The final publication is available at Springer via http://dx.doi.org/10.1007/s10479-018-2915-y

On a 2-class polling model with reneging and k_i-limited service

Kevin Granville*† Steve Drekic*‡

June 2018

Abstract This paper analyzes a 2-class, single-server polling model operating under a k_i-limited service discipline with class-dependent switchover times. Arrivals to each class are assumed to follow a Poisson process with phase-type distributed service times. Within each queue, customers are impatient and renege (i.e., abandon the queue) if the time before entry into service exceeds an exponentially distributed patience time. We model the queueing system as a level-dependent quasi-birth-and-death process, and the steady-state joint queue length distribution as well as the per-class waiting time distributions are computed via the use of matrix analytic techniques. The impacts of reneging and choice of service time distribution are investigated through a series of numerical experiments, with a particular focus on the determination of (k_1, k_2) which minimizes a cost function involving the expected time a customer spends waiting in the queue and an additional penalty cost should reneging take place.

Keywords Polling model · k_i-limited service discipline · Reneging · Quasi-birth-and-death process · Switchover times · Phase-type distribution

1 Introduction

A typical polling model consists of multiple queues attended by a single server in cyclic order. Due to its wide use in the areas of public health systems, transportation, and communication and computer networks, it has drawn considerable attention over the past fifty years. As a case in point, Levy and Sidi (1990) focused on polling model applications in various fields of operations research. An early survey conducted by Takagi (1988) summarized the important criteria needed to characterize a polling model in a queueing context, including a description of different possible service disciplines encountered in practice. Vishnevskii and Semenova

*Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, ON, Canada
†kgranville@uwaterloo.ca
‡sdrekic@uwaterloo.ca (✉)

(2006) provided an extensive review of the existing literature on polling models. In particular, new approaches and more general models, including several related optimization problems, were outlined and discussed in their review. For more information concerning recent efforts and current progress in the research of polling systems, the interested reader is directed to Boon (2011) and Boon et al. (2011).

Most of the polling model literature focuses on the determination of two fundamental performance measures – namely, the queue length distribution and the waiting time distribution for each queue. Explicit, closed-form expressions for the associated distribution functions are generally not obtainable, due mainly to the mathematical complexity inherent in the analysis of these kinds of queueing models. While much attention has been paid to determining the Laplace-Stieltjes transform of the waiting time distribution (e.g., see Winands et al. 2009), some authors have employed Markov chains (e.g., see MacPhee et al. 2007) and/or matrix analytic techniques (e.g., see Mishkoy et al. 2012; Perel and Yechiali 2017) as a tool to construct and analyze a wide class of polling models. Other authors (e.g., see van Vuuren and Winands 2007) have taken an alternate route and developed approximation procedures to accurately compute these distribution functions.

In this paper, we utilize the concept of absorption times within a continuous-time Markov chain setting to obtain, in an algorithmically tractable fashion, the distribution function of the waiting time in a single-server polling system consisting of two queues operating under a k_i-limited service discipline with Poisson arrivals and phase-type distributed service times. A great deal of work has been done in analyzing polling models with the k_i-limited service discipline (e.g., see Chang and Down 2002; Boon and Winands 2014), in which the server, when visiting class i, serves (in a non-idling fashion) up to k_i class-i customers before switching over to another class. In addition, we model server switchover times and the possibility of customers reneging from their queues. Queues with reneging (or abandonment) occur in many facets of everyday life, most notably in situations encountered in manufacturing systems of perishable goods (e.g., see Graves 1982) and telecommunication systems (e.g., see Gromoll et al. 2006). Although numerous papers in the polling model literature have incorporated the notion of switchover times (e.g., see Boon et al. 2010 and references therein), it is somewhat surprising that there is a dearth of results for reneging in polling models. In fact, we are aware of only a few papers (i.e., Vishnevskii and Semenova 2008, 2009; Boon 2012) which have integrated this feature within a polling model framework.

The rest of the paper is organized as follows. Section 2 introduces notation and provides a detailed description of the polling model. In Section 3, we build the fundamental components underlying our Markov chain based approach, which turns out to involve a level-dependent quasi-birth-and-death (QBD) process, in order to develop an algorithm for computing the steady-state joint queue length distribution. In Section 4, we modify the Markov chain structure of Section 3 to construct a phase-type framework which inevitably yields an explicit formula for the per-class waiting time distribution function and moments. In Section 5, we apply our results and provide a detailed numerical analysis of a defined cost function depending on the expected time waiting in system, probability of reneging, and the arrival rates of both classes of customers. Optimal combinations of k_1 and k_2 are found which

minimize the cost function for a pair of scenarios over a range of reneging rates and service time distributions, under a constraint which limits the maximum number of services in a single cycle. By way of varying the cost parameters and contrasting the two scenarios, several observations are made and discussed. We end the paper with some concluding remarks in Section 6.

2 Model Description

We consider a polling model in which a single server provides service to two distinct classes of customers, each having its own respective queue. Customers are served on a first-come, first-served basis within their own queue. Let $b_i < \infty$ be the class-i buffer size, $i = 1, 2$. Customers of classes 1 and 2 arrive to the system according to independent Poisson processes with rates λ_1 and λ_2, respectively. Service times for class-i customers, $i = 1, 2$, are assumed to be distributed as $PH(\underline{\beta}_i, S_i)$, referring to a phase-type distribution with rate matrix S_i (of dimension $s_i \times s_i$) and initial probability row vector $\underline{\beta}_i = (\beta_{i,1}, \beta_{i,2}, \ldots, \beta_{i,s_i})$, where $\sum_{j=1}^{s_i} \beta_{i,j} = 1$. For $i = 1, 2$, let μ_i denote the mean class-i service time. We assume that a customer's service time is independent of all other service times as well as the arrival processes.

Service is administered according to the k_i-limited service discipline, in which the server serves up to k_i customers of class i, switching over to the other class once the class-i queue empties or the maximum number of services has been reached. Note that by letting $k_i \to \infty$ for $i = 1, 2$, it is possible to model a 2-class polling model with *exhaustive* service. Moreover, we can capture the non-preemptive priority service discipline (with switchovers) by letting $k_1 \to \infty$ and $k_2 = 1$ (i.e., class 1 has higher priority over class 2). Once the decision to switch out of class i has been made, the server initiates an exponentially distributed switchover time with rate v_i, which must complete before service can begin on the other class. Switchover times are independent of each other, and independent of service times and the arrival processes. Furthermore, we assume that the server is unable to determine whether the other class is empty before initiating a switchover, so it is possible for multiple switchovers to take place before the server finally encounters a customer waiting to be served. As a result, the server is never truly idle in the system, even when both queues are empty.

We also incorporate the notion of class-dependent reneging and assume that when an arriving class-i customer enters the system, it leaves the system following an (independent) exponentially distributed amount of time with rate α_i and is subsequently lost. Once a customer does reach the server, however, we assume that customer is no longer subject to reneging. A graphical illustration of the polling model is given in Fig. 1, in which we designate the colours red and blue to represent characteristics of classes 1 and 2, respectively. The left (right) half of the figure depicts the system during a sojourn of the server to the class-1 (class-2) queue. The solid red and blue boxes denote customers who are either waiting, or in service, and the empty red and blue boxes represent open slots in either queue available to future arriving customers.

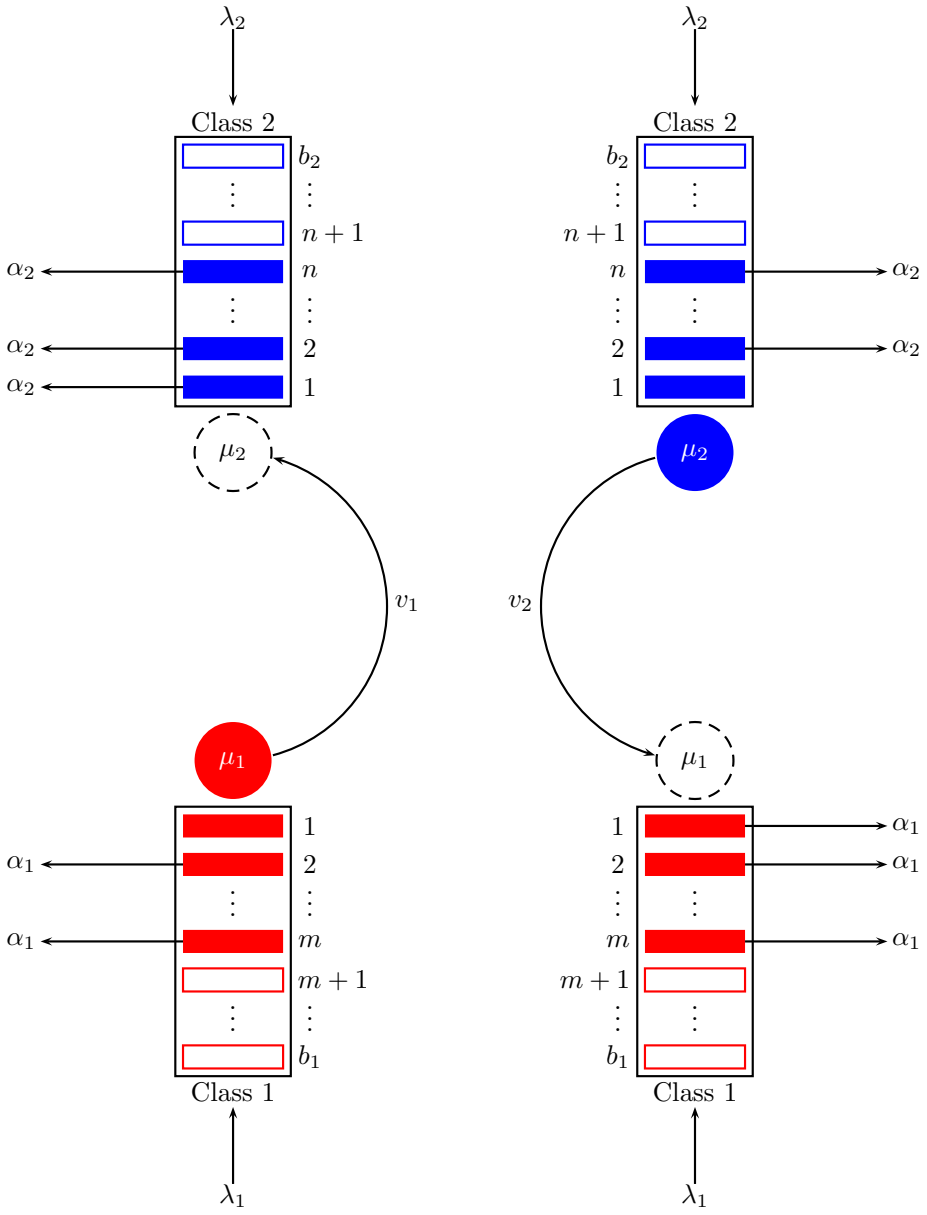

Figure 1: Depiction of the polling model during a sojourn of the server to either queue.

3 Determination of the Steady-state Probabilities

For $i = 1, 2$, let X_i represent the number of class-i customers present in the system (at stationarity), so that $0 \leq X_i \leq b_i$. Our first objective is to determine $P_{m,n}$, the steady-state joint probability that $X_1 = m$ and $X_2 = n$ for $m = 0, 1, \ldots, b_1$ and $n = 0, 1, \ldots, b_2$. To aid in this regard, we define an associated quantity $\pi_{m,n,l,y}$ representing the steady-state joint probability that $X_1 = m$, $X_2 = n$, the server occupies *position* l, and the current phase of

service is y (with $y = 0$ indicating that the system is in switchover mode). In particular, the possible values of l depend on the corresponding values of m and n in the following manner:

$$\begin{array}{lcll} m = 0 \text{ and } n = 0 & \Longrightarrow & l = & k_1 + k_2 + 1, k_1 + k_2 + 2, \\ m \neq 0 \text{ and } n = 0 & \Longrightarrow & l = & 1, 2, \ldots, k_1, k_1 + k_2 + 1, k_1 + k_2 + 2, \\ m = 0 \text{ and } n \neq 0 & \Longrightarrow & l = & k_1 + 1, k_1 + 2, \ldots, k_1 + k_2, k_1 + k_2 + 1, \\ & & & k_1 + k_2 + 2, \\ m \neq 0 \text{ and } n \neq 0 & \Longrightarrow & l = & 1, 2, \ldots, k_1, k_1 + 1, k_1 + 2, \ldots, k_1 + k_2, \\ & & & k_1 + k_2 + 1, k_1 + k_2 + 2. \end{array} \tag{1}$$

In other words, when $l = 1, 2, \ldots, k_1$, the server is known to be serving its l^{th} customer from the class-1 queue. On the other hand, when $l = k_1+1, k_1+2, \ldots, k_1+k_2$, the server is known to be serving its $(l - k_1)^{th}$ customer from the class-2 queue. Also, $l = k_1 + k_2 + i$, $i = 1, 2$, indicates that the server is conducting a switchover out of the class-i queue. Similarly, the possible values of y depend on l as follows:

$$\begin{array}{lcll} l = & 1, 2, \ldots, k_1 & \Longrightarrow & y = 1, 2, \ldots, s_1, \\ l = & k_1 + 1, k_1 + 2, \ldots, k_1 + k_2 & \Longrightarrow & y = 1, 2, \ldots, s_2, \\ l = & k_1 + k_2 + 1, k_1 + k_2 + 2 & \Longrightarrow & y = 0. \end{array} \tag{2}$$

Note that in the case when $m = n = 0$ (i.e., the queue is empty), the system can only be in one of the two possible switchover modes (as there are no customers to serve in either queue), and so $P_{0,0} = \pi_{0,0,k_1+k_2+1,0} + \pi_{0,0,k_1+k_2+2,0}$. Furthermore, it is an immediate consequence that

$$P_{0,n} = \sum_{l=k_1+1}^{k_1+k_2} \sum_{y=1}^{s_2} \pi_{0,n,l,y} + \sum_{l=k_1+k_2+1}^{k_1+k_2+2} \pi_{0,n,l,0} \ , \ n \geq 1,$$

$$P_{m,0} = \sum_{l=1}^{k_1} \sum_{y=1}^{s_1} \pi_{m,0,l,y} + \sum_{l=k_1+k_2+1}^{k_1+k_2+2} \pi_{m,0,l,0} \ , \ m \geq 1,$$

and

$$P_{m,n} = \sum_{l=1}^{k_1} \sum_{y=1}^{s_1} \pi_{m,n,l,y} + \sum_{l=k_1+1}^{k_1+k_2} \sum_{y=1}^{s_2} \pi_{m,n,l,y} + \sum_{l=k_1+k_2+1}^{k_1+k_2+2} \pi_{m,n,l,0} \ , \ m, n \geq 1.$$

We define the 0^{th} steady-state probability row vector to be

$$\underline{\pi}_0 = (\underline{\pi}_{0,0}, \underline{\pi}_{0,1}, \ldots, \underline{\pi}_{0,b_2}),$$

where $\underline{\pi}_{0,0} = (\pi_{0,0,k_1+k_2+1,0}, \pi_{0,0,k_1+k_2+2,0})$ and

$$\begin{aligned} \underline{\pi}_{0,n} = (&\pi_{0,n,k_1+1,1}, \ldots, \pi_{0,n,k_1+1,s_2}, \pi_{0,n,k_1+2,1}, \ldots, \pi_{0,n,k_1+k_2,s_2}, \\ &\pi_{0,n,k_1+k_2+1,0}, \pi_{0,n,k_1+k_2+2,0}) \end{aligned}$$

is a row vector of size $z_1 = k_2 s_2 + 2$ for $n = 1, 2, \ldots, b_2$. For $m \geq 1$, the m^{th} steady-state probability row vector is defined as

$$\underline{\pi}_m = (\underline{\pi}_{m,0}, \underline{\pi}_{m,1}, \ldots, \underline{\pi}_{m,b_2}),$$

where

$$\underline{\pi}_{m,0} = (\pi_{m,0,1,1}, \ldots, \pi_{m,0,1,s_1}, \pi_{m,0,2,1}, \ldots, \pi_{m,0,k_1,s_1}, \pi_{m,0,k_1+k_2+1,0}, \pi_{m,0,k_1+k_2+2,0})$$

is a row vector of size $k_1 s_1 + 2$ and

$$\underline{\pi}_{m,n} = (\pi_{m,n,1,1}, \ldots, \pi_{m,n,1,s_1}, \pi_{m,n,2,1}, \ldots, \pi_{m,n,k_1,s_1}, \pi_{m,n,k_1+1,1}, \ldots, \pi_{m,n,k_1+1,s_2}, \pi_{m,n,k_1+2,1}, \ldots, \pi_{m,n,k_1+k_2,s_2}, \pi_{m,n,k_1+k_2+1,0}, \pi_{m,n,k_1+k_2+2,0})$$

is a row vector of size $z_2 = k_1 s_1 + z_1$ for $n = 1, 2, \ldots, b_2$. Referring to X_1 as the *level* of the process, we remark that level 0 is comprised of $n_1 = b_2 z_1 + 2$ sub-levels, whereas each non-zero level consists of a total of $n_2 = b_2 z_2 + k_1 s_1 + 2$ sub-levels.

Let $\underline{\pi} = (\underline{\pi}_0, \underline{\pi}_1, \ldots, \underline{\pi}_{b_1})$ be the concatenated steady-state probability row vector having a total of $b_1 + 1$ levels. To determine $\underline{\pi}_m$ for $m \geq 0$, we need to solve $\tilde{\underline{0}} = \underline{\pi} Q$ where Q is the infinitesimal generator of the process and $\tilde{\underline{0}} = (\underline{0}_{n_1}, \underline{0}_{n_2}, \ldots, \underline{0}_{n_2})$ is an appropriately partitioned row vector with a total of $b_1 + 1$ levels (throughout the rest of the paper, $\underline{0}_i$ denotes a $1 \times i$ row vector of zeros). We note that Q is block-structured with blocks $Q_{i,j}$ containing all transitions where X_1 changes from i to j. Due to the presence of reneging in our model, we end up with a level-dependent QBD process having infinitesimal generator of the form

$$Q = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ \vdots \\ b_1 - 2 \\ b_1 - 1 \\ b_1 \end{array} \begin{array}{c} \begin{array}{ccccccc} 0 & 1 & 2 & \cdots & b_1 - 2 & b_1 - 1 & b_1 \end{array} \\ \begin{pmatrix} Q_{0,0} & Q_{0,1} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ Q_{1,0} & Q_{1,1} & Q_{1,2} & \ddots & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & Q_{2,1} & Q_{2,2} & \ddots & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \vdots & \ddots & \ddots & \ddots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & Q_{b_1-2,b_1-2} & Q_{b_1-2,b_1-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & Q_{b_1-1,b_1-2} & Q_{b_1-1,b_1-1} & Q_{b_1-1,b_1} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & Q_{b_1,b_1-1} & Q_{b_1,b_1} \end{pmatrix} \end{array}. \quad (3)$$

In (3), $\mathbf{0}$ denotes an appropriately dimensioned zero matrix. The overall dimension of Q is $n_1 + b_1 n_2$, as $Q_{0,0}$ is an $n_1 \times n_1$ sub-matrix, $Q_{0,1}$ is an $n_1 \times n_2$ sub-matrix, $Q_{1,0}$ is an $n_2 \times n_1$ sub-matrix, and all remaining sub-matrices are of size $n_2 \times n_2$.

We first observe that $Q_{1,2} = Q_{2,3} = \cdots = Q_{b_1-1,b_1} = \lambda_1 I_{n_2}$, where, in general, I_i denotes an $i \times i$ identity matrix. In what follows, let $\otimes$ denote the Kronecker product operator and

let $\delta_{i,j}$ be the standard Kronecker delta function. Also, let $\underline{e}_{i,j}$ be a row vector of length i with 1 as the j^{th} entry and zeros everywhere else, and let $\underline{e}_i$ be a row vector of i ones. In addition, let $\underline{S}'_{0,i} = -S_i\underline{e}'_{s_i}$, where the prime symbol, $'$, denotes vector transpose. Finally, for further notational convenience, define $\lambda = \lambda_1 + \lambda_2$, $\underline{v} = (v_1, v_2)$, $V = \text{diag}(\underline{v})$, $V_1 = v_1\underline{e}'_{2,1}\underline{e}_{2,2}$, and $V_2 = v_2\underline{e}'_{2,2}\underline{e}_{2,1}$. Based on this notation, the diagonal components of Q can be expressed as

$$Q_{0,0} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ \vdots \\ b_2-1 \\ b_2 \end{array} \begin{array}{c} \begin{array}{ccccccc} 0 & & 1 & 2 & \dots & b_2-1 & b_2 \end{array} \\ \begin{pmatrix} -(\lambda I_2 + V - V_1 - V_2) & \begin{bmatrix} \mathbf{0} & \lambda_2 I_2 \end{bmatrix} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} \\ \begin{bmatrix} \underline{e}'_{k_2}\underline{e}_{2,2} \otimes \underline{S}'_{0,2} \\ \alpha_2 I_2 \end{bmatrix} & \Delta_1 & \lambda_2 I_{z_1} & \ddots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \Gamma_2 & \Delta_2 & \ddots & \mathbf{0} & \mathbf{0} \\ \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \Delta_{b_2-1} & \lambda_2 I_{z_1} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \Gamma_{b_2} & \Delta_{b_2} \end{pmatrix} \end{array}$$

and

$$Q_{i,i} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ \vdots \\ b_2-1 \\ b_2 \end{array} \begin{array}{c} \begin{array}{ccccccc} 0 & & 1 & & 2 & \dots & b_2-1 & b_2 \end{array} \\ \begin{pmatrix} C_{i,0} & \begin{bmatrix} [\lambda_2 I_{k_1 s_1} & \underline{0}'_{k_1 s_1}\underline{0}_{k_2 s_2}] & \mathbf{0} \\ \mathbf{0} & & \lambda_2 I_2 \end{bmatrix} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} \\ \begin{bmatrix} \alpha_2 I_{k_1 s_1} & \mathbf{0} \\ \mathbf{0} & \underline{e}'_{k_2}\underline{e}_{2,2} \otimes \underline{S}'_{0,2} \\ \mathbf{0} & \alpha_2 I_2 \end{bmatrix} & C_{i,1} & \lambda_2 I_{z_2} & \ddots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & B_2 & C_{i,2} & \ddots & \mathbf{0} & \mathbf{0} \\ \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & C_{i,b_2-1} & \lambda_2 I_{z_2} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & B_{b_2} & C_{i,b_2} \end{pmatrix} \end{array},$$

for $i = 1, 2, \dots, b_1$, where

$$\Gamma_i = \begin{bmatrix} (i-1)\alpha_2 I_{k_2 s_2} + U_2 & \underline{e}'_{k_2,k_2}\underline{e}_{2,2} \otimes \underline{S}'_{0,2} \\ \mathbf{0} & i\alpha_2 I_2 \end{bmatrix},$$

$$\Delta_j = \begin{bmatrix} -I_{k_2} \otimes ((\lambda - \lambda_2\delta_{j,b_2} + (j-1)\alpha_2)I_{s_2} - S_2) & \mathbf{0} \\ \underline{e}'_{2,1}\underline{e}_{k_2,1} \otimes v_1\underline{\beta}_2 & -((\lambda - \lambda_2\delta_{j,b_2} + j\alpha_2)I_2 + V - V_2) \end{bmatrix},$$

$$B_i = \begin{bmatrix} i\alpha_2 I_{k_1 s_1} & \mathbf{0} \\ \mathbf{0} & \Gamma_i \end{bmatrix},$$

$$U_i = \begin{cases} \mathbf{0} & \text{if } k_i = 1, \\ \begin{bmatrix} \underline{0}'_{k_i-1} & I_{k_i-1} \\ 0 & \underline{0}_{k_i-1} \end{bmatrix} \otimes \underline{S}'_{0,i}\underline{\beta}_i & \text{if } k_i \geq 2, \end{cases}$$

$$C_{i,j} = \begin{cases} \begin{bmatrix} -I_{k_1} \otimes ((\lambda - \lambda_1 \delta_{i,b_1} + (i-1)\alpha_1) I_{s_1} - S_1) & \mathbf{0} \\ \underline{e}'_{2,2} \underline{e}_{k_1,1} \otimes v_2 \underline{\beta}_{-1} & -((\lambda - \lambda_1 \delta_{i,b_1} + i\alpha_1) I_2 + V - V_1) \end{bmatrix} & \text{if } j = 0, \\ \begin{bmatrix} \zeta_{1,i,j} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \zeta_{2,i,j} & \mathbf{0} \\ \underline{e}'_{2,2} \underline{e}_{k_1,1} \otimes v_2 \underline{\beta}_{-1} & \underline{e}'_{2,1} \underline{e}_{k_2,1} \otimes v_1 \underline{\beta}_{-2} & -((\lambda - \lambda_1 \delta_{i,b_1} - \lambda_2 \delta_{j,b_2} + i\alpha_1 + j\alpha_2) I_2 + V) \end{bmatrix} & \text{if } j = 1, 2, \ldots, b_2, \end{cases}$$

and

$$\zeta_{x,i,j} = -I_{k_x} \otimes ((\lambda - \lambda_1 \delta_{i,b_1} - \lambda_2 \delta_{j,b_2} + (i - \delta_{x,1})\alpha_1 + (j - \delta_{x,2})\alpha_2) I_{s_x} - S_x).$$

In addition, we have that

$$Q_{0,1} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ \vdots \\ b_2 \end{array} \begin{array}{c} \begin{array}{ccccc} 0 & 1 & 2 & \cdots & b_2 \end{array} \\ \begin{pmatrix} \begin{bmatrix} \underline{0}'_2 \underline{0}_{k_1 s_1} & \lambda_1 I_2 \end{bmatrix} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \begin{bmatrix} \underline{0}'_{z_1} \underline{0}_{k_1 s_1} & \lambda_1 I_{z_1} \end{bmatrix} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \begin{bmatrix} \underline{0}'_{z_1} \underline{0}_{k_1 s_1} & \lambda_1 I_{z_1} \end{bmatrix} & \ddots & \mathbf{0} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \begin{bmatrix} \underline{0}'_{z_1} \underline{0}_{k_1 s_1} & \lambda_1 I_{z_1} \end{bmatrix} \end{pmatrix} \end{array},$$

$$Q_{1,0} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ \vdots \\ b_2 \end{array} \begin{array}{c} \begin{array}{ccccc} 0 & 1 & 2 & \cdots & b_2 \end{array} \\ \begin{pmatrix} \begin{bmatrix} \underline{e}'_{k_1} \underline{e}_{2,1} \otimes \underline{S}'_{0,1} \\ \alpha_1 I_2 \end{bmatrix} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \begin{bmatrix} \underline{e}'_{k_1} \underline{e}_{z_1,z_1-1} \otimes \underline{S}'_{0,1} \\ \alpha_1 I_{z_1} \end{bmatrix} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \begin{bmatrix} \underline{e}'_{k_1} \underline{e}_{z_1,z_1-1} \otimes \underline{S}'_{0,1} \\ \alpha_1 I_{z_1} \end{bmatrix} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \begin{bmatrix} \underline{e}'_{k_1} \underline{e}_{z_1,z_1-1} \otimes \underline{S}'_{0,1} \\ \alpha_1 I_{z_1} \end{bmatrix} \end{pmatrix} \end{array},$$

and

$$Q_{i,i-1} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ \vdots \\ b_2 - 1 \\ b_2 \end{array} \begin{array}{c} \begin{array}{cccccc} 0 & 1 & 2 & \cdots & b_2 - 1 & b_2 \end{array} \\ \begin{pmatrix} A_{i,0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & A_{i,1} & \mathbf{0} & \ddots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & A_{i,1} & \ddots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & A_{i,1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & A_{i,1} \end{pmatrix} \end{array}, \ i = 2, 3, \ldots, b_1,$$

where

$$A_{i,j} = \begin{bmatrix} (i-1)\alpha_1 I_{k_1 s_1} + U_1 & \underline{e}'_{k_1,k_1} \underline{e}_{k_2 s_2 \delta_{j,1}+2, k_2 s_2 \delta_{j,1}+1} \otimes \underline{S}'_{0,1} \\ \mathbf{0} & i\alpha_1 I_{k_2 s_2 \delta_{j,1}+2} \end{bmatrix}.$$

Level-dependent QBD processes are well-studied in the literature (e.g., see Bright and Taylor 1995) and it is possible to adapt a computational procedure proposed by Gaver et al. (1984) to calculate the steady-state probabilities associated with our model, which we quickly summarize below. First of all, from $\underline{\tilde{0}} = \underline{\pi} Q$, we immediately obtain the equilibrium equations in block form as follows:

$$\underline{0}_{n_1} = \underline{\pi}_0 Q_{0,0} + \underline{\pi}_1 Q_{1,0}, \tag{4}$$

$$\underline{0}_{n_2} = \underline{\pi}_0 Q_{0,1} + \underline{\pi}_1 Q_{1,1} + \underline{\pi}_2 Q_{2,1}, \tag{5}$$

$$\underline{0}_{n_2} = \lambda_1 \underline{\pi}_{m-1} + \underline{\pi}_m Q_{m,m} + \underline{\pi}_{m+1} Q_{m+1,m},\ m = 2, 3, \ldots, b_1 - 1, \tag{6}$$

$$\underline{0}_{n_2} = \lambda_1 \underline{\pi}_{b_1-1} + \underline{\pi}_{b_1} Q_{b_1,b_1}. \tag{7}$$

Solving (5) through (7) in a backward fashion readily yields

$$\underline{\pi}_m = \underline{\pi}_0 \prod_{j=1}^{m} \mathcal{S}_j,\ m = 1, 2, \ldots, b_1, \tag{8}$$

where the set of matrices $\{\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_{b_1}\}$ satisfy the recursive relation

$$\mathcal{S}_j = -\lambda_1 (Q_{j,j} + \mathcal{S}_{j+1} Q_{j+1,j})^{-1},\ j = 2, 3, \ldots, b_1 - 1,$$

with

$$\mathcal{S}_{b_1} = -\lambda_1 Q_{b_1,b_1}^{-1}$$

and

$$\mathcal{S}_1 = -Q_{0,1} (Q_{1,1} + \mathcal{S}_2 Q_{2,1})^{-1}.$$

In addition, if we define $\mathcal{S}_0 = Q_{0,0} + \mathcal{S}_1 Q_{1,0}$, then (4) becomes $\underline{\pi}_0 \mathcal{S}_0 = \underline{0}_{n_1}$. Since all steady-state probabilities must sum to 1, we ultimately end up with the following system of linear equations which must be solved to determine $\underline{\pi}_0$, namely:

$$\underline{\pi}_0 \left[\begin{array}{cc} \mathcal{S}_0 & \underline{u}' \end{array} \right] = (\underline{0}_{n_1}, 1), \tag{9}$$

where

$$\underline{u}' = \underline{e}'_{n_1} + \sum_{m=1}^{b_1} \prod_{j=1}^{m} \mathcal{S}_j \underline{e}'_{n_2}.$$

In (9), $(\underline{0}_{n_1}, 1)$ represents the concatenated row vector of size $n_1 + 1$. Once $\underline{\pi}_0$ is calculated, we then proceed to obtain $\underline{\pi}_m$, $m = 1, 2, \ldots, b_1$, via (8).

With the determination of these steady-state probabilities, we introduce two important quantities of interest associated with this particular queueing system. First of all, $P_{b_1,\bullet} = \sum_{j=0}^{b_2} P_{b_1,j}$ represents the probability that an arbitrarily arriving class-1 customer is turned away at entry (and subsequently lost) due to the class-1 queue being full, and is referred to as the *class-1 blocking probability*. Likewise, the *class-2 blocking probability* is given by $P_{\bullet,b_2} = \sum_{m=0}^{b_1} P_{m,b_2}$, and it represents the probability that an arbitrarily arriving class-2 customer is denied entry to the system due to the class-2 queue being full. We remark that $P_{b_1,\bullet}$ and $P_{\bullet,b_2}$ are particularly useful in helping choose values of b_1 and b_2 so as to ensure negligible blocking probabilities are obtained for both queues.

4 Determination of the Waiting Time Distribution

We derive the steady-state distribution of the random variable W_i, $i = 1, 2$, representing the duration of time from the (non-blocked) arrival of an arbitrary class-i customer to the system until the server is reached. For reasons that will become evident shortly, we refer to W_i as the *nominal* class-i waiting time, as it does not consider the reneging behaviour of this customer. Without loss of generality, we focus our analysis only on W_1 as the characteristics of the two queues are essentially indifferent. In other words, the approach we develop below to obtain the distribution of W_1 can readily be adapted (via a simple relabeling of classes 1 and 2) to obtain the distribution of W_2.

Let us first define the modified steady-state probabilities

$$\phi_{0,0,l,0} = \frac{\pi_{0,0,l,0}}{1 - P_{b_1,\bullet}}$$

and

$$\phi_{m,n,l,y} = \frac{\pi_{m,n,l,y}}{1 - P_{b_1,\bullet}},$$

where $m = 1, 2, \ldots, b_1 - 1$, $n = 1, 2, \ldots, b_2$, and the components l and y are as defined in (1) and (2), respectively. Based on the above definitions, we introduce several row vectors required in the subsequent analysis. Specifically, define

$$\underline{\phi}_{0,n} = \frac{\underline{\pi}_{0,n}}{1 - P_{b_1,\bullet}},\ 1 \leq n \leq b_2\,,$$

$$\underline{\phi}_{m,0} = \frac{\underline{\pi}_{m,0}}{1 - P_{b_1,\bullet}},\ 1 \leq m \leq b_1 - 1\,,$$

and

$$\underline{\phi}_{m,n} = \frac{\underline{\pi}_{m,n}}{1 - P_{b_1,\bullet}},\ 1 \leq m \leq b_1 - 1, 1 \leq n \leq b_2\,.$$

Furthermore, let

$$\underline{\phi}_0 = (\phi_{0,0,k_1+k_2+1,0}, \phi_{0,0,k_1+k_2+2,0}, \underline{\phi}_{0,1}, \underline{\phi}_{0,2}, \ldots, \underline{\phi}_{0,b_2})$$

and

$$\underline{\phi}_m = (\underline{\phi}_{m,0}, \underline{\phi}_{m,1}, \ldots, \underline{\phi}_{m,b_2}),\ m = 1, 2, \ldots, b_1 - 1\,.$$

If we now construct

$$\underline{\Phi} = (\underline{\phi}_{b_1-1}, \underline{\phi}_{b_1-2}, \ldots, \underline{\phi}_1, \underline{\phi}_0) \tag{10}$$

to be the concatenated row vector of dimension

$$\ell = (b_1 - 1)n_2 + n_1, \tag{11}$$

then $\underline{\Phi}\,\underline{e}'_\ell = 1$ due to our earlier observation that, even when both queues are empty, the server is still busy in the midst of completing a switchover (and thus the wait time will be non-zero).

Upon successful entry into a busy system (i.e., to one of the ℓ possible states above), the *PASTA property* (e.g., see Tijms 2003, Theorem 2.4.1) ensures that our target Poisson-arriving class-1 customer finds the system in state (m, n, l, y) with probability $\phi_{m,n,l,y}$. For the moment, we assume that our target class-1 customer is not subject to reneging (later on, we will incorporate the reneging behaviour of this specific customer back into the problem). While waiting in the class-1 queue, the number of customers in the class-2 queue potentially changes, not to mention the service indicator component used to identify how many customers have completed service within the active serving cycle. On the other hand, as the number of customers in the class-1 queue changes, the ones arriving later have no impact on the waiting time of the target class-1 customer. Therefore, if we effectively think of the arrival rate for the class-1 queue to be equal to 0, the distribution of W_1 can in fact be modeled as the distribution of the time to absorption in a Markov chain with infinitesimal generator of the form

$$\begin{bmatrix} \mathcal{R} & -\mathcal{R}\underline{e}'_{\ell} \\ \underline{0}_{\ell} & 0 \end{bmatrix},$$

where

$$\mathcal{R} = \begin{array}{c} \\ b_1-1 \\ b_1-2 \\ b_1-3 \\ \vdots \\ 2 \\ 1 \\ 0 \end{array} \begin{array}{c} \begin{array}{ccccccc} b_1-1 & b_1-2 & b_1-3 & \cdots & 2 & 1 & 0 \end{array} \\ \begin{pmatrix} \widetilde{Q}_{b_1-1,b_1-1} & Q_{b_1-1,b_1-2} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \widetilde{Q}_{b_1-2,b_1-2} & Q_{b_1-2,b_1-3} & \ddots & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \widetilde{Q}_{b_1-3,b_1-3} & \ddots & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \widetilde{Q}_{2,2} & Q_{2,1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \widetilde{Q}_{1,1} & \widetilde{Q}_{1,0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & \widetilde{Q}_{0,0} \end{pmatrix} \end{array}. \quad (12)$$

In (12), the sub-matrices $Q_{2,1}, Q_{3,2}, \ldots, Q_{b_1-1,b_1-2}$ are identical to those defined in Section 3 and $\widetilde{Q}_{m,m} = Q_{m,m} + \lambda_1 I_{n_2}, m = 1, 2, \ldots, b_1 - 1$. Moreover, the levels $\{0, 1, \ldots, b_1 - 1\}$ of $\mathcal{R}$ represent how many possible customers are in the class-1 queue in front of our target customer upon arrival. Using the same notation from Section 3 whenever possible, it readily follows that

$$\tilde{Q}_{1,0} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ \vdots \\ b_2 \end{array} \begin{array}{c} \begin{array}{ccccc} 0 & 1 & 2 & \ldots & b_2 \end{array} \\ \begin{pmatrix} \begin{bmatrix} \underline{e}'_{k_1,k_1}\underline{e}_{2,1} \otimes \underline{S}'_{0,1} \\ \alpha_1 I_2 \end{bmatrix} & \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \begin{bmatrix} \underline{e}'_{k_1,k_1}\underline{e}_{z_1,z_1-1} \otimes \underline{S}'_{0,1} \\ \alpha_1 I_{z_1} \end{bmatrix} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \begin{bmatrix} \underline{e}'_{k_1,k_1}\underline{e}_{z_1,z_1-1} \otimes \underline{S}'_{0,1} \\ \alpha_1 I_{z_1} \end{bmatrix} & \ddots & \mathbf{0} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \ldots & \begin{bmatrix} \underline{e}'_{k_1,k_1}\underline{e}_{z_1,z_1-1} \otimes \underline{S}'_{0,1} \\ \alpha_1 I_{z_1} \end{bmatrix} \end{pmatrix} \end{array}$$

and

$$\tilde{Q}_{0,0} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ \vdots \\ b_2-1 \\ b_2 \end{array} \begin{array}{c} \begin{array}{ccccccc} 0 & & 1 & 2 & \dots & b_2-1 & b_2 \end{array} \\ \begin{pmatrix} -(\lambda_2 I_2 + V - V_1) & \begin{bmatrix} \mathbf{0} & \lambda_2 I_2 \end{bmatrix} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} \\ \begin{bmatrix} \underline{e}'_{k_2} \underline{e}_{2,2} \otimes \underline{S}'_{0,2} \\ \alpha_2 I_2 \end{bmatrix} & \tilde{\Delta}_1 & \lambda_2 I_{z_1} & \ddots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \Gamma_2 & \tilde{\Delta}_2 & \ddots & \mathbf{0} & \mathbf{0} \\ \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \tilde{\Delta}_{b_2-1} & \lambda_2 I_{z_1} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \Gamma_{b_2} & \tilde{\Delta}_{b_2} \end{pmatrix} \end{array},$$

where $\tilde{\Delta}_i = \Delta_i + \mathrm{diag}(\lambda_1 I_{k_2 s_2}, \lambda_1 I_2 - V_2)$.

According to the structure of the rate matrix $\mathcal{R}$, once our target customer enters the class-1 queue, the Markov chain will progressively make transitions from higher levels to lower ones, indicating the fact that the number of customers in front of the target customer reduces over time. The time to absorption distribution of such a Markov chain has received extensive attention in the literature (e.g., see Latouche and Ramaswami 1999, Chapter 2), and it is well-known that the distribution function of W_1, denoted by $F_1(\omega)$, is given by

$$F_1(\omega) = 1 - \underline{\Phi} \exp\{\mathcal{R}\omega\} \underline{e}'_\ell, \; \omega \geq 0,$$

which is of phase-type form with representation $PH(\underline{\Phi}, \mathcal{R})$. If we now proceed to include the reneging behaviour of our target class-1 customer by defining W_1^* to be the *actual* class-1 waiting time (i.e., the arriving class-1 customer's total time spent in system prior to *successfully* entering service), then it clearly follows that $G_1(\omega) = \Pr(W_1^* \leq \omega) = \Pr(W_1 \leq \omega \mid W_1 \leq R_1)$, where R_1 denotes an exponentially distributed random variable, independent of W_1, with rate α_1. Making use of fundamental matrix algebraic techniques, the following expression for $G_1(\omega)$ is obtained:

$$\begin{aligned} G_1(\omega) &= 1 - \Pr(W_1 > \omega \mid W_1 \leq R_1) \\ &= 1 - \frac{\Pr(\omega < W_1 \leq R_1)}{\Pr(W_1 \leq R_1)} \\ &= 1 - \frac{\int_\omega^\infty \Pr(W_1 > \omega) \alpha_1 e^{-\alpha_1 x} dx - \int_\omega^\infty \Pr(W_1 > x) \alpha_1 e^{-\alpha_1 x} dx}{1 - \int_0^\infty \Pr(W_1 > x) \alpha_1 e^{-\alpha_1 x} dx} \\ &= 1 - \frac{\underline{\Phi}\left[I_\ell - \alpha_1(\alpha_1 I_\ell - \mathcal{R})^{-1}\right] \exp\{\mathcal{R}\omega\} \underline{e}'_\ell e^{-\alpha_1 \omega}}{1 - \alpha_1 \underline{\Phi}(\alpha_1 I_\ell - \mathcal{R})^{-1} \underline{e}'_\ell} \\ &= 1 - \frac{\underline{\Phi}\left[I_\ell - \alpha_1(\alpha_1 I_\ell - \mathcal{R})^{-1}\right]}{1 - \alpha_1 \underline{\Phi}(\alpha_1 I_\ell - \mathcal{R})^{-1} \underline{e}'_\ell} \exp\{(\mathcal{R} - \alpha_1 I_\ell)\omega\} \underline{e}'_\ell \,, \; \omega \geq 0 \,, \end{aligned} \tag{13}$$

which implies that

$$W_1^* \sim PH\left(\frac{\underline{\Phi}\left[I_\ell - \alpha_1(\alpha_1 I_\ell - \mathcal{R})^{-1}\right]}{1 - \alpha_1 \underline{\Phi}(\alpha_1 I_\ell - \mathcal{R})^{-1} \underline{e}'_\ell}, \mathcal{R} - \alpha_1 I_\ell \right).$$

As a consequence (e.g., see Latouche and Ramaswami 1999, Equation 2.13), it is straightforward to show that

$$\mathrm{E}[W_1^{*r}] = \frac{r!\underline{\Phi}(\alpha_1 I_\ell - \mathcal{R})^{-(r+1)}\underline{\rho}'}{1 - \alpha_1 \underline{\Phi}(\alpha_1 I_\ell - \mathcal{R})^{-1}\underline{e}'_\ell}\,,\ r = 1, 2, \ldots\ , \tag{14}$$

where $\underline{\rho}' = -\mathcal{R}\underline{e}'_\ell$ denotes the column vector of absorption rates associated with the phase-type representation of W_1. Finally, as noted at the outset of this section, corresponding results for W_2^* can be obtained in a completely analogous fashion.

5 Numerical Analysis

In this section, we investigate the selection of service discipline parameters k_1 and k_2 in order to optimize the system by way of minimizing a particular cost function. From the previous section, W_i is the nominal class-i waiting time, representing the total time a class-i customer must wait before reaching the server. On the other hand, if R_i denotes the (exponentially distributed) amount of time a class-i customer is willing to wait to reach the server, then $W_i^{\#} = \min\{W_i, R_i\}$ represents the actual time a class-i customer spends waiting in the system. Since W_1 was shown to have a $PH(\underline{\Phi}_1, \mathcal{R}_1)$ representation (with $\underline{\Phi}_1$ and $\mathcal{R}_1$ given by (10) and (12), respectively), it readily follows that $W_1^{\#}$ is also phase-type distributed, but with representation $PH(\underline{\Phi}_1, \mathcal{R}_1 - \alpha_1 I_{\ell_1})$, where ℓ_1 is given by (11). Likewise, $W_2^{\#}$ is phase-type distributed with representation $PH(\underline{\Phi}_2, \mathcal{R}_2 - \alpha_2 I_{\ell_2})$, where ℓ_2, $\underline{\Phi}_2$, and $\mathcal{R}_2$ are similarly determined.

In what follows, we consider the cost function given by

$$\mathrm{Cost} = \mathrm{Cost}_1 + \mathrm{Cost}_2,$$

where

$$\mathrm{Cost}_i = c_i\lambda_i \mathrm{E}[W_i^{\#}] + r_i\lambda_i \mathrm{Pr}(R_i < W_i)$$

and c_i (the waiting cost parameter associated with class i) and r_i (the penalty cost parameter associated with a class-i customer who reneges), $i = 1, 2$, are assumed to be non-negative constants. Due to the phase-type representation for $W_i^{\#}$, it can ultimately be shown that

$$\mathrm{Cost}_i = \lambda_i(c_i + r_i\alpha_i)\underline{\Phi}_i(\alpha_i I_{\ell_i} - \mathcal{R}_i)^{-1}\underline{e}'_{\ell_i}. \tag{15}$$

We remark that this choice of cost function is inspired by the work of Borst et al. (1995), in which the authors studied a cyclic polling model with infinite buffers (but no reneging), and sought to determine optimal k_i values so as to minimize the mean waiting cost of customers, subject to a constraint limiting the number of services per cycle. In particular, by setting the reneging rates α_1 and α_2 both equal to zero, our cost function reduces to their waiting cost function. Moreover, as a means of testing the accuracy of our results, we were able to replicate the choice of optimal (k_1, k_2) in Table I.a, p. 607, of Borst et al. (1995) by setting $\alpha_1 = \alpha_2 = 0$, choosing $b_1 = 45$ and $b_2 = 90$, and calculating the cost function for

all $k_1 = 1, 2, \ldots, 11$, $k_2 = 1, \ldots, 12 - k_1$. These buffer sizes yielded blocking probabilities no larger than 0.0002663 for class 1 and 0.003710 for class 2, which occurred in the most extreme combinations of (k_1, k_2) – namely, $(1, 11)$ or $(11, 1)$. If we only consider those (k_1, k_2) combinations with $k_i \geq 2$, $i = 1, 2$, then the blocking probabilities were no larger than 6.890×10^{-6} for class 1, occurring at $(k_1, k_2) = (2, 10)$, and 7.521×10^{-4} for class 2, occurring at $(k_1, k_2) = (10, 2)$.

Similar to the study conducted by Borst et al. (1995), we investigate the behaviour of our proposed cost function and how optimal (k_1, k_2) combinations might change in the presence of reneging and varying service time distributions, subject to the constraint $k_1 + k_2 \leq K$ which limits the number of services per cycle. As a point of comparison, we consider two specific parametric cases which are both drawn from Section IV of Borst et al. (1995) with $K = 12$. In Case 1, we assume equal arrival rates ($\lambda_1 = \lambda_2 = 0.75$), equal switchover rates ($v_1 = v_2 = 1/0.1$), and mean service times of $\mu_1 = 0.9$ and $\mu_2 = 0.1$. In Case 2, we assume $\mu_1 = \mu_2 = 1$ along with differing arrival/switchover rates according to $\lambda_1 = 0.5$, $\lambda_2 = 0.25$, $v_1 = 1/0.1$, and $v_2 = 1/0.2$. In both cases, we consider reneging rates α_1 and α_2 chosen from the set $\{0.025, 0.05, 0.25\}$. Furthermore, the service time distribution for a given class could either be "Exp" (for exponential), "E_3" (for Erlang-3), or "H_2" (for hyperexponential-2, referring to a mixture of two exponential distributions with selected weights of 0.001 and 0.999). Regardless of which service time distribution is in place, mean service times adhered to the values specified above for Cases 1 and 2. We remark, however, that the H_2 distribution we used possesses 1000 times the variance of the Exp distribution, whereas the E_3 distribution we used possesses 1/3 times the variance of the Exp distribution.

The various parameter combinations resulted in a range of observed blocking probabilities, and the maximum blocking probability per class (over the different possible pairs of k_1 and k_2) for each combination of reneging rate and service time distribution was compared (setting $b_1 = b_2 = 20$). Of these local maxima, class-1 blocking probabilities under Case 1 (Case 2) had a median of 8.431×10^{-6} (1.741×10^{-6}) and a global maximum of 0.1312 (0.0577). With respect to class 2, the local maxima under Case 1 (Case 2) possessed a median of 2.518×10^{-4} (2.772×10^{-10}) and a global maximum of 0.1242 (0.00075). Although our model, with buffer sizes of $b_1 = b_2 = 20$ used throughout, falls short as a precise emulation of the corresponding infinite buffer system for a few combinations of k_i, α_i, and service time distribution (particularly in situations involving the variance-inflated H_2 distribution and low reneging rates), it does a more than adequate job when using only the Exp or E_3 distributions, or when reneging rates are moderate to high. If the goal is to emulate an infinite buffer system with as high of an accuracy as possible, under those aforementioned conditions (e.g., extremely large service time variance), we would recommend increasing b_1 and b_2, computational resources permitting, to achieve more tolerable blocking probabilities.

In Table 1, we investigate the relationship between selected buffer size and blocking probabilities, as well as run times using a 4.00 GHz i7-6700K processor. Here, we simultaneously increase both buffers, while considering (Exp, Exp), (H_2, H_2), and (E_3, E_3) service time distribution combinations. Case 1 parameters with $\alpha_1 = \alpha_2 = 0.025$ were chosen as they produce the highest blocking probabilities of our considered parameter ranges, while $(k_1, k_2) = (6, 6)$

Table 1: Blocking probabilities and run times (in seconds) for varying buffer sizes $b_1 = b_2 = b$ and (Exp, Exp), (H_2, H_2), or (E_3, E_3) service time distributions, with $\alpha_1 = \alpha_2 = 0.025$, $(k_1, k_2) = (6, 6)$, and Case 1 model parameters. $[\lambda_1 = \lambda_2 = 0.75,\ v_1 = v_2 = 1/0.1,\ \mu_1 = 0.9, \mu_2 = 0.1]$

Buffer	Service Time Distributions (Exp, Exp)			(H_2, H_2)			(E_3, E_3)		
b	$P_{b_1,\bullet}$	$P_{\bullet,b_2}$	Run Time	$P_{b_1,\bullet}$	$P_{\bullet,b_2}$	Run Time	$P_{b_1,\bullet}$	$P_{\bullet,b_2}$	Run Time
5	5.1×10^{-2}	8.1×10^{-2}	0.11	3.0×10^{-1}	2.9×10^{-1}	0.19	2.9×10^{-2}	5.6×10^{-2}	0.34
10	4.4×10^{-3}	7.2×10^{-3}	0.45	2.4×10^{-1}	2.3×10^{-1}	1.64	1.3×10^{-3}	2.2×10^{-3}	4.49
15	2.5×10^{-4}	3.9×10^{-4}	1.88	1.8×10^{-1}	1.7×10^{-1}	8.94	3.3×10^{-5}	4.4×10^{-5}	25.73
20	9.1×10^{-6}	1.3×10^{-5}	5.89	1.3×10^{-1}	1.2×10^{-1}	31.38	5.2×10^{-7}	5.9×10^{-7}	94.89
25	2.2×10^{-7}	3.1×10^{-7}	15.30	8.4×10^{-2}	7.7×10^{-2}	86.81	5.4×10^{-9}	5.4×10^{-9}	296.52
30	3.4×10^{-9}	4.9×10^{-9}	34.34	4.6×10^{-2}	4.0×10^{-2}	205.29	$< 1\times10^{-10}$	$< 1\times10^{-10}$	767.57
35	$< 1\times10^{-10}$	$< 1\times10^{-10}$	69.48	1.9×10^{-2}	1.6×10^{-2}	456.00	$< 1\times10^{-10}$	$< 1\times10^{-10}$	1649.16
40	$< 1\times10^{-10}$	$< 1\times10^{-10}$	128.29	5.4×10^{-3}	4.1×10^{-3}	941.35	$< 1\times10^{-10}$	$< 1\times10^{-10}$	3162.91
45	$< 1\times10^{-10}$	$< 1\times10^{-10}$	224.64	9.3×10^{-4}	6.3×10^{-4}	1719.31	$< 1\times10^{-10}$	$< 1\times10^{-10}$	5630.37
50	$< 1\times10^{-10}$	$< 1\times10^{-10}$	369.50	9.6×10^{-5}	5.9×10^{-5}	2930.37	$< 1\times10^{-10}$	$< 1\times10^{-10}$	9365.38

was selected so as to maximize the dimension of the state space of our process, subject to $k_1 + k_2 = 12$. It is clear that under Exp or E_3 service, blocking probabilities are not an issue at our selected $b_1 = b_2 = 20$. However, as stated earlier, the blocking probabilities are not ideal at this range under H_2 service. Unfortunately, to reduce these blocking probabilities to under a percentage, we would need to increase the buffer sizes beyond 35, which greatly increases the computational load for the H_2 service time distribution, and drastically more so for the E_3 distribution. Thus, in an effort to keep computation times manageable, we elected to accept these blocking probabilities and use buffer sizes of 20 apiece over all parameter combinations. To check the impact of this decision, we also considered only the (H_2, H_2) combination for both cases with $b_1 = b_2 = 40$ in Table 4, which we will discuss shortly.

Tables 2 and 3 display the optimal (k_1, k_2) pairs, along with their corresponding cost values, for each combination of reneging rate and service time distribution under Cases 1 and 2, respectively. In each table, we present results corresponding to $c_1 = 2$, $c_2 = 1$, $r_1 = 1$, and $r_2 = 0.5$, as well as results for select combinations of service time distribution when $r_1 = r_2 = 40$. In looking at the optimal values of k_1 and k_2 under Case 1 over a range of cost parameters, we observed that the limit of the optimal choice of (k_1, k_2) is $(11, 1)$ as c_1 or r_1 approaches ∞, or $(1, 11)$ as c_2 or r_2 approaches ∞. An example of this convergence is illustrated in Fig. 2, where we plotted the optimal values of k_1 against c_1 (with c_2, r_1, and r_2 held constant). The rates of convergence (to $k_1 = 11$) appear to be largely dependent on the relative values of α_1 and α_2. Note that k_1 converges faster when class-2 customers are more impatient, causing fewer of them to reach service and resulting in relatively longer class-1 queues. This causes class 1 to dominate the expected time waiting in system portion of the cost function, whereas class-2 customers dominate the probability of reneging portion. Since we are plotting against the class-1 waiting cost parameter (while keeping reneging costs constant), it is easy to see why the $(0.025, 0.25)$ combination converges the fastest and

Table 2: Optimal (k_1, k_2) and minimum cost values for Case 1 with $c_1 = 2$, $c_2 = 1$, and $r_1 = 1$, $r_2 = 0.5$ or $r_1 = r_2 = 40$. [$\lambda_1 = \lambda_2 = 0.75$, $v_1 = v_2 = 1/0.1$, $\mu_1 = 0.9, \mu_2 = 0.1$]

Reneging Rates		Service Time Distributions							
		(Exp, Exp)		(Exp, H_2)		(Exp, E_3)		(Exp, Exp)	
α_1	α_2	(k_1, k_2)	Cost	(k_1, k_2)	Cost	(k_1, k_2)	Cost	(k_1, k_2)	Cost
0.025	0.025	(3, 9)	4.3398	(3, 9)	6.2866	(3, 9)	4.3281	(3, 9)	6.9361
	0.05	(4, 8)	4.2581	(3, 9)	5.9977	(4, 8)	4.2468	(2, 10)	7.7429
	0.25	(7, 5)	3.7352	(9, 3)	4.8325	(7, 5)	3.7269	(2, 10)	11.9875
0.05	0.025	(3, 9)	3.6482	(3, 9)	5.2460	(3, 9)	3.6386	(3, 9)	7.1422
	0.05	(3, 9)	3.5947	(3, 9)	4.9824	(3, 9)	3.5855	(2, 10)	7.8847
	0.25	(6, 6)	3.2543	(6, 6)	4.0882	(6, 6)	3.2470	(1, 11)	11.9519
0.25	0.025	(2, 10)	2.1520	(2, 10)	3.0167	(2, 10)	2.1464	(3, 9)	9.0264
	0.05	(2, 10)	2.1334	(2, 10)	2.8169	(2, 10)	2.1279	(3, 9)	9.7272
	0.25	(2, 10)	2.0230	(2, 10)	2.2667	(2, 10)	2.0183	(1, 11)	13.3357
		(H_2, Exp)		(H_2, H_2)		(H_2, E_3)		(H_2, H_2)	
α_1	α_2	(k_1, k_2)	Cost	(k_1, k_2)	Cost	(k_1, k_2)	Cost	(k_1, k_2)	Cost
0.025	0.025	(4, 8)	20.3486	(3, 9)	21.7547	(4, 8)	20.3441	(3, 9)	35.8657
	0.05	(5, 7)	18.2711	(4, 8)	19.5065	(5, 7)	18.2666	(3, 9)	36.3322
	0.25	(8, 4)	14.6158	(9, 3)	15.5758	(8, 4)	14.6114	(2, 10)	33.8738
0.05	0.025	(3, 9)	16.4205	(2, 10)	17.4343	(3, 9)	16.4171	(2, 10)	34.1935
	0.05	(4, 8)	14.3710	(3, 9)	15.2235	(4, 8)	14.3676	(3, 9)	34.6786
	0.25	(7, 5)	10.7526	(8, 4)	11.3615	(7, 5)	10.7490	(2, 10)	32.2619
0.25	0.025	(1, 11)	8.8681	(1, 11)	9.4376	(1, 11)	8.8681	(3, 9)	27.1216
	0.05	(1, 11)	6.9769	(1, 11)	7.3890	(1, 11)	6.9756	(3, 9)	27.6632
	0.25	(6, 6)	3.5293	(5, 7)	3.7245	(6, 6)	3.5263	(2, 10)	25.4136
		(E_3, Exp)		(E_3, H_2)		(E_3, E_3)		(E_3, E_3)	
α_1	α_2	(k_1, k_2)	Cost	(k_1, k_2)	Cost	(k_1, k_2)	Cost	(k_1, k_2)	Cost
0.025	0.025	(4, 8)	3.4530	(3, 9)	5.5164	(4, 8)	3.4401	(3, 9)	5.5277
	0.05	(4, 8)	3.3965	(4, 8)	5.2494	(4, 8)	3.3839	(3, 9)	6.2533
	0.25	(7, 5)	3.0303	(8, 4)	4.2083	(7, 5)	3.0210	(2, 10)	10.1034
0.05	0.025	(3, 9)	2.9907	(3, 9)	4.6552	(3, 9)	2.9800	(3, 9)	5.8344
	0.05	(3, 9)	2.9587	(3, 9)	4.4143	(3, 9)	2.9483	(2, 10)	6.5219
	0.25	(5, 7)	2.7179	(6, 6)	3.6040	(5, 7)	2.7096	(1, 11)	10.1745
0.25	0.025	(2, 10)	1.8771	(2, 10)	2.7582	(2, 10)	1.8710	(4, 8)	7.8647
	0.05	(2, 10)	1.8677	(2, 10)	2.5692	(2, 10)	1.8618	(3, 9)	8.4997
	0.25	(2, 10)	1.8065	(2, 10)	2.0618	(2, 10)	1.8013	(1, 11)	11.8951
(r_1, r_2)		(1, 0.5)		(1, 0.5)		(1, 0.5)		(40, 40)	

Table 3: Optimal (k_1, k_2) and minimum cost values for Case 2 with $c_1 = 2$, $c_2 = 1$, and $r_1 = 1$, $r_2 = 0.5$ or $r_1 = r_2 = 40$. $[\lambda_1 = 0.5, \lambda_2 = 0.25, v_1 = 1/0.1, v_2 = 1/0.2, \mu_1 = \mu_2 = 1]$

Reneging Rates		Service Time Distributions							
		(Exp, Exp)		(Exp, H_2)		(Exp, E_3)		(Exp, Exp)	
α_1	α_2	(k_1, k_2)	Cost	(k_1, k_2)	Cost	(k_1, k_2)	Cost	(k_1, k_2)	Cost
0.025	0.025	(10, 2)	2.6649	(10, 2)	7.8090	(10, 2)	2.4761	(10, 2)	4.3972
	0.05	(11, 1)	2.4083	(11, 1)	6.8184	(11, 1)	2.2577	(9, 3)	4.6890
	0.25	(11, 1)	1.8357	(11, 1)	5.1854	(11, 1)	1.7432	(8, 4)	5.9054
0.05	0.025	(10, 2)	2.3812	(9, 3)	5.6575	(10, 2)	2.2247	(10, 2)	4.6660
	0.05	(10, 2)	2.2030	(10, 2)	4.8100	(10, 2)	2.0669	(10, 2)	4.9625
	0.25	(11, 1)	1.6934	(11, 1)	3.5890	(11, 1)	1.6127	(8, 4)	6.1953
0.25	0.025	(6, 6)	1.5047	(5, 7)	2.9877	(6, 6)	1.4353	(11, 1)	6.4562
	0.05	(7, 5)	1.4550	(6, 6)	2.2615	(7, 5)	1.3907	(10, 2)	6.7245
	0.25	(11, 1)	1.2323	(10, 2)	1.5454	(11, 1)	1.1879	(8, 4)	7.8861
		(H_2, Exp)		(H_2, H_2)		(H_2, E_3)		(H_2, H_2)	
α_1	α_2	(k_1, k_2)	Cost	(k_1, k_2)	Cost	(k_1, k_2)	Cost	(k_1, k_2)	Cost
0.025	0.025	(11, 1)	11.9085	(10, 2)	15.6149	(11, 1)	11.8306	(10, 2)	24.8349
	0.05	(11, 1)	10.6576	(11, 1)	13.9119	(11, 1)	10.5843	(10, 2)	23.2350
	0.25	(11, 1)	9.5631	(11, 1)	12.3096	(11, 1)	9.5085	(9, 3)	21.9772
0.05	0.025	(10, 2)	8.2620	(9, 3)	10.5575	(10, 2)	8.1927	(9, 3)	20.6821
	0.05	(11, 1)	7.0367	(10, 2)	8.9056	(11, 1)	6.9730	(9, 3)	19.1262
	0.25	(11, 1)	5.9644	(11, 1)	7.4260	(11, 1)	5.9166	(8, 4)	17.9732
0.25	0.025	(4, 8)	3.9129	(4, 8)	5.0239	(4, 8)	3.8846	(8, 4)	15.7298
	0.05	(9, 3)	2.8207	(6, 6)	3.4547	(9, 3)	2.7818	(8, 4)	14.2528
	0.25	(11, 1)	1.8533	(9, 3)	2.1237	(11, 1)	1.8198	(7, 5)	13.2301
		(E_3, Exp)		(E_3, H_2)		(E_3, E_3)		(E_3, E_3)	
α_1	α_2	(k_1, k_2)	Cost	(k_1, k_2)	Cost	(k_1, k_2)	Cost	(k_1, k_2)	Cost
0.025	0.025	(10, 2)	2.2940	(10, 2)	7.5463	(10, 2)	2.0953	(10, 2)	3.4708
	0.05	(11, 1)	2.0828	(11, 1)	6.5993	(11, 1)	1.9246	(9, 3)	3.7560
	0.25	(11, 1)	1.5665	(11, 1)	5.0376	(11, 1)	1.4696	(7, 5)	5.0495
0.05	0.025	(10, 2)	2.0877	(9, 3)	5.4309	(10, 2)	1.9226	(10, 2)	3.7678
	0.05	(10, 2)	1.9375	(10, 2)	4.6075	(10, 2)	1.7938	(9, 3)	4.0603
	0.25	(11, 1)	1.4762	(11, 1)	3.4371	(11, 1)	1.3913	(7, 5)	5.3424
0.25	0.025	(6, 6)	1.3829	(5, 7)	2.8826	(6, 6)	1.3106	(11, 1)	5.6070
	0.05	(7, 5)	1.3415	(7, 5)	2.1606	(7, 5)	1.2744	(10, 2)	5.8895
	0.25	(11, 1)	1.1422	(10, 2)	1.4620	(11, 1)	1.0960	(8, 4)	7.0570
(r_1, r_2)		(1, 0.5)		(1, 0.5)		(1, 0.5)		(40, 40)	

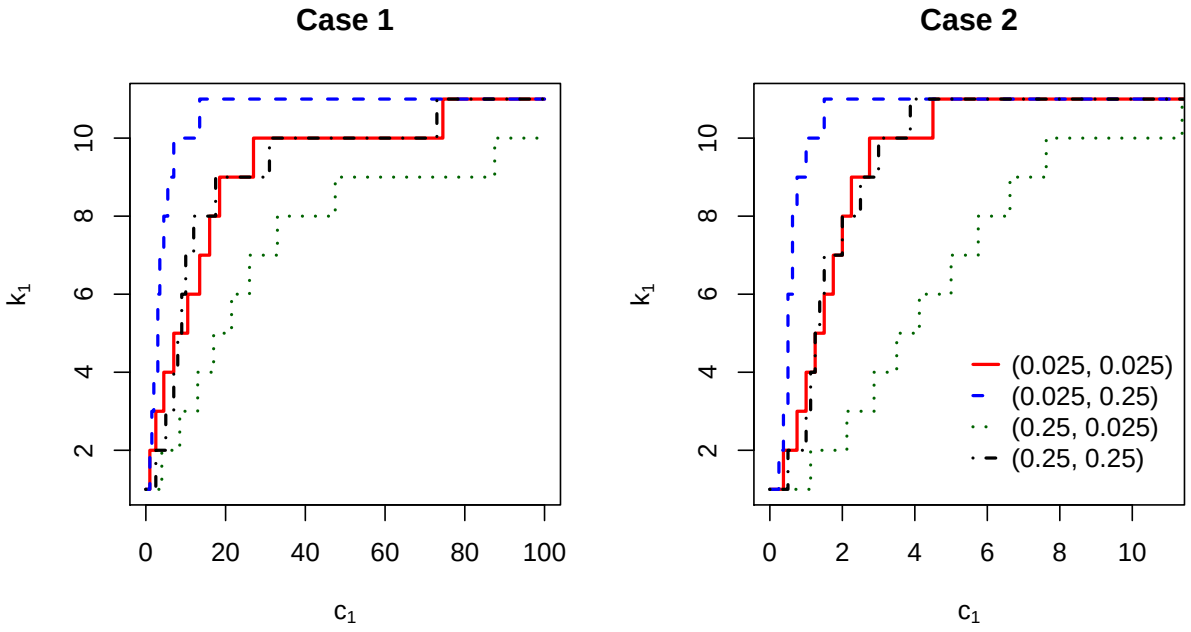

Figure 2: Plots of k_1 vs. c_1 under both Cases 1 and 2 with Exp service times, $c_2 = 2$, $r_1 = r_2 = 1$, and four combinations of reneging rates.

$(0.25, 0.025)$ the slowest, whereas equal reneging rate combinations tend to be comparable to one another. This result is consistent between Cases 1 and 2. However, we note that in Case 2, since the class-1 arrival rate is twice that of class 2, costs associated with class 1 dominate the cost function sooner. As a result, we observed that Case 2's system converges to $(11, 1)$ faster and $(1, 11)$ slower in comparison to Case 1's system.

Tables 2 and 3 also suggest that when the waiting and reneging cost parameters are of a comparable size (or the waiting cost parameters are much larger), the size of k_i is inversely proportional to α_i (while keeping the other class's reneging rate constant). When the reneging cost parameters are much larger than the waiting cost parameters, such as in the $r_1 = r_2 = 40$ examples, this relationship may invert, as serving less class-i customers per cycle, in combination with a larger α_i that increases the probability of a class-i customer reneging before service becomes available, becomes more costly. In general, the system appears to be more sensitive to smaller changes in the waiting cost parameters. This is an intuitive result considering the final form of the cost function in (15), in which the product of r_i and α_i is present. Depending on the choice of reneging rates, this may lead to the reneging cost being much smaller than the waiting cost.

For a given pair of α_1 and α_2, we observed that changing the service time distribution could affect the optimal choice of k_1 and k_2. Our results in Tables 2 and 3 indicate that it is possible to vary the optimal (k_1, k_2) values by switching only one (or both) of the service

Table 4: Optimal (k_1, k_2) with corresponding minimum cost values, as well as maximal class-1 and class-2 blocking probabilities over all possible (k_1, k_2) combinations, for (H_2, H_2) service time distributions under Case 1 and Case 2 model parameters with $b_1 = b_2 = 40$, $c_1 = 2$, $c_2 = 1$, $r_1 = 1$, and $r_2 = 0.5$.

Reneging Rates		Case 1				Case 2			
α_1	α_2	(k_1, k_2)	Cost	$\max P_{b_1,\bullet}$	$\max P_{\bullet,b_2}$	(k_1, k_2)	Cost	$\max P_{b_1,\bullet}$	$\max P_{\bullet,b_2}$
0.025	0.025	(3, 9)	29.8757	5.4×10^{-3}	4.1×10^{-3}	(10, 2)	17.1809	1.3×10^{-5}	$< 1\times10^{-10}$
	0.05	(5, 7)	25.2273	5.4×10^{-3}	1.3×10^{-8}	(11, 1)	15.4646	1.3×10^{-5}	$< 1\times10^{-10}$
	0.25	(9, 3)	21.1141	5.4×10^{-3}	$< 1\times10^{-10}$	(11, 1)	13.8489	1.3×10^{-5}	$< 1\times10^{-10}$
0.05	0.025	(2, 10)	20.3077	3.1×10^{-8}	4.0×10^{-3}	(9, 3)	10.5754	$< 1\times10^{-10}$	$< 1\times10^{-10}$
	0.05	(3, 9)	15.7629	3.1×10^{-8}	1.3×10^{-8}	(10, 2)	8.9186	$< 1\times10^{-10}$	$< 1\times10^{-10}$
	0.25	(8, 4)	11.7568	3.2×10^{-8}	$< 1\times10^{-10}$	(11, 1)	7.7391	$< 1\times10^{-10}$	$< 1\times10^{-10}$
0.25	0.025	(1, 11)	11.8187	$< 1\times10^{-10}$	3.8×10^{-3}	(4, 8)	5.0286	$< 1\times10^{-10}$	$< 1\times10^{-10}$
	0.05	(1, 11)	7.5269	$< 1\times10^{-10}$	1.3×10^{-8}	(6, 6)	3.4547	$< 1\times10^{-10}$	$< 1\times10^{-10}$
	0.25	(5, 7)	3.7245	$< 1\times10^{-10}$	$< 1\times10^{-10}$	(9, 3)	2.1237	$< 1\times10^{-10}$	$< 1\times10^{-10}$

time distributions. The larger the difference between the variances of the previous and new service time distributions, the more likely we were to observe changes in the optimal (k_1, k_2) values. In many situations, there was no discernible difference when comparing the Exp and E_3 distributions, except for a decrease in the optimal cost when using the E_3 distribution. However, when comparing either Exp or E_3 against the H_2 distribution, it was common to find different optimal (k_1, k_2) pairs and we always observed an increase in the optimal cost.

Although there is some evidence to suggest that the optimal (k_1, k_2) values are, more or less, insensitive to the second moment of the service time distribution in our model (and this is consistent with the remarks in Borst et al. 1995), we did capture varying results by inflating the relative variance difference between the two service time distributions to a large enough degree. One may be inclined to attribute the presence of these observed changes in our optimal results to only the occasional high blocking probability rather than the service time distribution, but we must emphasize that some of these variations were still present in settings with negligible blocking probabilities, such as, for example, when $\alpha_1 = \alpha_2 = 0.25$ and H_2 service times are used for both classes. In order to confirm that these differences are due to the service time distributions for lower reneging rates, and not just by negative expected waiting time bias as a result of non-negligible blocking probabilities, we reran the experiment for (H_2, H_2) at $b_1 = b_2 = 40$, which guaranteed maximal blocking probabilities of less than a percent for Cases 1 and 2. This updated data is presented in Table 4. While the aforementioned negative bias in $E[W_i^{\#}]$ is observable in the optimal costs, all optimal (k_1, k_2) are unchanged except when $(\alpha_1, \alpha_2) = (0.025, 0.05)$ in Case 1. Here, however, we observe a shift from $(4, 8)$ to $(5, 7)$, bringing it in line with (H_2, Exp) and (H_2, E_3), rather than having the same optimal (k_1, k_2) as (Exp, Exp). This supports our claim that these observed deviations are due to our choice of H_2 distribution, and not solely due to the blocking probabilities.

So while blocking probabilities can contribute to the variability in the optimal (k_1, k_2)

combinations, we cannot conclude that the selection is completely insensitive, as differences may exist even when variances are similar (for instance, compare (Exp, Exp) with (E_3, Exp) in Table 2). Furthermore, based on the form of (15), another conclusion may be made. By selecting a larger arrival rate for a class, the expected time waiting in the system will increase, as well as the probability of reneging, while simultaneously raising that class' weight in the cost function. This will heighten the system's sensitivities to the service time distribution of that class, and may lead to more variation in the selection of optimal (k_1, k_2) when comparing combinations of service time distribution for that class with smaller differences in variance.

To illustrate the impact of the service time distribution's second moment, we plot in Fig. 3 the reneging probability for a customer of either class against M_{SM} under Case 2 parameters and $\alpha_1 = \alpha_2 = 0.05$ for $(k_1, k_2) \in \{(6,6), (11,1), (1,11)\}$ with $(\mathrm{H}_2(M_{SM}), \mathrm{Exp})$, $(\mathrm{Exp}, \mathrm{H}_2(M_{SM}))$, or $(\mathrm{H}_2(M_{SM}), \mathrm{H}_2(M_{SM}))$ service time distributions. Here, $\mathrm{H}_2(M_{SM})$ denotes the phase-type distribution

$$PH\left(\left(\frac{0.25}{M_{SM}-1+0.25}, \frac{M_{SM}-1}{M_{SM}-1+0.25}\right), \frac{1}{\mu_i}\begin{bmatrix} \frac{-0.5}{M_{SM}-0.5} & 0 \\ 0 & -2 \end{bmatrix}\right),$$

which ensures a mean class-i service time of μ_i, and a second moment of $M_{SM}(2\mu_i^2)$, for $M_{SM} > 1$. In words, M_{SM} acts as a multiplier on the second moment, relative to the Exp service time distribution for that class. We remark that the following results also apply to $\mathrm{E}[W_i^{\#}]$, since as seen in (15), the reneging probability for class i equals $\alpha_i \mathrm{E}[W_i^{\#}]$, resulting in plots of $\mathrm{E}[W_i^{\#}]$ against M_{SM} having the same shape.

From Fig. 3, we first observe that a customer's reneging probability increases with M_{SM}, but becomes insensitive to further increases after the second moment becomes large. Moreover, increasing the second moment of class 2's service time distribution only results in a smaller increase in reneging probability (for either class) than the same increase for only class 1's service time distribution. This is due to $\lambda_1 = 2\lambda_2$ under Case 2, which results in twice as many (on average) class-1 customers arriving to the system than class-2 customers over a period of time, thereby creating more opportunities for the server to be involved in a particularly long service. In addition, it is not surprising that increasing the second moment for both classes simultaneously results in a larger increase in reneging probability than increasing a single class in isolation. Finally, in the corresponding plots for Case 1 (which we have omitted here), we observed that there was much less sensitivity when increasing the second moment for class 2, as a result of its smaller mean service time of 0.1 (and scaling class 1 was similar as in Case 2, since their expected values only differ by 0.1). We can therefore conclude that in the presence of reneging, it would be quite inappropriate to approximate a highly variable service time distribution with a simpler exponential distribution.

Next, Figs. 4 and 5 present plots of $G_i(\omega)$, the distribution function of the actual waiting time random variable W_i^*, as defined in Section 4. These functions were evaluated via (13) for both classes under a particular pair of reneging rates (namely, $\alpha_1 = 0.025$ and $\alpha_2 = 0.25$) and four combinations of Exp and H_2 service times, with Cases 1 and 2 presented in Figs. 4 and 5, respectively. For each combination of service time distribution, the optimal values

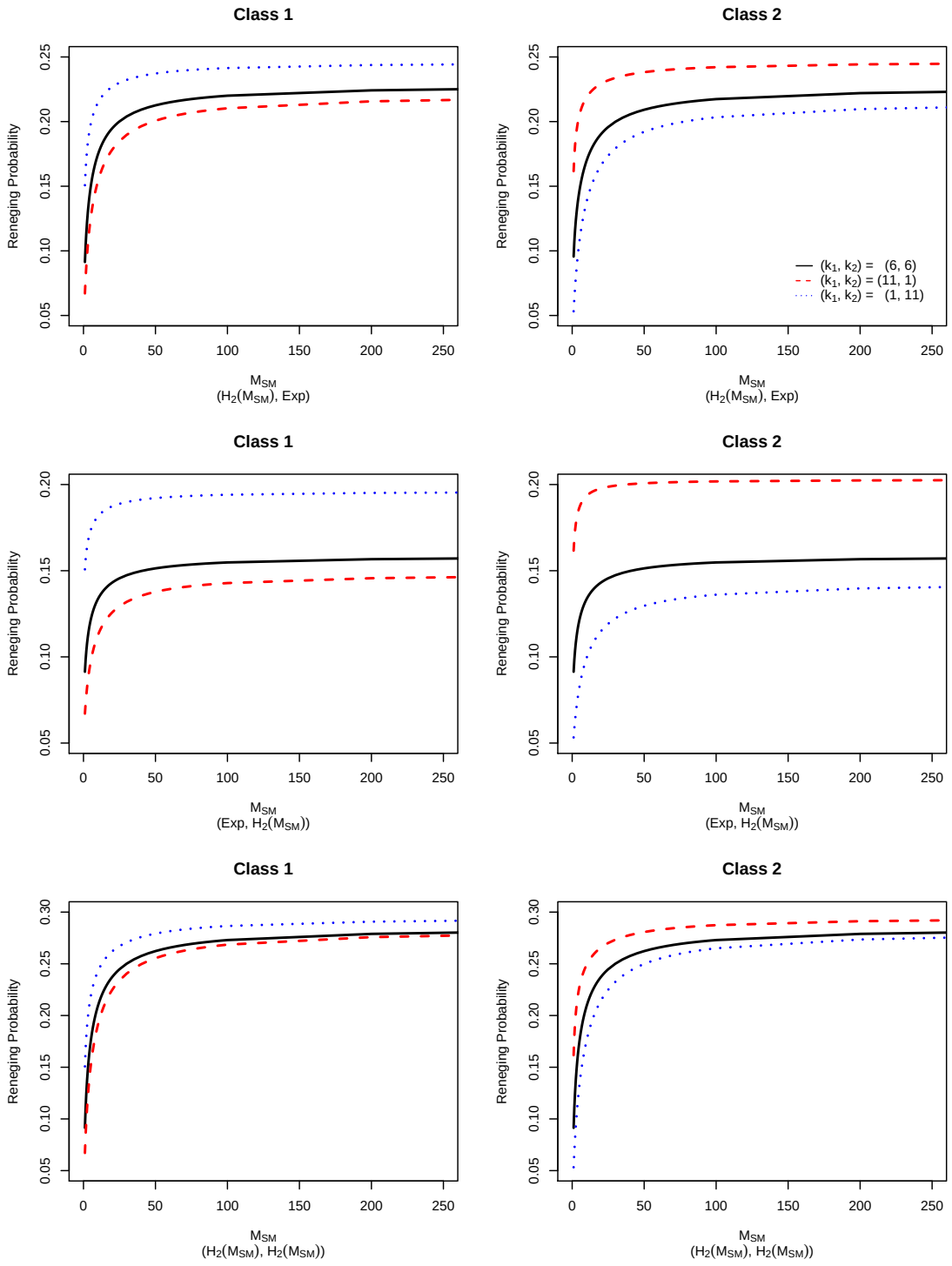

Figure 3: Plots of reneging probability vs. M_{SM} for both classes under Case 2 with $\alpha_1 = \alpha_2 = 0.05$ and $(k_1, k_2) \in \{(6,6), (11,1), (1,11)\}$.

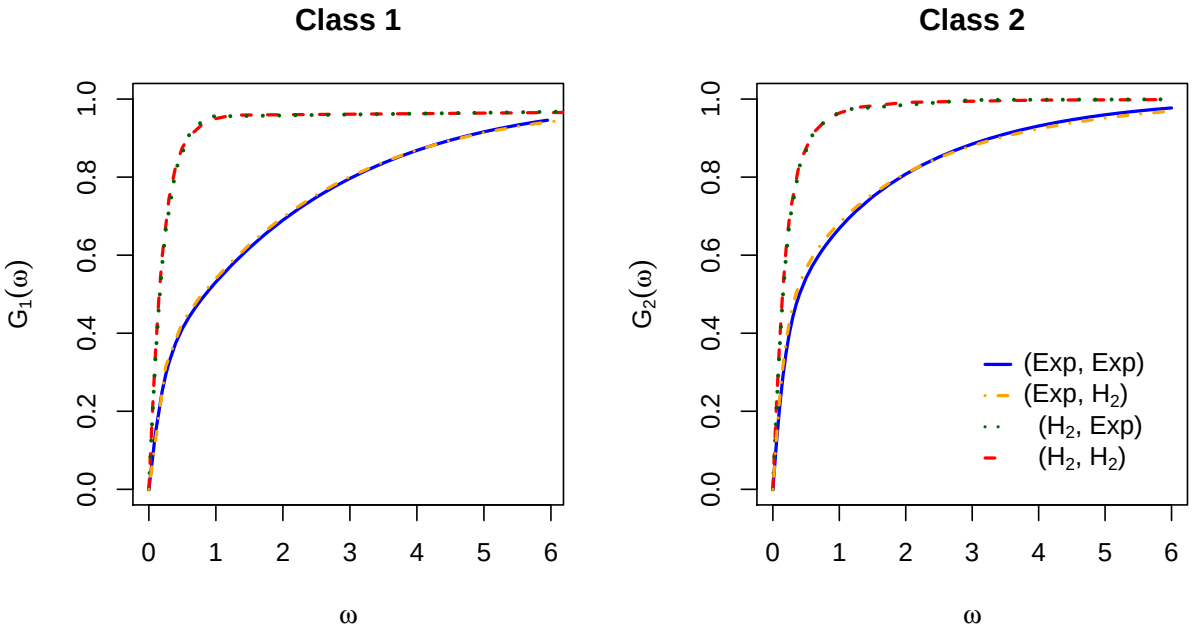

Figure 4: Plots of $G_i(\omega)$ vs. ω for both classes under Case 1 with $\alpha_1 = 0.025$, $\alpha_2 = 0.25$, either Exp or H_2 service times, and optimal (k_1, k_2) from Table 2.

of k_1 and k_2 were selected for use from Tables 2 and 3. It is interesting to note that the H_2 distribution in these cases typically yielded shorter actual waiting times than the Exp distribution. This is due to the fact that the actual waiting time distribution is conditional on the customer reaching service before reneging. In order to have the same mean as the Exp distribution (while inflating the variance), the H_2 distribution was constructed as a mixture of two exponential distributions, one with a higher rate and a great likelihood of occurrence (namely, 99.9%) and the other with a very low rate and a rare chance of occurrence (namely, 0.1%). The conditional nature of W_i^* results in an exponentially distributed bound on the total time for the preceding customers' service/reneging times, implying that if the reneging rate of the target customer is high enough (so that the bound on the total time is short enough), we only really observe services which follow the higher rate. This essentially reduces the H_2 distribution to an exponential distribution with faster service times (and hence shorter actual waiting times).

In order to better understand the behaviour of the actual waiting time distribution, we also calculated $\mathrm{E}[W_i^*]$ via (14) for a variety of scenarios. Specifically, we considered each reneging rate pair (as in Tables 2 and 3), values of $k_i \in \{1, 2, \ldots, 11\}$ satisfying $k_1 + k_2 = 12$, and combinations of Exp and H_2 service times. We found that the H_2 distribution yielded shorter mean actual waiting times in comparison to the Exp distribution (all else being equal), except in Case 1 when the distribution we were changing was that of class 2, and the

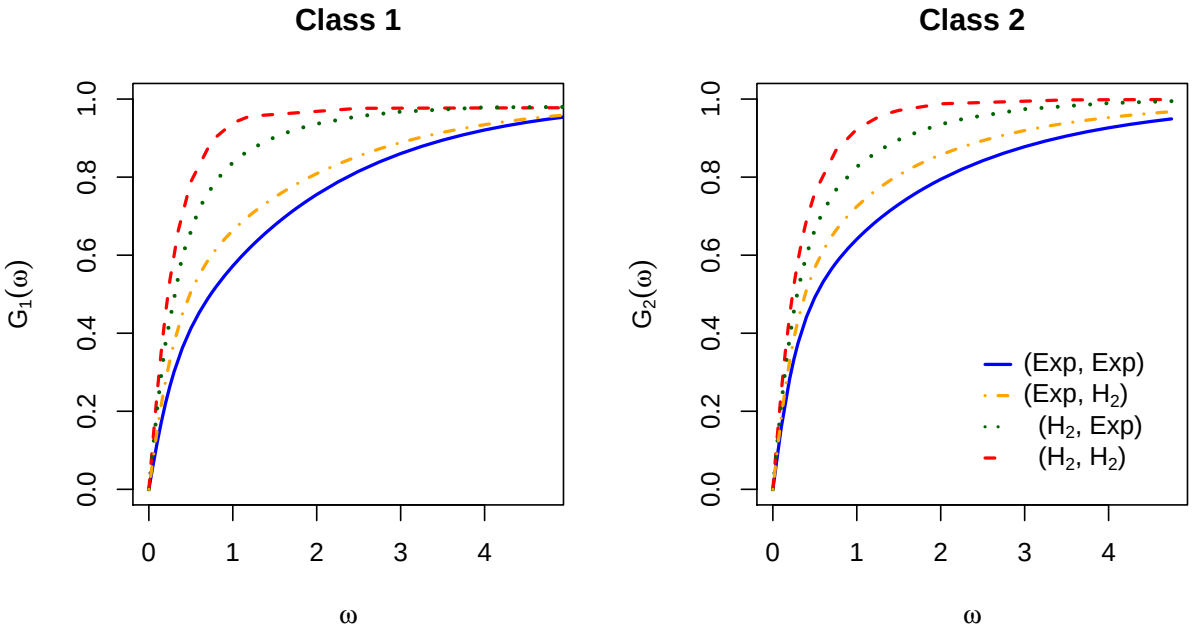

Figure 5: Plots of $G_i(\omega)$ vs. ω for both classes under Case 2 with $\alpha_1 = 0.025$, $\alpha_2 = 0.25$, either Exp or H_2 service times, and optimal (k_1, k_2) from Table 3.

reneging rate of the target customer was not large (i.e., 0.025 or 0.05). This is likely due to the combination of a small mean service time for class 2 and longer exponentially distributed time bounds failing to reproduce the "reducing" effect as previously noted, thereby allowing the H_2 distribution to act (approximately) unchanged for class 2. Thus, the higher service time variance resulted in higher expected actual waiting times. In some of these instances, however, when k_2 was small and the target customer came from class 2, it was still possible for the mean actual waiting time to decrease. Since the small value of k_2 increases the probability of class-2 customers waiting in front of the target customer to renege (instead of reaching service, despite the small value of α_2), this reduces the effect of the service time distribution in place.

Overall, we observed that the mean actual waiting times in Case 1 were primarily dependent on the class-1 service time distribution due to its larger mean service time (i.e., $\mu_1 = 0.9$ vs. $\mu_2 = 0.1$). Combined with the fact that a high reneging rate for a particular class reduces the influence of that class's service time distribution, we witnessed the rather extreme situation seen in Fig. 4, where the distribution is almost entirely dependent on class 1 (since $\alpha_1 = 0.025$ and $\alpha_2 = 0.25$). While the assumption of equal mean service times helped balance the dependence between classes in Case 2, the fact that the class-1 arrival rate is twice that of class 2 still resulted in a larger influence from class 1, as seen in Fig. 5. Finally, as expected, increasing k_i decreased class-i's mean actual waiting time, since

it increased the number of customers (queued ahead of the target customer) that could be served during a single visit by the server.

6 Concluding Remarks

We have presented a 2-class, single-server polling model with reneging customers operating under a k_i-limited service discipline in order to analyze the queueing performance of the system in terms of customer loss and delay. Using matrix analytic techniques, we derived important performance measures of interest including the joint queue length distribution as well as the per-class waiting time distributions and their moments. We utilized our mathematical results to investigate a constrained optimization problem under a variety of scenarios involving the system parameters. All in all, our results revealed a number of interesting and non-trivial combinations of (k_1, k_2) which minimize our proposed objective function.

Future work will proceed along two fronts. On the one hand, our goal is to generalize the model beyond the exponential interarrival/reneging assumption to something more wide-ranging such as a phase-type or Coxian distribution. On the other, we plan to extend the existing results to allow for a third class of customers. The approach would parallel the methodology used in the 2-class case, although we anticipate that the dimensionality of the phase-type representation for the waiting time distribution would increase considerably due to the presence of the extra queue. In general, we expect that using this exact method for a system having N queues (with finite buffer sizes $b_1, b_2, ..., b_N$) would require

$$\sum_{t_1=0}^{1}\sum_{t_2=0}^{1}\cdots\sum_{t_N=0}^{1}\left(\prod_{i=1}^{N} b_i^{t_i}\right)\left(N+\sum_{j=1}^{N} t_j k_j s_j\right)$$

total states, and thus the rate matrix of the underlying phase-type distribution for the waiting time would be of dimension

$$\sum_{t_1=0}^{1}\sum_{t_2=0}^{1}\cdots\sum_{t_N=0}^{1}\left((b_1-1)^{t_1}\prod_{i=2}^{N} b_i^{t_i}\right)\left(N+\sum_{j=1}^{N} t_j k_j s_j\right).$$

As such, the particular structure of the block matrices comprising the associated infinitesimal generator may require further exploitation in order to reduce the memory storage requirements inherent in this model formulation.

Acknowledgements Steve Drekic and Kevin Granville thank the anonymous referee and the editor-in-chief for their supportive comments and helpful suggestions. Steve Drekic and Kevin Granville also acknowledge the financial support from the Natural Sciences and Engineering Research Council of Canada through its Discovery Grants program (#RGPIN-2016-03685) and Postgraduate Scholarship-Doctoral program, respectively.

Conflict of Interest The authors declare no conflict of interest.

References

[1] Boon, M.A.A. (2011). Polling Models: From Theory to Traffic Intersections. Doctoral dissertation, Eindhoven: Technische Universiteit Eindhoven, 190 pages.

[2] Boon, M.A.A., van der Mei, R.D., & Winands, E.M.M. (2011). Applications of polling systems. *Surveys in Operations Research and Management Science*, 16(2), 67-82.

[3] Boon, M.A.A. (2012). A polling model with reneging at polling instants. *Annals of Operations Research*, 198(1), 5–23.

[4] Boon, M.A.A., van Wijk, A.C.C., Adan, I.J.B.F., & Boxma, O.J. (2010). A polling model with smart customers. *Queueing Systems*, 66(3), 239–274.

[5] Boon, M.A.A. & Winands, E.M.M. (2014). Heavy-traffic analysis of k-limited polling systems. *Probability in the Engineering and Informational Sciences*, 28(4), 451–471.

[6] Borst, S.C., Boxma, O.J., & Levy, H. (1995). The use of service limits for efficient operation of multistation single-medium communication systems. *IEEE/ACM Transactions on Networking*, 3(5), 602–612.

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

[8] Chang, W. & Down, D.G. (2002). Exact asympototics for k_i-limited exponential polling models. *Queueing Systems*, 42(4), 401–419.

[9] Gaver, D.P., Jacobs, P.A., & Latouche, G. (1984). Finite birth-and-death models in randomly changing environments. *Advances in Applied Probability*, 16(4), 715–731.

[10] Graves, S.C. (1982). The application of queueing theory to continuous perishable inventory systems. *Management Science*, 28(4), 400–406.

[11] Gromoll, H.C., Robert, P., Zwart, B., & Bakker, R. (2006). The impact of reneging in processor sharing queues. In Proceedings of the Joint International Conference on Measurement and Modeling of Computer Systems, Saint Malo, France, pp. 87–96.

[12] Horng, S.-C. & Lin, S.-Y. (2009). Ordinal optimization of $G/G/1/K$ polling systems with k-limited service discipline. *Journal of Optimization*, 140(2), 213–231.

[13] Latouche, G. & Ramaswami, V. (1999). *Introduction to Matrix Analytic Methods in Stochastic Modeling*. Philadelphia, PA: ASA SIAM.

[14] Levy, H. & Sidi, M. (1990). Polling systems: applications, modeling and optimization. *IEEE Transactions on Communications*, COM-38(10), 1750–1760.

[15] MacPhee, I., Menshikov, M., Petritis, D., & Popov, S. (2007). A Markov chain model of a polling system with parameter regeneration. *Annals of Applied Probability*, 17(5/6), 1447–1473.

[16] Mishkoy, G., Krieger, U.R., & Bejenari, D. (2012). Matrix algorithm for polling models with PH distribution. *Buletinul Academiei de Ştiinţe a Republicii Moldova. Matematica*, 68(1), 70–80.

[17] Perel, E. & Yechiali, U. (2017). Two-queue polling systems with switching policy based on the queue that is not being served. *Stochastic Models*, 33(3), 430-450.

[18] Takagi, H. (1988). Queueing analysis of polling models. *ACM Computing Surveys*, 20(1), 5–28.

[19] Tijms, H.C. (2003). *A First Course in Stochastic Models*. Chichester, UK: John Wiley & Sons.

[20] van Vuuren, M. & Winands, E.M.M. (2007). Iterative approximation of k-limited polling systems. *Queueing Systems*, 55(3), 161–178.

[21] Vishnevskii, V.M. & Semenova, O.V. (2006). Mathematical methods to study the polling systems. *Automation and Remote Control*, 67(2), 173-220.

[22] Vishnevskii, V.M. & Semenova, O.V. (2008). The power-series algorithm for two-queue polling system with impatient customers. *Proceedings of ICT 2008 – 15th International Conference on Telecommunications*, Saint-Petersburg, Russia, pp. 1–3.

[23] Vishnevskii, V.M. & Semenova, O.V. (2009). The power-series algorithm for $M/M/1$-type polling system with impatient customers. *Proceedings of EUROCON 2009 – International Conference on Computer as a Tool*, Saint-Petersburg, Russia, pp. 1915–1918.

[24] Winands, E.M.M., Adan, I.J.B.F., van Houtum, J., & Down, D.G. (2009). A state-dependent polling model with k-limited service. *Probability in the Engineering and Informational Sciences*, 23(2), 385–408.