+\mu_2 P(0,n+1,2), \quad n\geq 1,\\
(\lambda_2+\mu_2)P(1,n,2) &= \lambda_1 P(0,n,2)+\lambda_2 P(1,n-1,2)I(n\geq N+1)\\
&\quad +\mu_2 P(1,n+1,2), \quad n\geq N. && (9)
\end{aligned}$$

Note that the states $(1,n,2)$ for $n=1,\ldots,N-1$ do not occur. For the non-preemptive case, minor adaptations have to be made, cf. Section 8.
It follows from (7) that

$$P(1,n,1) \;=\; (\frac{\lambda_2}{\lambda_2+\mu_1})^{n-N+1}P(1,N-1,1), \quad n\geq N-1, \qquad (10)$$

with generating function

$$P_{11}(z) \;:=\; \sum_{n=N-1}^{\infty} P(1,n,1)z^n \;=\; \frac{(\lambda_2+\mu_1)z^{N-1}}{\mu_1+\lambda_2(1-z)}P(1,N-1,1). \qquad (11)$$

We also introduce the generating functions $P_{02}(z) := \sum_{n=1}^{\infty} P(0,n,2)z^n$ and $P_{12}(z) := \sum_{n=N}^{\infty} P(1,n,2)z^n$. Taking generating functions in (8) results in

$$\begin{aligned}
\left(\lambda_1+\lambda_2(1-z)+\mu_2(1-\frac{1}{z})\right)P_{02}(z) &= \lambda_2 z P(0,0)-\mu_2 P(0,1,2)+\mu_1 P_{11}(z)\\
&\quad + \mu_1\sum_{n=1}^{N-2} P(1,n,1)z^n, && (12)
\end{aligned}$$

and hence

$$\begin{aligned}\lambda_2(z-z_+)(z-z_-)P_{02}(z) &= \mu_2 z P(0,1,2) - \lambda_2 z^2 P(0,0) - \mu_1 z P_{11}(z) \\ &- \mu_1 z \sum_{n=1}^{N-2} P(1,n,1) z^n, \end{aligned} \tag{13}$$

where

$$z_\pm = \frac{\lambda_1+\lambda_2+\mu_2 \pm \sqrt{(\lambda_1+\lambda_2+\mu_2)^2 - 4\lambda_2\mu_2}}{2\lambda_2}. \tag{14}$$

Observe (cf. Section 2) that z_- equals the LST $\Gamma(\lambda_1) = \mathbb{E}(\mathrm{e}^{-\lambda_1 B_2})$ of a busy period B_2 of $M/M/1$ queue Q_2 in isolation (i.e., with arrival rate λ_2 and service rate μ_2), evaluated at λ_1 (cf. (2)); alternatively,
$z_- = \mathbb{P}(B_2$ ends before an arrival at $Q_1)$. Furthermore, $z_+ z_- = \mu_2/\lambda_2 > 1$, and hence $z_+ > 1$. Since $P_{02}(z)$ is finite for $|z| \leq 1$, the righthand side of (13) must be zero for $z = z_-$, yielding

$$\mu_2 P(0,1,2) = \lambda_2 z_- P(0,0) + \mu_1 P_{11}(z_-) + \mu_1 \sum_{n=1}^{N-2} P(1,n,1) z_-^n. \tag{15}$$

We can thus divide both sides of (13) by a factor $z - z_-$, and using (11) we conclude that

$$\begin{aligned} P_{02}(z) &= \frac{z}{\lambda_2(z_+ - z)}[\lambda_2 P(0,0) \\ &+ \mu_1 \frac{\mu_1+\lambda_2}{\mu_1+\lambda_2(1-z_-)} P(1,N-1,1) \frac{1}{\mu_1+\lambda_2(1-z)} \\ &\times [(\mu_1+\lambda_2)\frac{z^{N-1}-z_-^{N-1}}{z-z_-} - \lambda_2 z z_- \frac{z^{N-2}-z_-^{N-2}}{z-z_-}] \\ &+ \mu_1 \sum_{n=1}^{N-2} P(1,n,1) \frac{z^n - z_-^n}{z-z_-}]. \end{aligned} \tag{16}$$

After a lengthy calculation we obtain, for $n \geq 1$,

$$\begin{aligned} P(0,n,2) &= (\frac{1}{z_+})^n [P(0,0) + \frac{\mu_1}{\lambda_2} \sum_{j=0}^{N-3} \sum_{h=j+1}^{N-2} P(1,h,1) z_-^{h-1-j} z_+^j \\ &+ \frac{\mu_1}{\lambda_2} \frac{\mu_1+\lambda_2}{\mu_1+\lambda_2(1-z_-)} P(1,N-1,1) \times \\ &\times \{\sum_{m=0}^{n-1} \sum_{j=0}^{N-2} z_-^{N-2-j} (\frac{\lambda_2}{\mu_1+\lambda_2})^{m-j} z_+^m - \frac{\mu_2}{\mu_1+\lambda_2} \sum_{m=0}^{n-2} \sum_{j=0}^{N-3} z_-^{N-3-j} (\frac{\lambda_2}{\mu_1+\lambda_2})^{m-j} z_+^m\}]. \end{aligned} \tag{17}$$

Observe that the product of the two zeroes of the quadratic equation $\lambda_2 z^2 - (\lambda_2 + \lambda_1 + \mu_2)z + \mu_2 = 0$ equals μ_2/λ_2, and hence $1/z_+$ can also be written as $\frac{\lambda_2}{\mu_2} z_- = \frac{\lambda_2}{\mu_2}\Gamma(\lambda_1)$.

Let us now turn to the $P(1,n,2)$. Taking generating functions in (9) results in

$$\left(\lambda_2(1-z) + \mu_2(1-\frac{1}{z})\right) P_{12}(z) \;=\; \lambda_1 \sum_{n=N}^{\infty} P(0,n,2)z^n - \mu_2 P(1,N,2)z^{N-1}. \tag{18}$$

Taking $z = 1$ yields $\mu_2 P(1,N,2) = \lambda_1 \sum_{n=N}^{\infty} P(0,n,2)$ (which can be achieved directly by taking a horizontal 'cut' between the second and third lines in each graph of Figure 1), and hence we have

$$P_{12}(z) = \frac{\lambda_1 z}{\mu_2 - \lambda_2 z} \frac{\sum_{n=N}^{\infty} P(0,n,2)(z^{N-1} - z^n)}{1-z}. \tag{19}$$

A straightforward inversion of the generating function allows us to express $P(1,n,2)$ in terms of the just obtained $P(0,k,2)$, $k \geq N$: For $n \geq N$,

$$P(1,n,2) = \frac{\lambda_1}{\mu_2 - \lambda_2}\left[-\sum_{k=N}^{n} P(0,k,2)(1 - (\frac{\lambda_2}{\mu_2})^{n-k}) + \sum_{k=N}^{\infty} P(0,k,2)(1 - (\frac{\lambda_2}{\mu_2})^{n-N+1}) \right]. \tag{20}$$

It remains to determine the $N+2$ unknown constants $P(0,0)$, $P(1,k,1)$, $k = 0,\ldots,N-1$, and $P(0,1,2)$. For this we use the following $N+1$ equations, which have not yet been used: (4), (5), (6), and the normalizing condition as a final equation.

4 The busy period

In this section we study the length B of the busy period, the time between an arrival in an empty system and the first time thereafter that the server becomes idle again. First some notation: the LST of B is denoted by $\Psi(\alpha)$. The LST of the time to go from some state (i,j,k) (with $i = 0,1; j = 0,1,\ldots; k = 1,2$) to state $(0,0)$ is denoted by $\phi_{i,j,k}(\alpha)$. We shall demonstrate how all $\phi_{i,j,k}(\alpha)$ can be determined. In particular, this will yield $\Psi(\alpha)$, as (condition on the type of arrival that starts a busy period)

$$\Psi(\alpha) = \frac{\lambda_1}{\lambda_1 + \lambda_2}\phi_{1,0,1}(\alpha) + \frac{\lambda_2}{\lambda_1 + \lambda_2}\phi_{0,1,2}(\alpha). \tag{21}$$

In some applications, for example when a type-1 customer indicates an interruption, one may be just interested in the time to go from $(0,1,2)$ to

$(0,0)$, and hence just in $\phi_{0,1,2}(\alpha)$.
Below we shall study $\phi_{0,1,2}(\alpha)$; but let us first consider $\phi_{1,0,1}(\alpha)$. There are two possible transitions out of $(1,0,1)$, viz., to $(0,0)$ and to $(1,1,1)$; hence, using that the minimum of an $\exp(\lambda_2)$ and an $\exp(\mu_1)$ random variable is an $\exp(\lambda_2+\mu_1)$ random variable,

$$\begin{aligned}\phi_{1,0,1}(\alpha) &= \frac{\mu_1}{\mu_1+\lambda_2}\frac{\mu_1+\lambda_2}{\mu_1+\lambda_2+\alpha}+\frac{\lambda_2}{\lambda_2+\mu_1}\frac{\lambda_2+\mu_1}{\lambda_2+\mu_1+\alpha}\phi_{1,1,1}(\alpha)\\ &= \frac{\mu_1}{\lambda_2+\mu_1+\alpha}+\frac{\lambda_2}{\lambda_2+\mu_1+\alpha}\phi_{1,1,1}(\alpha). \qquad (22)\end{aligned}$$

From $(1,1,1)$ the process will eventually reach one of the states $(0,n+1,2)$, for $n=0,1,\ldots$. If there are n type-2 arrivals before the completion of the type-1 service, we end up in state $(0,n+1,2)$. Hence

$$\phi_{1,1,1}(\alpha)=\sum_{n=0}^{\infty}(\frac{\lambda_2}{\lambda_2+\mu_1+\alpha})^n\frac{\mu_1}{\lambda_2+\mu_1+\alpha}\phi_{0,n+1,2}(\alpha). \qquad (23)$$

The same reasoning in fact reveals that, more generally,

$$\phi_{1,k,1}(\alpha)=\sum_{n=0}^{\infty}(\frac{\lambda_2}{\lambda_2+\mu_1+\alpha})^n\frac{\mu_1}{\lambda_2+\mu_1+\alpha}\phi_{0,n+k,2}(\alpha). \qquad (24)$$

For (21) we still need $\phi_{0,1,2}(\alpha)$; but from (23) we see that for $\phi_{1,1,1}(\alpha)$ we actually need not only $\phi_{0,1,2}(\alpha)$ but all $\phi_{0,n+1,2}(\alpha)$, $n=0,1,\ldots$. We shall determine these in two steps. In Step 1 we express $\phi_{0,N+k,2}(\alpha)$ in terms of $\phi_{0,N,2}(\alpha)$ and $\phi_{1,N,2}(\alpha)$. In Step 2 we obtain N linear equations for $\phi_{0,k,2}(\alpha)$ for $k=1,\ldots,N$.
Step 1. Now comes the hardest part of the analysis: Starting in $(0,N+k,2)$ the process always ends up in one of the two states $(0,N,2)$ or $(1,N,2)$ (remember that a switch to Q_1 cannot occur until the number of type-2 customers has dropped below N). Hence our aim is now to express $\phi_{0,N+k,2}(\alpha)$ in terms of $\phi_{0,N,2}(\alpha)$ and $\phi_{1,N,2}(\alpha)$. For this we need two additional functions, which will be determined in two lemma's. Firstly, let $\phi^*(\alpha)$ denote the *improper* LST of the time to go from $(0,n+1,2)$ to $(0,n,2)$ *without passing through* $(1,n+1,2)$, for any $n\geq N$. This LST is the same for all those n, because of the homogeneity of the transition probabilities. We determine it in the next lemma.

Lemma 1

$$\phi^*(\alpha)=\Gamma(\alpha+\lambda_1). \qquad (25)$$

Proof. Remember that one can interpret $\Gamma(\alpha)$ as the probability that B_2 (a busy period of Q_2 in isolation) is smaller than an $\exp(\alpha)$ random variable (see below (14)). Using the same reasoning, the LST of the time to go from $(0, n+1, 2)$ to $(0, n, 2)$ *without having a type*-1 *arrival* can be interpreted as the probability that B_2 is smaller than the minimum of an $\exp(\alpha)$ and an $\exp(\lambda_1)$ random variable, i.e., an $\exp(\alpha+\lambda_1)$ random variable. Accordingly, this LST equals $\Gamma(\alpha+\lambda_1)$. ■

Now we introduce a second improper LST: $\phi_*(\alpha)$ is the LST of the time to go from $(0, n+1, 2)$ to $(1, n, 2)$ *without passing through* $(0, n, 2)$*, for any* $n \geq N$. Again, this LST is the same for all those n. We determine it in the next lemma.

Lemma 2

$$\begin{aligned} \phi_*(\alpha) &= \frac{\lambda_1\Gamma(\alpha)}{\alpha+\lambda_1+\mu_2+\lambda_2(1-\phi^*(\alpha)-\Gamma(\alpha))} \\ &= \frac{\lambda_1\Gamma(\alpha)}{\alpha+\lambda_1+\mu_2+\lambda_2(1-\Gamma(\alpha+\lambda_1)-\Gamma(\alpha))}. \end{aligned} \tag{26}$$

Proof. Formula (26) follows once we have proven the following relation:

$$\begin{aligned} \phi_*(\alpha) &= \frac{\lambda_1}{\lambda_1+\lambda_2+\mu_2+\alpha}\Gamma(\alpha) + \frac{\lambda_2}{\lambda_1+\lambda_2+\mu_2+\alpha}\phi^*(\alpha)\phi_*(\alpha) \\ &+ \frac{\lambda_2}{\lambda_1+\lambda_2+\mu_2+\alpha}\phi_*(\alpha)\Gamma(\alpha). \end{aligned} \tag{27}$$

The latter relation is proven by distinguishing the three disjoint trajectories to go from $(0, n+1, 2)$ to $(1, n, 2)$ with $(0, n, 2)$ as *taboo* state; they correspond to the three terms in the righthand side of (27).
The first trajectory starts with a step from $(0, n+1, 2)$ to $(1, n+1, 2)$. It subsequently takes a busy period with LST $\Gamma(\alpha)$ to reach $(1, n, 2)$ from $(1, n+1, 2)$.
The second and third trajectory both begin with a step from $(0, n+1, 2)$ to $(0, n+2, 2)$. The second trajectory goes from $(0, n+2, 2)$ to $(0, n+1, 2)$ (not necessarily in one step) *without hitting taboo state* $(1, n+1, 2)$. That is represented by the improper LST $\phi^*(\alpha)$. We are now back in the state $(0, n+1, 2)$; hence the multiplicative factor $\phi_*(\alpha)$.
The third trajectory goes from $(0, n+2, 2)$ to $(1, n+1, 2)$ *without passing through taboo state* $(0, n+1, 2)$. That is represented by the improper LST $\phi_*(\alpha)$. Finally, the LST of the time to go from $(1, n+1, 2)$ to $(1, n, 2)$ equals $\Gamma(\alpha)$.

The second line of (26) follows by using Lemma 1. ■

It should be noticed that $\phi^*(0)$ represents the probability to go from some state $(0, n+1, 2)$ to $(0, n, 2)$ without passing through $(1, n+1, 2)$, for any $n \geq N$, and that $\phi_*(0) = 1 - \phi^*(0)$.
We now have the tools to express $\phi_{0,N+k,2}(\alpha)$ in terms of $\phi_{0,N,2}(\alpha)$ and $\phi_{1,N,2}(\alpha)$.

Lemma 3

$$\phi_{0,N+k,2}(\alpha) = (\phi^*(\alpha))^k \phi_{0,N,2}(\alpha) + \frac{(\Gamma(\alpha))^k - \phi^*(\alpha)^k}{\Gamma(\alpha) - \phi^*(\alpha)} \phi_*(\alpha) \phi_{1,N,2}(\alpha). \quad (28)$$

Proof. We argue that the following relation holds (from which (28) immediately follows):

$$\phi_{0,N+k,2}(\alpha) = (\phi^*(\alpha))^k \phi_{0,N,2}(\alpha) + \sum_{j=0}^{k-1} (\phi^*(\alpha))^j \phi_*(\alpha) \Gamma(\alpha)^{k-1-j} \phi_{1,N,2}(\alpha). \quad (29)$$

The first term in the righthand side of (29) corresponds to the case in which $(0, N, 2)$ is reached without a type-1 arrival. The second term distinguishes between the states $(0, N+k-j, 2)$, for $j = 0, \ldots, k-1$, in which a type-1 arrival occurs. The $(\phi^*(\alpha))^j$ factor represents the move from $(0, N+k, 2)$ to $(0, N+k-j, 2)$ without a type-1 arrival. The factor $\phi_*(\alpha)$ represents the transition to $(1, N+k-1-j, 2)$ without reaching $(0, N+k-1-j, 2)$. It finally takes a sum of $k-1-j$ independent ordinary type-2 busy periods, with LST $\Gamma(\alpha)$, to reach state $(1, N, 2)$. ■

Above we have expressed $\phi_{0,N+k,2}(\alpha)$ in terms of $\phi_{0,N,2}(\alpha)$ and $\phi_{1,N,2}(\alpha)$. In Step 2 we shall determine $\phi_{0,k,2}(\alpha)$ for $k = 1, \ldots, N$. At this end of Step 1 we consider $\phi_{1,N,2}(\alpha)$. That term is easily expressed in $\phi_{1,N-1,1}(\alpha)$:

$$\phi_{1,N,2}(\alpha) = \Gamma(\alpha) \phi_{1,N-1,1}(\alpha), \quad (30)$$

and in fact

$$\phi_{1,N+k,2}(\alpha) = \Gamma(\alpha)^{k+1} \phi_{1,N-1,1}(\alpha), \quad k = 0, 1, \ldots \ . \quad (31)$$

Indeed, referring to the bottom row of the state-transition diagram in Figure 1, observe that the queue length process in Q_2 in these states behaves like the queue length process in an $M/M/1$ queue with arrival rate λ_2 and

service rate μ_2, until state $(1, N, 2)$ is left because of a service completion; the process then moves to state $(1, N-1, 1)$. It therefore takes the sum of $k+1$ independent $M/M/1$ busy periods (each with LST $\Gamma(\alpha)$) to go from state $(1, N+k, 2)$ to state $(1, N-1, 1)$.

Step 2. It remains to determine $\phi_{0,k,2}(\alpha)$ for $k = 1, \ldots, N$. By once more distinguishing the various transition probabilities, we derive N linear equations in these unknowns:

$$\begin{aligned} \phi_{0,k,2}(\alpha) &= \frac{\mu_2}{\lambda_1+\lambda_2+\mu_2+\alpha}\phi_{0,k-1,2}(\alpha) + \frac{\lambda_2}{\lambda_1+\lambda_2+\mu_2+\alpha}\phi_{0,k+1,2}(\alpha) \\ &+ \frac{\lambda_1}{\lambda_1+\lambda_2+\mu_2+\alpha}\phi_{1,k,1}(\alpha), \qquad k = 1, \ldots, N-1, \end{aligned} \tag{32}$$

where we define $\phi_{0,0,2}(\alpha) \equiv 1$. The Nth equation differs from the above ones in the last term:

$$\begin{aligned} \phi_{0,N,2}(\alpha) &= \frac{\mu_2}{\lambda_1+\lambda_2+\mu_2+\alpha}\phi_{0,N-1,2}(\alpha) + \frac{\lambda_2}{\lambda_1+\lambda_2+\mu_2+\alpha}\phi_{0,N+1,2}(\alpha) \\ &+ \frac{\lambda_1}{\lambda_1+\lambda_2+\mu_2+\alpha}\phi_{1,N,2}(\alpha). \end{aligned} \tag{33}$$

This concludes the busy period analysis.
To summarize:
* All $\phi_{1,k,1}(\alpha)$ (including $\phi_{1,1,1}(\alpha)$ that we need to evaluate (21)) are expressed in terms of $\phi_{0,n+1,2}(\alpha)$; cf. (23) and (24).
* All $\phi_{1,N+k,2}(\alpha)$ are expressed in terms of $\phi_{1,N-1,1}(\alpha)$, and hence they are also expressed in terms of the $\phi_{0,n+1,2}(\alpha)$; cf. (31).
* All $\phi_{0,N+k,2}(\alpha)$ are expressed in terms of $\phi_{0,N,2}(\alpha)$ (and $\phi_{1,N,2}(\alpha)$, but see the sentence above).
* Finally, the remaining N unknown functions $\phi_{0,k,2}(\alpha)$, for $k = 1, \ldots, N$, are determined by solving the N linear equations (32), (33).

Remark 2 *Observe that, from renewal theory,* $P(0,0) = \frac{\frac{1}{\lambda_1+\lambda_2}}{\frac{1}{\lambda_1+\lambda_2}+\mathbb{E}B}$ *(the ratio of mean idle period and sum of mean idle period and mean busy period, where the latter mean follows from the above busy period analysis). This offers the possibility to determine* $P(0,0)$ *without using the normalizing condition, as done in the previous section.*

5 The sojourn times

In this section we analyze the sojourn time LST's of a tagged type-1 customer and of a tagged type-2 customer. The former is easy, but the latter is challenging because of the threshold feature. Indeed, when a type-2 customer is taken into service, it makes a difference whether or not (s)he has at least $N-1$ customers behind her/him. If yes, then (s)he is "immune" to the presence of a type-1 customer, i.e., protected against a possible preemption, and her/his remaining sojourn time equals her/his service time. If not, various interruptions might occur. Hence later arrivals may affect the sojourn time of the type-2 customer, and we need to keep track of those arrivals.

5.1 The sojourn time of an admitted type-1 customer

Theorem 1 *The sojourn time LST $\Phi_1(\alpha)$ of a type-1 customer who is admitted to the system is given by*

$$\Phi_1(\alpha) = \frac{\mu_1}{\mu_1+\alpha}\left[P(0,0) + \sum_{k=1}^{N-1} P(0,k,2) + \sum_{k=N}^{\infty} P(0,k,2)\Gamma^{k-N+1}(\alpha)\right]. \tag{34}$$

Proof. By PASTA, the queue-length arrival probabilities are the same as the steady-state queue-length probabilities. When (s)he sees state $(0,0)$, or states $(0,k,2)$ with less than N type-2 customers, (s)he immediately is taken into service (and remember that type-1 customers are never interrupted); hence the $\frac{\mu_1}{\mu_1+\alpha}$ factor. When (s)he finds $k \geq N$ type-2 customers present, (s)he waits until that number drops to $N-1$, i.e., $k-N+1$ type-2 $M/M/1$ busy periods, with LST $\Gamma(\alpha)$. The type-1 customer then gets priority and is served. ■

5.2 The sojourn time of a type-2 customer

We first present a similar relation as (34) for the sojourn time LST $\Phi_2(\alpha)$ of a type-2 customer, but in that relation we have the still unknown LST's $f_{i,j+1,k}(\alpha)$, which represent the LST of the sojourn time of a type-2 customer when upon her/his arrival (s)he sees the system in state (i,j,k):

$$\begin{aligned}\Phi_2(\alpha) &= P(0,0)f_{0,1,2}(\alpha) + \sum_{j=1}^{N-1} P(0,j,2)f_{0,j+1,2}(\alpha) + \sum_{j=N}^{\infty} P(0,j,2)f_{0,j+1,2}(\alpha) \\ &+ \sum_{j=0}^{\infty} P(1,j,1)f_{1,j+1,1}(\alpha) + \sum_{j=N}^{\infty} P(1,j,2)f_{1,j+1,2}(\alpha). \end{aligned} \tag{35}$$

We now derive recurrence relations involving those $f_{i,j+1,k}(\alpha)$. In order to keep track of the number of customers *behind* the tagged type-2 customer, we add a fourth index m, writing $f_{i,j+1,k;m}(\alpha)$; $f_{i,j+1,k;0}(\alpha)$ equals $f_{i,j+1,k}(\alpha)$. Here $f_{i,j+1,k;m}(\alpha)$ is the LST of the remaining sojourn time of the tagged type-2 customer when there are i type-1 customers, j type-2 customers ahead of her/him and m behind her/him, and the server serves a type-k customer.

The terms $f_{1,j+1,1;m}(\alpha)$.

Distinguishing between the various possible events, we can write for $j = 0, 1, \ldots,\ m = 0, 1, \ldots, N-2$:

$$f_{1,j+1,1;m}(\alpha) = \frac{\lambda_2}{\lambda_2+\mu_1+\alpha} f_{1,j+1,1;m+1}(\alpha) + \frac{\mu_1}{\lambda_2+\mu_1+\alpha} f_{0,j+1,2;m}(\alpha), \tag{36}$$

while for $j = 0, 1, \ldots$:

$$f_{1,j+1,1;N-1}(\alpha) = \frac{\mu_1}{\mu_1+\alpha}(\frac{\mu_2}{\mu_2+\alpha})^{j+1}. \tag{37}$$

The rationale behind this last formula is that the tagged type-2 customer is, in her/his remaining sojourn, immune to type-1 customers. (S)he still needs to wait for the completion of the type-1 customer in service, but her/his remaining sojourn time is the sum of $j+1$ service times.

The terms $f_{0,j+1,2;m}(\alpha)$.

We have, for $j = 0, 1, \ldots,\ m = 0, 1, \ldots, N-2$, when $j+1+m \geq N$ (so that the type-2 customer in service is immune):

$$\begin{aligned} f_{0,j+1,2;m}(\alpha) &= \frac{\lambda_2}{\lambda_1+\lambda_2+\mu_2+\alpha} f_{0,j+1,2;m+1}(\alpha) + \frac{\mu_2}{\lambda_1+\lambda_2+\mu_2+\alpha} f_{0,j,2;m}(\alpha) \\ &+ \frac{\lambda_1}{\lambda_1+\lambda_2+\mu_2+\alpha} f_{1,j+1,2;m}(\alpha), \end{aligned} \tag{38}$$

and when $j+1+m < N$ (no immunity) the last term is different:

$$\begin{aligned} f_{0,j+1,2;m}(\alpha) &= \frac{\lambda_2}{\lambda_1+\lambda_2+\mu_2+\alpha} f_{0,j+1,2;m+1}(\alpha) + \frac{\mu_2}{\lambda_1+\lambda_2+\mu_2+\alpha} f_{0,j,2;m}(\alpha) \\ &+ \frac{\lambda_1}{\lambda_1+\lambda_2+\mu_2+\alpha} f_{1,j+1,1;m}(\alpha), \end{aligned} \tag{39}$$

while, for $j = 0, 1, \ldots$:

$$f_{0,j+1,2;N-1}(\alpha) = (\frac{\mu_2}{\mu_2+\alpha})^{j+1}. \tag{40}$$

The terms $f_{1,j+1,2;m}(\alpha)$.

Notice that $j+1+m$ now should be at least N: if a type-2 customer is being served in the presence of a type-1 customer, (s)he apparently has immunity. We have, for such j, m, with $m = 0, 1, \ldots, N-2$, and in the first equation taking $j+1+m \geq N+1$:

$$f_{1,j+1,2;m}(\alpha) = \frac{\lambda_2}{\lambda_2+\mu_2+\alpha} f_{1,j+1,2;m+1}(\alpha) + \frac{\mu_2}{\lambda_2+\mu_2+\alpha} f_{1,j,2;m}(\alpha), \quad (41)$$

$$f_{1,j+1,2;N-1-j}(\alpha) = \frac{\lambda_2}{\lambda_2+\mu_2+\alpha} f_{1,j+1,2;N-j}(\alpha) + \frac{\mu_2}{\lambda_2+\mu_2+\alpha} f_{1,j,1;N-1-j}(\alpha), \quad (42)$$

where we took $j = 0, 1, \ldots, N-1$, and finally

$$f_{1,j+1,2;m}(\alpha) = (\frac{\mu_2}{\mu_2+\alpha})^{j+1}, \qquad m = N-1, N, \ldots \ . \quad (43)$$

Let us now take stock.
Equations (36) and (37) allow us to express all $f_{1,j+1,1;m}(\alpha)$ in terms of $f_{0,j+1,2;k}(\alpha)$, $k = 0, \ldots, N-2$.
Equations (38), (39) and (40) allow us to express all $f_{0,j+1,2;m}(\alpha)$, $m = 0, 1, \ldots, N-2$, in terms of $f_{1,r+1,2;k}(\alpha)$, $k = 0, \ldots, N-2$, $r \leq j$, and, when $j+1+m < N$, in $f_{1,j+1,1;m}(\alpha)$.
Equations (41)-(43) allow us to express all $f_{1,j+1,2;m}(\alpha)$ in terms of the $N-1$ unknown $f_{1,r,1;N-1-r}(\alpha)$, $r = 1, \ldots, N-1$. Using (36), they are expressed in terms of the $N-1$ unknown $f_{0,r,2;N-1-r}(\alpha)$ (for $N = 2$, see below, this would just be one equation relating $f_{1,1,1}(\alpha)$ and $f_{0,1,2}(\alpha)$). Via (39) with $m = N-2-j$, we have another $N-1$ relations, that express the $f_{0,r,2;N-1-r}(\alpha)$ in terms of the $f_{1,r,1;N-1-r}(\alpha)$ (for $N = 2$, that would be just one equation). We thus obtain $2(N-1)$ equations in the remaining $2(N-1)$ unknowns. Having a procedure to determine all conditional sojourn time LST's, we can determine the unconditional sojourn time LST of a type-2 customer via (35).

The case $N = 2$.
As the above analysis is quite intricate, let us elaborate the procedure for a threshold $N = 2$. In this case we can immediately determine the terms $f_{i,j+1,k;1}(\alpha)$, because in all these terms the tagged type-2 customer is immune for a preemption (and there is no need to consider $f_{i,j+1,k;m}(\alpha)$ for $m \geq 2$). We have, for $j = 0, 1, \ldots$:

$$f_{1,j+1,1;0}(\alpha) = \frac{\lambda_2}{\lambda_2+\mu_1+\alpha} \frac{\mu_1}{\mu_1+\alpha} (\frac{\mu_2}{\mu_2+\alpha})^{j+1} + \frac{\mu_1}{\lambda_2+\mu_1+\alpha} f_{0,j+1,2;0}(\alpha), \quad (44)$$

and for $j = 1, 2, \ldots$:

$$\begin{aligned} f_{0,j+1,2;0}(\alpha) &= \frac{\lambda_2}{\lambda_1+\lambda_2+\mu_2+\alpha}(\frac{\mu_2}{\mu_2+\alpha})^{j+1} \\ &+ \frac{\mu_2}{\lambda_1+\lambda_2+\mu_2+\alpha} f_{0,j,2;0}(\alpha) + \frac{\lambda_1}{\lambda_1+\lambda_2+\mu_2+\alpha} f_{1,j+1,2;0}(\alpha), \end{aligned} \tag{45}$$

and for $j = 2, 3, \ldots$:

$$f_{1,j+1,2;0}(\alpha) = \frac{\mu_2}{\lambda_2+\mu_2+\alpha} f_{1,j,2;0}(\alpha) + \frac{\lambda_2}{\lambda_2+\mu_2+\alpha}(\frac{\mu_2}{\mu_2+\alpha})^{j+1}, \tag{46}$$

while

$$f_{1,2,2;0}(\alpha) = \frac{\mu_2}{\lambda_2+\mu_2+\alpha} f_{1,1,1;0}(\alpha) + \frac{\lambda_2}{\lambda_2+\mu_2+\alpha}(\frac{\mu_2}{\mu_2+\alpha})^{2}. \tag{47}$$

It follows from (46) and (47) that, for $j = 1, 2, \ldots$:

$$f_{1,j+1,2;0}(\alpha) = \frac{\mu_2}{\mu_2+\alpha}\left[(\frac{\mu_2}{\mu_2+\alpha})^j - (\frac{\mu_2}{\lambda_2+\mu_2+\alpha})^j\right] + (\frac{\mu_2}{\lambda_2+\mu_2+\alpha})^j f_{1,1,1;0}(\alpha). \tag{48}$$

Let us now take stock.
Equation (44) allows us to express all $f_{1,j+1,1;0}(\alpha)$ in terms of $f_{0,j+1,2;0}(\alpha)$.
Equation (45) allows us to express all $f_{0,j+1,2;0}(\alpha)$ $(j \geq 1)$ in terms of the various $f_{1,k,2;0}(\alpha)$, $k = 2, \ldots, j+1$.
Equation (48) allows us to express all $f_{1,j+1,2;0}(\alpha)$ $(j \geq 1)$ in terms of $f_{1,1,1;0}(\alpha)$.
It remains to determine $f_{1,1,1;0}(\alpha)$ and $f_{0,1,2;0}(\alpha)$. For $j = 0$, (45) has to be replaced by the following equation:

$$\begin{aligned} f_{0,1,2;0}(\alpha) &= \frac{\lambda_2}{\lambda_1+\lambda_2+\mu_2+\alpha}\frac{\mu_2}{\mu_2+\alpha} + \frac{\mu_2}{\lambda_1+\lambda_2+\mu_2+\alpha} \\ &+ \frac{\lambda_1}{\lambda_1+\lambda_2+\mu_2+\alpha} f_{1,1,1;0}(\alpha), \end{aligned} \tag{49}$$

and (44) for $j = 0$ states that

$$f_{1,1,1;0}(\alpha) = \frac{\lambda_2}{\lambda_2+\mu_1+\alpha}\frac{\mu_1}{\mu_1+\alpha}\frac{\mu_2}{\mu_2+\alpha} + \frac{\mu_1}{\lambda_2+\mu_1+\alpha} f_{0,1,2;0}(\alpha). \tag{50}$$

The two equations (49) and (50) are trivially solved to determine the two unknowns $f_{1,1,1;0}(\alpha)$ and $f_{0,1,2;0}(\alpha)$. Knowledge of $f_{1,1,1;0}(\alpha)$ then leads to

knowledge of all $f_{1,j+1,2;0}(\alpha)$ via (48). Finally, writing (45) as a recurrence equation

$$f_{0,j+1,2;0}(\alpha) = \frac{\mu_2}{\lambda_1+\lambda_2+\mu_2+\alpha} f_{0,j,2;0}(\alpha) + A_j(\alpha), \quad j = 1, 2, \ldots, \quad (51)$$

where all $A_j(\alpha) := \frac{\lambda_2}{\lambda_1+\lambda_2+\mu_2+\alpha}(\frac{\mu_2}{\mu_2+\alpha})^{j+1} + \frac{\lambda_1}{\lambda_1+\lambda_2+\mu_2+\alpha} f_{1,j+1,2;0}(\alpha)$ are now explicitly known, yields for $j = 0, 1, \ldots$:

$$f_{0,j+1,2;0}(\alpha) = (\frac{\mu_2}{\lambda_1+\lambda_2+\mu_2+\alpha})^j f_{0,1,2;0}(\alpha) + \sum_{k=1}^{j} (\frac{\mu_2}{\lambda_1+\lambda_2+\mu_2+\alpha})^{j-k} A_k(\alpha). \quad (52)$$

Thus, all conditional sojourn time LST are known, and (35) provides the unconditional sojourn time LST of a type-2 customer.

6 Numerical results

In this section we present some numerical results, in eight tables. Each table gives the probability $P(0,0)$ of an empty system, the mean busy period length $E[B]$, and the mean numbers of customers $E[L_i]$ and mean sojourn times $E[W_i]$ of class-i customers, $i = 1, 2$. Tables 1 and 2 study the impact of varying λ_1, for $N = 3$ resp. $N = 10$; Tables 3 and 4 do the same for λ_2; in Tables 5 and 6, μ_1 is varied, and in Tables 7 and 8 we vary μ_2.

First some general observations: (i) The following relation exists between $P(0,0)$ and $E[B]$: $P(0,0) = \frac{\frac{1}{\lambda_1+\lambda_2}}{\frac{1}{\lambda_1+\lambda_2}+E[B]}$ (where $1/(\lambda_1+\lambda_2)$ is the mean idle period). (ii) We have $E[W_2] = E[L_2]/\lambda_2$ (Little's law), and $E[W_1] = \frac{E[L_1]}{\lambda_1(1-E[L_1])}$. The latter result also is in agreement with Little's law, as $1 - E[L_1]$ here equals the probability of Q_1 being empty, and hence (PASTA) $\lambda_1(1-E[L_1]) = \lambda_1 P(L_1 = 0)$ is the arrival rate of accepted class-1 customers. (iii) Also notice that we have included some extreme values of the λ_i and μ_i, both to see the effect of extreme traffic conditions and as a check of (known) limiting cases.

Table 1: The impact of λ_1; $\lambda_2 = 3$, $\mu_1 = 3$, $\mu_2 = 4$ and $N = 3$

λ_1	$P(0,0)$	$E[B]$	$E[L_1]$	$E[L_2]$	$E[W_1]$	$E[W_2]$
0.01	0.2467	1.0143	0.0192	3.2010	1.9625	1.0670
0.1	0.2214	1.1343	0.1427	4.2898	1.6649	1.4299
0.5	0.1503	1.6148	0.4020	6.0003	1.3444	2.0001
1	0.1026	2.1875	0.5577	6.9081	1.2609	2.3027
2	0.0568	3.3241	0.7101	7.7872	1.2249	2.5957
4	0.0244	5.7064	0.8308	8.5069	1.2277	2.8356
10	0.0053	14.4769	0.9266	9.1362	1.2621	3.0454
100	$1.9329e^-5$	502.2853	0.9925	9.6748	1.3234	3.2249

Table 2: The impact of λ_1; $\lambda_2 = 3$, $\mu_1 = 3$, $\mu_2 = 4$ and $N = 10$

λ_1	$P(0,0)$	$E[B]$	$E[L_1]$	$E[L_2]$	$E[W_1]$	$E[W_2]$
0.01	0.2467	1.0145	0.0055	3.1203	0.5552	1.0401
0.1	0.2184	1.1544	0.0521	4.1241	0.5498	1.3747
0.5	0.1230	2.0369	0.2381	8.6542	0.6250	2.8847
1	0.0580	4.0567	0.4241	13.9994	0.7366	4.6665
2	0.0139	14.2280	0.6458	21.0619	0.9116	7.0206
4	0.0013	106.9819	0.8135	26.6259	1.0905	8.8753
10	$1.3631e^{-5}$	$5.6433e^3$	0.9250	30.2022	1.2334	10.0674
100	$2.1092e^{-13}$	$4.6030e^{10}$	0.9925	32.2670	1.3288	10.7557

Our main findings are the following. In Tables 1 and 2 (focusing on the impact of λ_1) we see that $P(0,0)$ quickly decreases from $1-\rho = 0.25$ ($\lambda_1 = 0$) to (very) low values when λ_1 increases, while the mean busy period quickly increases from $\frac{1}{\mu_2-\lambda_2} = 1$ to (very) large values. The effect of increasing λ_1 on $E[L_2]$ is relatively mild when $N = 3$, but it is more pronounced when $N = 10$; the explanation is that class-2 customers are only 'protected' against a switch to Q_1 when L_2 is at least N. A final remark regarding Tables 1 and 2: Interestingly, $E[W_1]$ is not monotone in λ_1, as it would have been for an $M/M/1/1$ queue in isolation. In line with this, we see for very small values of λ_1 that $E[L_2]$ is a bit *larger* for $N = 3$ than for $N = 10$; in virtually all other cases, the reverse is the case, caused by the protection of class-2 customers against class-1 customers.

In Tables 3 and 4 (focusing on the impact of λ_2) we see that $P(0,0)$ decreases from $1/(1+\lambda_1/\mu_1)$ (for $\lambda_2 = 0$) to very low values when λ_2 increases to its extreme value 3.75, while the mean busy period quickly increases from $1/\mu_1$ to large values (very large for $N = 10$). $E[L_1]$ is not very sensitive to λ_2, but $E[L_2]$ is much higher than $\frac{\lambda_2}{\mu_2-\lambda_2}$, the mean queue length for Q_2 in isolation – in particular when $N = 10$.

Table 3: The impact of λ_2; $\lambda_1 = 2$, $\mu_1 = 3$, $\mu_2 = 4$ and $N = 3$

λ_2	$P(0,0)$	$E[B]$	$E[L_1]$	$E[L_2]$	$E[W_1]$	$E[W_2]$
0.01	0.5975	0.3351	0.4000	0.0055	0.3333	0.5523
0.1	0.5750	0.3520	0.4000	0.0573	0.3334	0.5727
0.5	0.4763	0.4398	0.4020	0.3302	0.3361	0.6604
1	0.3602	0.5922	0.4152	0.7720	0.3551	0.7720
2	0.1724	1.2001	0.5086	2.2674	0.5175	1.1337
2.5	0.1058	1.8775	0.5962	3.9235	0.7384	1.5694
3	0.0568	3.3241	0.7101	7.7872	1.2249	2.5957
3.5	0.0225	7.8831	0.8463	21.2354	2.7534	6.0673
3.75	0.0100	17.1894	0.9213	50.0158	5.8512	13.3376

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

λ_2	$P(0,0)$	$E[B]$	$E[L_1]$	$E[L_2]$	$E[W_1]$	$E[W_2]$
0.01	0.5875	0.3351	0.4000	0.0055	0.3333	0.5523
0.1	0.5750	0.3520	0.4000	0.0574	0.3333	0.5739
0.5	0.4750	0.4421	0.4000	0.3474	0.3333	0.6947
1	0.3501	0.6188	0.4001	0.9405	0.3335	0.9405
2	0.1188	1.8537	0.4283	4.4829	0.3745	2.2415
2.5	0.0467	4.5314	0.5076	9.4443	0.5155	3.7777
3	0.0139	14.2280	0.6458	21.0619	0.9116	7.0206
3.5	0.0029	62.7323	0.8168	60.6934	2.2298	17.3410
3.75	$9.3201e^{-4}$	186.4251	0.9076	144.5221	4.9141	38.5392

Table 5: The impact of μ_1; $\lambda_1 = 2$, $\lambda_2 = 3$, $\mu_2 = 4$ and $N = 3$

μ_1	$P(0,0)$	$E[B]$	$E[L_1]$	$E[L_2]$	$E[W_1]$	$E[W_2]$
0.01	$1.5857e^{-6}$	$1.2612e^{5}$	0.9988	980.3697	399.5025	326.7899
0.1	$1.5444e^{-4}$	$1.2948e^{3}$	0.9875	102.8234	39.5247	34.2745
0.5	0.0034	58.3826	0.9384	24.6374	7.6108	8.2125
1	0.0116	16.9876	0.8808	14.6992	3.6953	4.8997
2	0.0339	5.7004	0.7839	9.5659	1.8137	3.1886
4	0.0771	2.3931	0.6543	6.8813	0.9462	2.2938
10	0.1490	1.1419	0.4952	5.2618	0.4905	1.7539
100	0.2369	0.6442	0.3450	4.4017	0.2634	1.4672

Table 6: The impact of μ_1; $\lambda_1 = 2$, $\lambda_2 = 3$, $\mu_2 = 4$ and $N = 10$

μ_1	$P(0,0)$	$E[B]$	$E[L_1]$	$E[L_2]$	$E[W_1]$	$E[W_2]$
0.01	$2.2342e^{-8}$	$8.9519e^{6}$	0.9988	$1.0031e^{3}$	399.5000	334.3639
0.1	$2.5417e^{-6}$	$7.8688e^{4}$	0.9875	125.2911	39.5004	41.7637
0.5	$1.0404e^{-4}$	$1.9222e^{3}$	0.9375	45.9242	7.5003	15.3081
1	$6.6236e^{-4}$	301.6296	0.8753	34.4077	3.5106	11.4692
2	0.0047	42.2918	0.7547	25.9629	1.5384	8.6543
4	0.0271	7.1876	0.5541	17.5398	0.6214	5.8466
10	0.1067	1.6744	0.2835	8.9742	0.1978	2.9914
100	0.2314	0.6644	0.0684	4.4448	0.0367	1.4816

Table 7: The impact of μ_2; $\lambda_1 = 2$, $\lambda_2 = 3$, $\mu_1 = 3$ and $N = 3$

μ_2	$P(0,0)$	$E[B]$	$E[L_1]$	$E[L_2]$	$E[W_1]$	$E[W_2]$
3.25	0.0132	14.9184	0.9045	36.3961	4.7333	12.1320
3.5	0.0273	7.1333	0.8266	16.9795	2.3839	5.6598
3.75	0.0419	4.5779	0.7628	10.7583	1.6078	3.5861
4	0.0568	3.3241	0.7101	7.7872	1.2249	2.5957
10	0.3157	0.4336	0.4235	1.3487	0.3673	0.4496
100	0.5701	0.1508	0.4001	0.4685	0.3334	0.1562
1000	0.5970	0.1350	0.4000	0.4066	0.3333	0.1355

Table 8: The impact of μ_2; $\lambda_1 = 2$; $\lambda_2 = 3$, $\mu_1 = 3$ and $N = 10$

μ_2	$P(0,0)$	$E[B]$	$E[L_1]$	$E[L_2]$	$E[W_1]$	$E[W_2]$
3.25	0.0012	166.3675	0.8864	110.1147	3.9020	36.7049
3.5	0.0036	54.9727	0.7912	49.7064	1.8941	16.5688
3.75	0.0077	25.6834	0.7116	30.3284	1.2336	10.1095
4	0.0139	14.2288	0.6458	21.0619	0.9116	7.0206
10	0.3010	0.4645	0.4015	1.7561	0.3354	0.5854
100	0.5700	0.1509	0.4000	0.4736	0.3333	0.1579
1000	0.5970	0.1350	0.4000	0.4070	0.3333	0.1357

In Tables 5 and 6 (focusing on the impact of μ_1), very small values of μ_1 lead to $E[L_1]$ being very close to one, and huge values of $E[L_2]$, $E[W_1]$ and $E[W_2]$. The mean busy period also is very large, even for $\mu_1 = 1$.

In Tables 7 and 8 (focusing on the impact of μ_2), small values of μ_2 lead to very small values of $P(0,0)$ and very large values of $E[B]$, in particular for $N = 10$. Again we observe that $E[L_2]$ and $E[W_2]$ are quite sensitive to the presence of class 1, in particular for $N = 10$. Further notice that $E[W_1]$ basically just is a mean service time when μ_2 is very large, while $E[L_1] = 1 - P(0,0) - \sum_{n=1}^{\infty} P(0,n,2) \approx 1 - P(0,0)$.

7 The non-preemptive case

In this section we briefly consider the same model, but with *non-preemptive* switching from Q_2 to Q_1. In Figure 2 we have displayed the state-transition diagram for this case, for $N = 2, 3, 4, 5$. A comparison with Figure 1 for the preemptive case shows that both diagrams are very similar. Notice that in this non-preemptive case the states $(1, n, 2)$ with $n < N$ do exist, whereas they don't in the preemptive case. A switch from $(1, n, 2)$ to $(1, n-1, 1)$ occurs for $n \leq N$ at a service completion in Q_2. The balance equations still hold, apart from a few trivial changes. Furthermore, the balance equations for the states $(1, n, 2)$ with $1 \leq n < N$ are:

$$(\lambda_2 + \mu_2)P(1,n,2) \;=\; \lambda_1 P(0,n,2) + \lambda_2 P(1,n-1,2)I(n \geq 2). \qquad (53)$$

In view of the strong similarities between both model variants, the analysis of queue lengths proceeds very similarly as in the preemptive case. Yet, in this section we sketch how the analysis of Sections 4 and 5 needs to be adapted to obtain the busy period and sojourn time distributions in this non-preemptive case. Analysis of the queue length distribution for this model variant is provided in [15].

The busy period. Very little changes in Section 4. Equations (21)-(31), Equation (33) and Lemmas 1-3 remain valid. In Equation (32) we have to replace the last term in the righthand side by $\frac{\lambda_1}{\lambda_1+\lambda_2+\mu_2+\alpha}\phi_{1,k,2}(\alpha)$. Furthermore, we get $N-1$ extra equations for the new states $\phi_{1,k,2}(\alpha)$: For $k = 1, \ldots, N-1$,

$$\begin{aligned} \phi_{1,k,2}(\alpha) \quad &= \quad \frac{\mu_2}{\lambda_2+\mu_2+\alpha}\phi_{1,k-1,2}(\alpha)I(k \geq 2) \qquad (54) \\ &+ \quad \frac{\lambda_2}{\lambda_2+\mu_2+\alpha}\phi_{1,k-1,1}(\alpha)I(k=1) + \frac{\mu_2}{\lambda_2+\mu_2+\alpha}\phi_{1,k+1,2}(\alpha). \end{aligned}$$

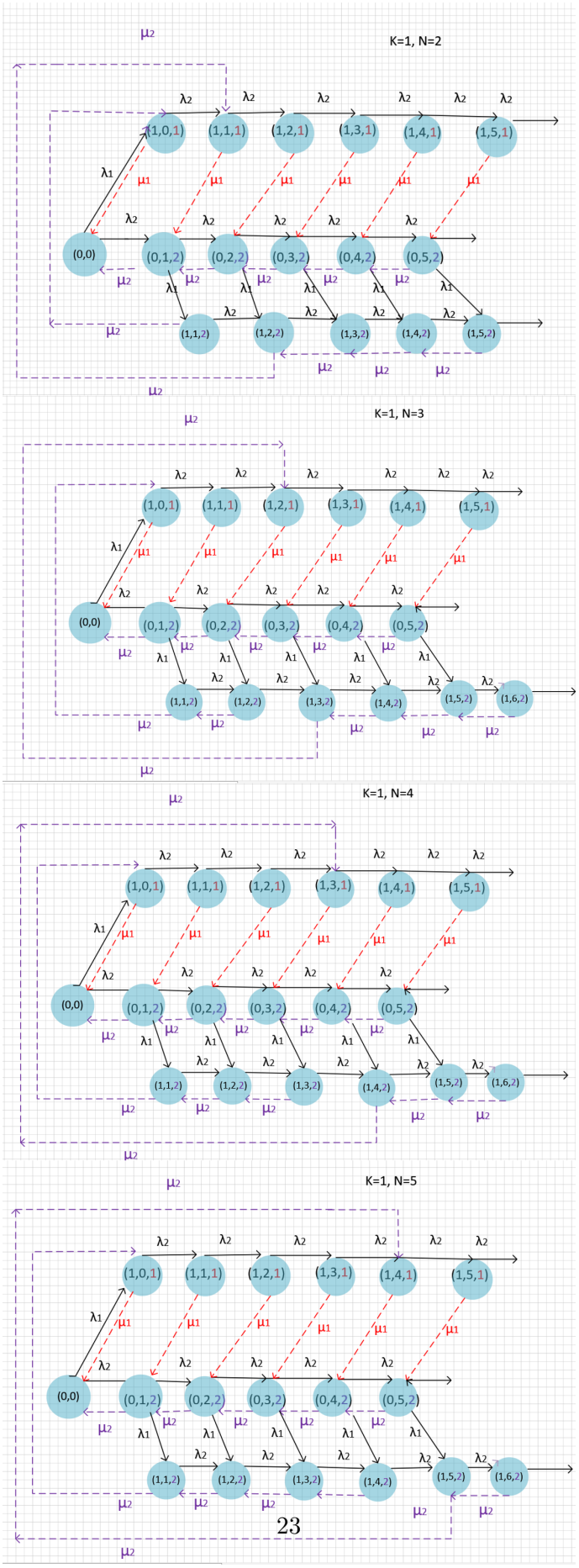

Figure 2: State-transition diagram, for $N = 2, 3, 4, 5$ (non-preemptive).

These $\phi_{1,k,2}(\alpha)$ terms are readily expressed in terms of $\phi_{1,0,1}(\alpha)$.

The sojourn times. Here the changes are more substantial. Firstly, the term $\frac{\mu_1}{\mu_1+\alpha}\sum_{k=1}^{N-1} P(0,k,2)$ has to be adapted. If a type-1 customer arrives and sees state $(0,k,2)$ with $k<N$, then she first waits until the service is completed; and thereafter she is taken into service if the number of type-2 arrivals is less than $N-k$. Otherwise she has to wait until that number again drops below N. The former option gives for each $P(0,k,2)$ term with $1\leq k\leq N-1$ a contribution

$$\begin{aligned} & \frac{\mu_1}{\mu_1+\alpha}\int_{t=0}^{\infty} \mathrm{e}^{-\alpha t}\mu \mathrm{e}^{-\mu t}\sum_{j=0}^{N-k}\frac{(\lambda_2 t)^j}{j!}\mathrm{e}^{-\lambda_2 t}\mathrm{d}t \\ = \quad & \frac{\mu_1}{\mu_1+\alpha}\sum_{j=0}^{N-k}\frac{\mu_2}{\lambda_2+\mu_2+\alpha}(\frac{\lambda_2}{\lambda_2+\mu_2+\alpha})^j \\ = \quad & \frac{\mu_1}{\mu_1+\alpha}\frac{\mu_2}{\mu_2+\alpha}[1-(\frac{\lambda_2}{\lambda_2+\mu_2+\alpha})^{N-k+1}]. \end{aligned} \tag{55}$$

The second option gives for each $P(0,k,2)$ term with $1\leq k\leq N-1$ a contribution

$$\begin{aligned} & \frac{\mu_1}{\mu_1+\alpha}\int_{t=0}^{\infty} \mathrm{e}^{-\alpha t}\mu_2 \mathrm{e}^{-\mu_2 t}\sum_{j=N-k+1}\frac{(\lambda_2 t)^j}{j!}\mathrm{e}^{-\lambda_2 t}\mathrm{d}t\Gamma^{j-(N-k)}(\alpha) \\ = \quad & \frac{\mu_1}{\mu_1+\alpha}\sum_{j=N-k+1}^{\infty}\frac{\mu_2}{\lambda_2+\mu_2+\alpha}(\frac{\lambda_2}{\lambda_2+\mu_2+\alpha})^j\ \Gamma^{k+j-N}(\alpha) \\ = \quad & \frac{\mu_1}{\mu_1+\alpha}\Gamma(\alpha)\frac{\mu_2}{\lambda_2(1-\Gamma(\alpha))+\mu_2+\alpha}(\frac{\lambda_2}{\lambda_2+\mu_2+\alpha})^{N-k+1}. \end{aligned} \tag{56}$$

In Equation (35) we need to add the new terms $\sum_{j=1}^{N-1} P(1,j,2)f_{1,j+1,2}(\alpha)$, where

$$\begin{aligned} f_{1,j+1,2;m}(\alpha) \quad &= \quad \frac{\lambda_2}{\lambda_2+\mu_2+\alpha}f_{1,j+1,2;m+1}(\alpha)+\frac{\mu_2}{\lambda_2+\mu_2+\alpha}f_{1,j,2;m}(\alpha), \quad j+m\geq N-1, \\ &= \quad \frac{\lambda_2}{\lambda_2+\mu_2+\alpha}f_{1,j+1,2;m+1}(\alpha)+\frac{\mu_2}{\lambda_2+\mu_2+\alpha}f_{1,j,1;m}(\alpha), \quad j+m< N-1. \end{aligned} \tag{57}$$

Equations (36)-(43) do not change, except for the last term in the righthand side of (39), which changes into $\frac{\lambda_1}{\lambda_1+\lambda_2+\mu_2+\alpha}f_{1,j+1,2;m}(\alpha)$, because the arrival of a type-1 customer does not lead to immediate preemption of the ongoing 2-service.

8 Conclusion and suggestions for further research

In this paper we have studied a single-server two-queue Markovian polling system with preemptive and with non-preemptive threshold switching. We have derived explicit expressions for the joint queue length distribution; we have analyzed the busy period distribution employing an original approach that uses taboo states; and we have determined the sojourn time distribution for customers in both queues.

Regarding further research, firstly, one could try to determine busy period and sojourn time distributions in case Q_1 has $K > 1$ buffer spaces; the queue lengths in this model were studied in [16]. Secondly, it might be possible to still analyze the model when the service time distribution at Q_1 is general – although it is then probably very challenging to determine, e.g., the sojourn time distribution of type-2 customers. Finally, the results of the present study might be used for optimization purposes, e.g. when holding and switching costs are involved, and when a reward is incurred for each served type-1 customer.

Acknowledgment The research of Onno Boxma is supported by the NWO Gravitation Programme NETWORKS (Grant number 024.002.003). The research of David Perry is partly supported by a grant of the Israel Science Foundation (Grant number 3274/19). The research of Uri Yechiali is partly supported by a grant of the Israel Science Foundation (Grant number 1968/23).

References

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

[2] F. Avram and A. Gómez-Corral (2006). On the optiml control of a two-queue polling model. *Oper. Res. Letters* **34**, 339-348.

[3] M.A.A. Boon, R.D. van der Mei and E.M.M. Winands (2011). Applications of polling systems. *Surv. Oper. Res. Manag. Sci.* **16**, 67-82.

[4] S.C. Borst and O.J. Boxma (2018). Polling: Past, present and perspective. *TOP* **26**, 335-369.

[5] O.J. Boxma and D.G. Down (1997). Dynamic server assignment in a two-queue model. *Eur. J. Oper. Res.* **103**, 595-609.

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

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

[8] S. Browne and U. Yechiali (1989). Dynamic priority rules for cyclic-type queues. *Adv. Appl. Probab.* **21**, 432-450.

[9] J.W. Cohen (1982). *The Single Server Queue.* North-Holland Publ. Cy., Amsterdam.

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

[11] N.K. Jaiswal (1968). *Priority Queues.* Academic Press, New York.

[12] A. Jolles, E. Perel and U. Yechiali (2018). Alternating server with non-zero switch-over times and opposite-queue threshold-based switching policy. *Performance Evaluation* **126**, 22-38.

[13] D.-S. Lee (1996). A two-queue model with exhaustive and limited service disciplines. *Stochastic Models* **12**, 285-305.

[14] D.-S. Lee and B. Sengupta (1993). Queueing analysis of a threshold based priority scheme for ATM networks. *IEEE/ACM Trans. Networking* **1**, 709-717.

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

[16] E. Perel and U. Yechiali (2017). Two-queue polling systems with switching policy based on the queue that is not being served. *Stochastic Models* **33**, 430-450.

[17] H. Takagi (1991). *Queueing Analysis, Vol. 1: Vacation and Priority Systems*. Elsevier Science Publishers, Amsterdam.

[18] S.F. Yashkov (1983). A derivation of response time distribution for a $M/G/1$ processor sharing queue. *Problems Control Inf. Theory* **12**, 133-148.

[19] U. Yechiali (1993). Analysis and control of polling systems. In: *Performance Evaluation of Computer and Communication Systems*, eds. L. Donatiello and R. Nelson, Springer, Berlin, pp. 630-650.