ELSEVIER

European Journal of Operational Research 103 (1997) 595–609

EUROPEAN JOURNAL OF OPERATIONAL RESEARCH

Theory and Methodology

Dynamic server assignment in a two-queue model

O.J. Boxma [a,b,*], D.G. Down [a]

[a] *CWI, P.O. Box 94079, 1090 GB Amsterdam, The Netherlands*

[b] *Tilburg University, Faculty of Economics, P.O. Box 90153, 5000 LE Tilburg, The Netherlands*

Received 1 December 1995; accepted 1 September 1996

Abstract

We consider a polling model of two $M/G/1$ queues, served by a single server. The service policy for this polling model is of threshold type. Service at queue 1 is exhaustive. Service at queue 2 is exhaustive unless the size of queue 1 reaches some level T during a service at queue 2; in the latter case the server switches to queue 1 at the end of that service. Both zero- and nonzero switchover times are considered. We derive exact expressions for the joint queue length distribution at customer departure epochs, and for the steady-state queue-length and sojourn time distributions. In addition, we supply a simple and very accurate approximation for the mean queue lengths, which is suitable for optimization purposes. © 1997 Elsevier Science B.V.

Keywords: Queueing; Polling; ATM; Threshold service; Queue length distribution

1. Introduction

In this paper we consider a model of two $M/G/1$ queues, which are served by a single server. The service policy for this polling model is of threshold type. When the server is at queue 1, it serves its customers until the queue is empty (exhaustive service). When it has emptied queue 1, it switches to queue 2. While it visits queue 2, one of two events occurs first: queue 2 becomes empty or the size of queue 1 reaches some level T. In the latter case, the server switches to queue 1 immediately after having completed the service in which it is involved. Both zero- and nonzero switchover times will be considered. It is assumed that the server does not idle if there are customers present at either queue.

The Poisson arrival processes have rates λ_1, λ_2, and the service time distribution at queue i is $B_i(\cdot)$ with mean β_i, second moment $\beta_i^{(2)}$ and Laplace-Stieltjes Transform (LST) $B_i^*(\cdot)$, $i = 1, 2$. The traffic load at queue i is $\rho_i := \lambda_i \beta_i$, $i = 1, 2$. All interarrival times, service times and switchover times are assumed to be independent. The ergodicity condition is assumed to hold, as we restrict ourselves to steady-state behaviour. In the case of zero switchover times, the ergodicity condition is obviously satisfied iff the traffic load $\rho := \rho_1 + \rho_2 < 1$. The ergodicity condition is more complicated—and has some interesting features—when switchover times are nonzero; see Section 5. We are interested in the queue length and sojourn time distributions of this model,

* Corresponding author.

PII S0377-2217(96)00300-1

our ultimate goal being to obtain insight into the influence of thresholds on system performance, and into the quality of threshold policies for polling models.

In Section 2 we consider the case of zero switchover times. We use an analytic approach to obtain the joint steady-state queue length distribution at customer departure epochs. These results also lead to exact results for the marginal steady-state queue length and sojourn time distributions at both queues. The relative intricacy of the obtained exact results has led us to investigate a simple approximation of the mean queue lengths. Such an approximation is discussed in Section 3; it is sufficiently simple to be suitable for optimization purposes. Its accuracy is investigated in Section 4, where also exact numerical results are presented to indicate the effect of the threshold level on the mean queue lengths. In Section 5 we discuss the ergodicity condition for the case of nonzero switchover times, and we outline how the analysis of Section 2 can be adapted to handle this case. In the remainder of the present section we survey the related literature, and give our motivation for this study.

Some threshold-based polling systems have recently been proposed and analysed by Lee and Sengupta [18,19], Haverkort et al. [15], and Boxma et al. [4,5]. In [18] Lee considers a single-server two-queue model where the high priority queue is served exhaustively; the low priority queue receives k-limited service. In [19] a customer of each queue is served alternately unless the queue length of the high priority queue exceeds a certain threshold level; then only customers from that queue are served until its queue length is back to the threshold level. Haverkort et al. [15] analyze that same model using stochastic Petri nets; they also suggest and analyze a variant in which, once the threshold is exceeded, the server serves the high priority queue until it is empty (thus reducing the number of switches). The latter model is similar to the model of [4,5] and of the present paper, in which queues are served *exhaustively* unless a threshold level is reached. In [4] service times are exponentially distributed, and services at queue 2 are preemptively interrupted as soon as the threshold at queue 1 is reached. In [5] service times are again exponentially distributed, but the service process at queue 2 is *nonpreemptively* interrupted when the threshold at queue 1 is reached. The present study considers the model of [5] with nonpreemptive interruptions of the service process at the low priority queue, and with *general* service times at both queues. A preliminary version of the present paper, with a restriction to zero switchover times, has been presented in the conference paper [10].

The motivation for this work is two-fold. The first one is application-oriented. In modern telecommunication networks employing ATM (Asynchronous Transfer Mode) switching technology, a key problem is to be able to meet the quality-of-service requirements for different types of traffic. One way of accomplishing this is to assign different priorities to real-time traffic (voice, video) and nonreal-time traffic (data). The stringent delay requirements for real-time traffic dictate the assignment of a higher priority to it, but one would like to be able to meet those delay requirements while simultaneously giving the best possible service to nonreal-time traffic. Threshold-type service disciplines seem appropriate for this purpose; thus one would like to obtain insight into their performance.

A second motivation for the present study is the interesting feature that the server behaviour is, in a non-trivial way, not only determined by the situation at the queue that is presently being visited, but also by the situation at the other queue. The rich polling literature contains only a few papers (see in particular [4,5,11,16,17,19,24,26]) that take this possibility into consideration. Koole [17] considers a two-queue model with exponential service times. For all queue length combinations in a truncated state space he determines via dynamic programming whether the server should stay or switch to the other queue. The optimal switching curve appears not to have a simple form, but it is closely approximated by a threshold policy: the queue with the highest μc-value should be served exhaustively, and if the number of customers in that queue exceeds a certain threshold level, then it pays to switch to it when serving the other queue (note that this leads to our model, with exponential service times). More limited results, but for general service times, were obtained by Duenyas and Van Oyen [11]. For the same model, Reiman and Wein [24] arrive at a similar policy using heavy traffic analysis. Yadin [26] presents an exact analysis of several threshold policies, including the present one. However, he limits his discussion to the behaviour of the queue length process during one visit to a queue.

2. Zero switchover times

In this section we consider the case of zero switchover times. Let t_n denote the epoch of the nth service completion after $t = 0$, $n = 1,2,\ldots$. Let $X_{1,n}$ and $X_{2,n}$ be the numbers of customers (jobs) in queue 1 and queue 2, respectively, immediately after t_n. Let K_n be a random variable which is equal to k if the server is at queue k immediately before t_n, $k = 1,2$. Clearly, the stochastic process $\{(X_{1,n}, X_{2,n}, K_n),\ n = 1,2,\ldots\}$ forms an embedded Markov chain. For $n = 1,2,\ldots$, define its stationary probabilities (we assume that $\rho < 1$):

$$p_1(i,j) := \lim_{n\to\infty} \Pr\{X_{1,n} = i, X_{2,n} = j, K_n = 1\}, \quad i \geqslant 0,\ j \geqslant 0,$$

$$p_2(i,j) := \lim_{n\to\infty} \Pr\{X_{1,n} = i, X_{2,n} = j, K_n = 2\}, \quad i \geqslant 0,\ j \geqslant 0,$$

and the generating functions (GF)

$$\Pi_1(z_1,z_2) := \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} z_1^i z_2^j\, p_1(i,j), \quad |z_1| \leqslant 1,\ |z_2| \leqslant 1,$$

$$\Pi_2(z_1,z_2) := \sum_{i=T}^{\infty}\sum_{j=0}^{\infty} z_1^i z_2^j\, p_2(i,j), \quad |z_1| \leqslant 1,\ |z_2| \leqslant 1,$$

$$P_{2,i}(z_2) := \sum_{j=0}^{\infty} z_2^j p_2(i,j), \quad 0 \leqslant i \leqslant T-1,\ |z_2| \leqslant 1.$$

We shall consecutively derive the equilibrium equations for these GF's (Section 2.1), solve those equations (Section 2.2), and derive the steady-state marginal queue length distributions and the sojourn time distributions at both queues (Section 2.3).

2.1. Equilibrium equations

Introduce the following short-hand notation, for $i = 1,2$, $|z_1| \leqslant 1$, $|z_2| \leqslant 1$:

$$B_i^*(z_1,z_2) := B_i^*(\lambda_1(1-z_1) + \lambda_2(1-z_2)),$$

$$r_i := \frac{\lambda_i}{\lambda_1 + \lambda_2}.$$

By expressing the Markov chain $(X_{1,n}, X_{2,n}, K_n)$ into $(X_{1,n-1}, X_{2,n-1}, K_{n-1})$, letting $n \longrightarrow \infty$, and taking generating functions, one obtains the following relations, with $|z_1| \leqslant 1$, $|z_2| \leqslant 1$:

$$\Pi_1(z_1,z_2) = \frac{B_1^*(z_1,z_2)}{z_1}\Big[\Pi_1(z_1,z_2) - \Pi_1(0,z_2) + (\Pi_1(0,0) + P_{2,0}(0))r_1 z_1 + \Pi_2(z_1,z_2) + \sum_{j=1}^{T-1} P_{2,j}(0) z_1^j\Big], \tag{1}$$

$$\Pi_2(z_1,z_2) + \sum_{j=0}^{T-1} P_{2,j}(z_2) z_1^j = \frac{B_2^*(z_1,z_2)}{z_2}\Big[\Pi_1(0,z_2) - \Pi_1(0,0) + (\Pi_1(0,0) + P_{2,0}(0))r_2 z_2 + \sum_{j=0}^{T-1}(P_{2,j}(z_2) - P_{2,j}(0)) z_1^j\Big], \tag{2}$$

and for $0 \leqslant i \leqslant T-1$:

$$P_{2,i}(z_2) = \frac{1}{z_2} \int_0^\infty \left[(\Pi_1(0,0) + P_{2,0}(0)) r_2 z_2 e^{-\lambda_1 t} \frac{(\lambda_1 t)^i}{i!} + [\Pi_1(0,z_2) - \Pi_1(0,0)] e^{-\lambda_1 t} \frac{(\lambda_1 t)^i}{i!} \right.$$
$$\left. + \sum_{k=0}^{i} [P_{2,k}(z_2) - P_{2,k}(0)] e^{-\lambda_1 t} \frac{(\lambda_1 t)^{i-k}}{(i-k)!} \right] e^{-\lambda_2(1-z_2)t} \, dB_2(t). \tag{3}$$

Furthermore, all probabilities should sum up to one:

$$\Pi_1(1,1) + \Pi_2(1,1) + \sum_{j=0}^{T-1} P_{2,j}(1) = 1. \tag{4}$$

To assist the reader, let us interpret (1); Eqs. (2) and (3) can be interpreted similarly. The factor $B_1^*(z_1, z_2)$ is the GF of the joint distribution of the numbers of arrivals at queues 1 and 2 during one service at queue 1; division by z_1 accounts for the departure of the customer that received that service. Now consider the expression between the large brackets in (1). There are four possibilities that result in serving queue 1 between t_{n-1} and t_n:

(i) At t_{n-1} a service at queue 1 was completed, and queue 1 was not yet empty. $\Pi_1(z_1, z_2) - \Pi_1(0, z_2)$ is the GF of the joint distribution of the numbers of customers at queues 1 and 2 left behind after the previous service completion, *restricting oneself to those cases in which queue* 1 *is not empty.*

(ii) Both queues were empty at t_{n-1}, and the first new arrival took place at queue 1. The corresponding term is $(\Pi_1(0,0) + P_{2,0}(0)) r_1 z_1$.

(iii) At t_{n-1} a service at queue 2 was completed, and in queue 1 the threshold had been reached: $\Pi_2(z_1, z_2)$.

(iv) At t_{n-1} a service at queue 2 was completed. The threshold in queue 1 had not been reached, but queue 1 was not empty whereas queue 2 was: $\sum_{j=1}^{T-1} P_{2,j}(0) z_1^j$.

2.2. Solution

From (1):

$$\Pi_1(z_1, z_2) \frac{z_1 - B_1^*(z_1, z_2)}{z_1} = \frac{B_1^*(z_1, z_2)}{z_1} \left[-\Pi_1(0, z_2) + \Pi_2(z_1, z_2) \right.$$
$$\left. + (\Pi_1(0,0) + P_{2,0}(0)) r_1 z_1 + \sum_{j=1}^{T-1} P_{2,j}(0) z_1^j \right]. \tag{5}$$

Observe that $z_1 - B_1^*(z_1, z_2)$ has exactly one zero $z_1 = \mu(z_2)$ in $|z_1| \leqslant 1$, for every $|z_2| \leqslant 1$. Here $\mu(z_2) = \mathrm{E}[e^{-\lambda_2(1-z_2)G_1}]$, with G_1 a stochastic variable with distribution the busy period distribution of queue 1 in isolation. This is a well known fact in queueing theory: see for example Ch. II.4 of [7]. Since the GF $\Pi_1(z_1, z_2)$ is analytic in $|z_1| \leqslant 1$, $|z_2| \leqslant 1$, the right-hand side of (5) should be zero for $z_1 = \mu(z_2)$:

$$-\Pi_1(0, z_2) + (\Pi_1(0,0) + P_{2,0}(0)) r_1 \mu(z_2) + \Pi_2(\mu(z_2), z_2) + \sum_{j=1}^{T-1} P_{2,j}(0) \mu^j(z_2) = 0. \tag{6}$$

Substituting (2) into (6), with $z_1 = \mu(z_2)$, gives a relation between $\Pi_1(0, z_2)$ and all $P_{2,j}(z_2)$:

$$\Pi_1(0,z_2) = (\Pi_1(0,0) + P_{2,0}(0))r_1\mu(z_2) - \sum_{j=1}^{T-1}(P_{2,j}(z_2) - P_{2,j}(0))\mu^j(z_2)$$

$$-P_{2,0}(z_2) + \frac{B_2^*(\mu(z_2),z_2)}{z_2}\Bigg[\Pi_1(0,z_2) - \Pi_1(0,0)$$

$$+(\Pi_1(0,0) + P_{2,0}(0))r_2z_2 + \sum_{j=0}^{T-1}(P_{2,j}(z_2) - P_{2,j}(0))\mu^j(z_2)\Bigg].$$

Bringing the $\Pi_1(0,z_2)$ terms to the left-hand side and dividing both sides by $1 - B_2^*(\mu(z_2),z_2)/z_2$ gives:

$$\Pi_1(0,z_2) = -\sum_{j=0}^{T-1}(P_{2,j}(z_2) - P_{2,j}(0))\mu^j(z_2)$$

$$+\frac{(\Pi_1(0,0) + P_{2,0}(0))r_1z_2\mu(z_2) - z_2P_{2,0}(0)}{z_2 - B_2^*(\mu(z_2),z_2)}$$

$$+\frac{B_2^*(\mu(z_2),z_2)[(\Pi_1(0,0) + P_{2,0}(0))r_2z_2 - \Pi_1(0,0)]}{z_2 - B_2^*(\mu(z_2),z_2)}. \tag{7}$$

If we put the right-hand side of (7) over the common denominator $z_2 - B_2^*(\mu(z_2),z_2)$, then as $z_2 = 1$ is a zero of this denominator, the analyticity of $\Pi_1(0,z_2)$ in $|z_2| \leqslant 1$ implies that the numerator is also zero for $z_2 = 1$, which yields:

$$\Pi_1(0,0) + P_{2,0}(0) = 1 - \rho. \tag{8}$$

This result is consistent with the observation that the whole system is empty with probability $1 - \rho$.

Introducing, for $0 \leqslant i \leqslant T-1$, $|z| \leqslant 1$,

$$c_i(z) := \int_0^\infty e^{-\lambda_1 t}\frac{(\lambda_1 t)^i}{i!}e^{-\lambda_2(1-z)t}\,dB_2(t),$$

we can rewrite (3) as follows: for $|z_2| \leqslant 1$, $0 \leqslant i \leqslant T-1$,

$$z_2P_{2,i}(z_2) = (1-\rho)r_2z_2c_i(z_2)$$

$$+[\Pi_1(0,z_2) - \Pi_1(0,0)]c_i(z_2) + \sum_{k=0}^{i}[P_{2,k}(z_2) - P_{2,k}(0)]c_{i-k}(z_2). \tag{9}$$

Express $\Pi_1(0,z_2)$ into the $P_{2,i}(z_2)$, using (7), and $\Pi_1(0,0)$ into $P_{2,i}(0)$, using (8); then (9) gives us T equations for the T unknown functions $P_{2,i}(z_2)$, $0 \leqslant i \leqslant T-1$. The structure of the equations shows that one can write:

$$A(z)P(z) = F(z), \tag{10}$$

with $P(z)$ denoting the column vector with elements $(P_{2,0}(z),\ldots,P_{2,T-1}(z))$; the matrix $A(z)$ has elements

$$A_{ij}(z) = \begin{cases} z - c_0(z) + c_i(z)\mu^i(z), & j = i, \\ c_i(z)\mu^j(z), & j > i, \\ c_i(z)\mu^j(z) - c_{i-j}(z), & j < i, \end{cases}$$

for $i, j = 0, 1, \ldots, T-1$, and $F(z)$ is a column vector with elements

$$F_i(z) = \frac{c_i(z)}{z - B_2^*(\mu(z), z)}\{(1-\rho)r_2 z^2 + z(1-\rho)r_1\mu(z) - z(1-\rho)\}$$
$$- \sum_{k=0}^{i} P_{2,k}(0)c_{i-k}(z) + c_i(z)\sum_{k=0}^{T-1} P_{2,k}(0)\mu^k(z),$$

for $i = 0, 1, \ldots, T-1$.

In solving (10) we have to pay special attention to the zeros of $\det(A(z))$ inside the unit circle $|z| = 1$. Suppose that there are K such zeros, $\zeta_1, \ldots, \zeta_K$ (in fact the first column of $A(z)$ immediately reveals that $\zeta_1 = 0$).

Call $A_i(z)$ the matrix obtained from $A(z)$ by replacing the ith column by the column vector $F(z)$. According to Cramer's rule, $P_{2,i}(z) = \det(A_i(z))/\det(A(z))$, $i = 0, \ldots, T-1$. The analyticity of $P_{2,i}(z)$ in $|z| \leqslant 1$ implies that

$$\det(A_i(\zeta_j)) = 0, \tag{11}$$

for $j = 1, \ldots, K$. Using (11), each zero ζ_j yields one equation relating the T unknown $P_{2,i}(0)$. (For the equation due to ζ_1, an application of l'Hospital's rule is required.) We shall argue that when $\rho < 1$, a set of T independent equations results, yielding the unique solution for the constants $P_{2,i}(0)$. Indeed, the Kolmogorov equations for the equilibrium distribution of the Markov chain $\{(X_{1,n}, X_{2,n}, K_n),\ n = 1, 2, \ldots\}$, along with the normalizing equation (4), have a unique absolutely convergent solution when $\rho < 1$; and using generating functions, we have transformed those Kolmogorov equations plus the normalizing equation into the T-dimensional matrix equation (10), plus the relations (7) and (8). If $K = T$, then as there exists a unique solution, the equations generated by (11) must be independent. Now suppose that $K < T$. Then we would obtain too few equations to determine all T unknown constants uniquely, and we would find multiple solutions for them—which is impossible. Finally, if $K > T$, then we would find too many equations for the T unknowns. Once again, as it is known that there is a unique solution, there must be exactly T independent equations amongst those derived using (11).

Remark 1. It is in principle possible that there are more than T zeros, but that the ensuing linear equations for the $P_{2,i}(0)$ are dependent; we expect that one can prove that this is not possible, but we have not yet done so. It is suspected that the proof of this may follow using techniques similar to those in Cohen and Down [8]. For related approaches to a similar problem, the reader is referred to Mitrani and Mitra [21], Neuts [22] and Gail et al. [13]. Regardless, this poses no difficulties for numerical analysis; in fact, in our numerical experiments we have yet to encounter a case with other than exactly T zeros of $\det(A(z))$ in $|z| < 1$.

Remark 2. When $T = 1$, this model is equivalent to the well-known M/G/1 queue with two customer classes and nonpreemptive priority for class 1. Indeed, our results for that case can be shown to agree with those in the literature (cf. [7], Section III.3.8)). When $T = \infty$, the model reduces to the classical two-queue model with exhaustive service at both queues, cf. Takács [25].

2.3. *The steady-state queue length and sojourn time distributions*

Clearly $G_1(z) := \Pi_1(z, 1)/r_1$ is the GF of the queue length distribution at queue 1 at customer departure epochs from that queue; similarly, $G_2(z) := [\sum_{i=0}^{T-1} P_{2,i}(z) + \Pi_2(1, z)]/r_2$ is the GF of the queue length distribution at queue 2 at customer departure epochs from that queue. An up-and-downcrossing argument implies that these are also the queue length distributions at customer arrival epochs at these queues; and

PASTA finally implies that they are also the marginal steady-state queue length distributions (but note that, unlike [4,5] which consider exponential service times, we have not obtained the *joint* steady-state queue length distribution). The mean steady-state queue length at queue i, to be denoted by EX_i, is easily obtained from $G_i(z)$ by differentiation at $z = 1$. In order to extract numerical results for queue length moments, or even for the distributions themselves, from the obtained GF's, one may also take recourse to algorithms such as recently have been developed in [6] and [1].

Having obtained $G_i(z)$, one can subsequently argue—as in the standard M/G/1 FCFS queue—that the number of customers left behind at queue i by a departing customer is exactly the number of arrivals at that queue during the sojourn time of this customer; hence, with T_i a stochastic variable with distribution the sojourn time distribution at queue i,

$$E[e^{-\omega T_i}] = G_i(1 - \omega/\lambda_i).$$

3. An approximation for the mean waiting times

In the previous section the two-queue threshold model has been analysed exactly, but this has not led to simple expressions for performance measures like the mean queue lengths and mean waiting times; they are expressed in T zeros that have to be determined numerically. For optimization purposes (e.g., choose T such that a weighted sum of the mean waiting times is minimized) it is important to have a simple, yet accurate, approximation for such performance measures. The goal of this section is to derive such an approximation for the mean waiting time $EW_1^{(T)}$, the superscript indicating the dependence on T. The conservation law

$$\rho_1 EW_1^{(T)} + \rho_2 EW_2^{(T)} = \rho \frac{\lambda_1 \beta_1^{(2)} + \lambda_2 \beta_2^{(2)}}{2(1-\rho)}, \tag{12}$$

(cf. Gelenbe and Mitrani [14], Ch. 6) then immediately yields $EW_2^{(T)}$, and the mean queue lengths follow using Little's formula.

In two special cases a detailed solution of the two-queue threshold model is known: $T = 1$ and $T = \infty$. For $T = 1$, the model reduces to a single server with two customer types and nonpreemptive priority for type 1. It is well-known (see e.g. Cohen [7], p. 458) that

$$EW_1^{(1)} = \frac{\sum_{i=1}^{2} \lambda_i \beta_i^{(2)}}{2(1-\rho_1)}.$$

For $T = \infty$, a model with exhaustive service at both queues results. The mean waiting time at queue 1 is now given by (cf. Takács [25])

$$EW_1^{(\infty)} = \frac{\lambda_1 \beta_1^{(2)}}{2(1-\rho_1)} + \frac{\lambda_1 \rho_2^2 \beta_1^{(2)} + \lambda_2 (1-\rho_1)^2 \beta_2^{(2)}}{2(1-\rho_1)(1-\rho)(1-\rho+2\rho_1\rho_2)}. \tag{13}$$

Our numerical experiments (cf. Section 4) suggest that $EW_1^{(T)}$ is monotonically decreasing in T. We have not been able to prove this intuitively obvious result. The experiments also suggest that $EW_1^{(T)}$ approaches $EW_1^{(\infty)}$ in exponential fashion. The geometric tail behaviour of the queue length distribution in the ordinary M/G/1 queue makes it likely that the probability of the queue length at queue 1 exceeding $T - 1$ decreases exponentially in T, which in its turn suggests an exponential behaviour of $EW_1^{(T)}$ in T. Hence we propose the following approximation for $EW_1^{(T)}$:

$$EW_1^{(T)} = EW_1^{(\infty)} + e^{-c(T-1)}(EW_1^{(1)} - EW_1^{(\infty)}). \tag{14}$$

Here the constant c still has to be specified. This is done by approximating $EW_1^{(T)}$ for $T = 2$. Hereto we reason as follows. Consider the arrival of an arbitrary customer C at queue 1. With probabilities $1-\rho$, ρ_1 and ρ_2 it finds the server idle, at queue 1 and at queue 2, respectively. We can write

$$EW_1^{(2)} = (1-\rho)0 + \rho_1 \frac{\beta_1^{(2)}}{2\beta_1} + \rho_2 EZ + EX_1^{\text{wait}}\beta_1.$$

Here EZ is the mean time from C's arrival until the server returns to queue 1, in the case that the server was at queue 2. The last term of the formula takes into account that all X_1^{wait} customers that are waiting at queue 1 will be served *before* C. Applying Little's formula to that last term, and splitting Z into the residual time of the service in queue 2 during which C arrived and the time U until the server subsequently switches to queue 1, we can write (compare with the expression for $EW_1^{(1)}$ above):

$$EW_1^{(2)} = \frac{\sum_{i=1}^{2} \lambda_i \beta_i^{(2)} + 2\rho_2 EU}{2(1-\rho_1)}.$$

Now observe that $U = 0$ unless all the following three events occur ($\tilde{B}_2$ denotes the service at queue 2 during which C arrived)

- E_1: at the start of $\tilde{B}_2$, queue 1 is empty;
- E_2: during $\tilde{B}_2$, apart from C nobody arrives at queue 1;
- E_3: after $\tilde{B}_2$, queue 2 is not yet empty.

We can write:

$$EU = \Pr\{E_1\} \Pr\{E_2\} E[UI_{E_3} \mid E_1, E_2],$$

where I_{E_3} denotes the indicator function of the event E_3. From renewal theory,

$$\Pr\{E_2\} = -B_2^{*'}(\lambda_1)/EB_2. \tag{15}$$

Reasoning that, for $T = 2$, queue 1 behaves quite similarly to an M/G/1 queue in isolation, we approximate $\Pr\{E_1\}$ by the probability that queue 1 in isolation is empty, given that it has at most one customer. M/G/1 theory then yields:

$$\Pr\{E_1\} \approx B_1^*(\lambda_1). \tag{16}$$

We apply a more bold approximation for the conditional expectation of UI_{E_3}. We interpolate between the case of light traffic at queue 2 (in which case this conditional expectation is almost zero) and the case of heavy traffic at queue 2 (in which case this conditional expectation equals $1/\lambda_1 + \beta_2^{(2)}/2\beta_2$):

$$E[UI_{E_3} \mid E_1, E_2] \approx \rho_2 [\frac{1}{\lambda_1} + \frac{\beta_2^{(2)}}{2\beta_2}].$$

Combination of the above formulas leads to the following approximation:

$$EW_1^{(2)} \approx \frac{\sum_{i=1}^{2} \lambda_i \beta_i^{(2)} - 2\rho_2 \lambda_2 B_2^{*'}(\lambda_1) B_1^*(\lambda_1)[1/\lambda_1 + \beta_2^{(2)}/2\beta_2]}{2(1-\rho_1)}. \tag{17}$$

Substitution in (14) yields c, and hence an approximation for $EW_1^{(T)}$. The conservation law (12) then gives $EW_2^{(T)}$.

Table 1
Mean numbers of customers for exponential service times. $\lambda_1 = \lambda_2 = 1; \beta_1 = \beta_2 = 1/3$

T	EX_1	EX_2	EX_1^{approx}
1	.6667	1.3333	
2	.7522	1.2478	.7604
3	.8254	1.1746	.8278
4	.8795	1.1205	.8762
5	.9176	1.0824	.9110
10	1.0000	1.0000	.9829

Table 2
Mean numbers of customers for deterministic service times. $\lambda_1 = \lambda_2 = 1; \beta_1 = \beta_2 = 1/3$

T	EX_1	EX_2	EX_1^{approx}
1	.5000	.8333	
2	.5757	.7577	.5820
3	.6214	.7119	.6237
4	.6447	.6887	.6448
5	.6561	.6773	.6555
10	.6667	.6667	.6610

4. Numerical results

To illustrate the procedure presented in Section 2, and to test the simple approximation that was proposed in the previous section, we examine some particular models for varying threshold values. We remark beforehand that the calculation of the numerical values for the actual system may be handled in a relatively straightforward manner. The main difficulty is in efficiently finding the zeros of $\det(A(z))$ inside $|z| = 1$, to a sufficiently high degree of accuracy, particularly for large values of T. It was found that Muller's algorithm [23] worked satisfactorily for the models in this section.

In our first case both service time distributions are exponential. The parameter values are $\lambda_1 = \lambda_2 = 1$ and $\beta_1 = \beta_2 = 1/3$. The expected number of customers at each queue is given in columns 2 and 3 of Table 1. This case is also studied in [4]; the results for the same case there are seen to agree with the results in Table 1. The numerical analysis in our case is somewhat more involved; in [4,5] all zeros can be determined explicitly, from T *separate* equations. The fourth column gives the approximate values for the mean number of customers at queue 1 using the results of the previous section (of course we have applied Little's formula to obtain mean queue lengths from mean waiting times). The entry for $T = 1$ is not given, as it was calculated exactly and then used in the approximation.

The second case differs from the first one only in the choice of service time distributions; these are degenerate (constant service times at both queues). The expected number of customers at each queue is given in Table 2, along with values obtained using the approximation of the previous section. The result for $T = 1$ has been checked against formulas given in Section III.3.8 of [7] for the M/D/1 queue with non-preemptive priority.

The extremely close agreement between the approximate and actual expected number of customers is remarkable. It indicates the usefulness of this simple approximation for examining the effects of varying T on system behaviour.

Tables 3 and 4 present results for variants of the case of Table 1: exponential service times but a higher mean service time at queue 1 or queue 2, leading to a heavily loaded system. Larger values of T had to be considered, and the approximation was somewhat less accurate. Furthermore, the effect of T on the mean queue

Table 3
Mean numbers of customers for exponential service times. $\lambda_1 = \lambda_2 = 1; \beta_1 = 1/2$ and $\beta_2 = 1/3$

T	EX_1	EX_2	EX_1^{approx}
1	1.2222	4.6667	
2	1.3999	4.4001	1.3333
3	1.5669	4.1497	1.4333
4	1.7084	3.9374	1.5233
5	1.8252	3.7622	1.6043
10	2.1543	3.2351	1.9028
15	2.2759	3.0873	2.0791
20	2.3328	3.0004	2.1832
30	2.3333	3.0000	2.2810

Table 4
Mean numbers of customers for exponential service times. $\lambda_1 = \lambda_2 = 1; \beta_1 = 1/3$ and $\beta_2 = 1/2$

T	EX_1	EX_2	EX_1^{approx}
1	0.8750	3.7500	
2	1.0801	3.6133	1.0625
3	1.2971	3.4686	1.2335
4	1.4994	3.3337	1.3893
5	1.6815	3.2124	1.5314
10	2.3188	2.7912	2.0747
15	2.6671	2.5679	2.4169
20	2.8114	2.4669	2.6326
30	3.0000	2.3333	2.8541

Table 5
Mean numbers of customers for exponential service times. $\lambda_1 = 1$ and $\lambda_2 = .5; \beta_1 = \beta_2 = 1/4$

T	EX_1	EX_2	EX_1^{approx}
1	.3750	.4500	
2	.3824	.4351	.3883
3	.3871	.4258	.3905
4	.3893	.4214	.3908
5	.3903	.4194	.3909
10	.3909	.4182	.3909

length was considerable.

Tables 5–7 present results for cases with exponential service times and varying λ_2. Again the approximation is good, the worst results occurring for the highest load ($\rho = 5/8$ in Table 7).

We have also experimented with different choices of service time distributions (results available from the authors upon request). For example, we have considered the case of Tables 1 and 2, but with exponential service times at queue 1 and an E_2, D or H_2 distribution at queue 2. The approximation is of comparable quality as in Tables 1 and 2; furthermore, as one might expect, all mean queue lengths appear to increase with increasing coefficient of variation of the service time distribution. As the exact results for $T = 1$ and $T = \infty$ (cf. (13)) suggest, the dependence on that coefficient of variation is rather weak for most parameter combinations.

In all cases that we have investigated, the expected number of customers at the high priority queue is

Table 6
Mean numbers of customers for exponential service times. $\lambda_1 = \lambda_2 = 1; \beta_1 = \beta_2 = 1/4$

T	EX_1	EX_2	EX_1^{approx}
1	.4167	.5833	
2	.4494	.5506	.4700
3	.4722	.5278	.4892
4	.4853	.5147	.4961
5	.4924	.5076	.4986
10	.4979	.5003	.5000
15	.5000	.5000	.5000

Table 7
Mean numbers of customers for exponential service times. $\lambda_1 = 1$ and $\lambda_2 = 1.5; \beta_1 = \beta_2 = 1/4$

T	EX_1	EX_2	EX_1^{approx}
1	.4583	.8056	
2	.5396	.7514	.5783
3	.6027	.7093	.6437
4	.6455	.6808	.6794
5	.6734	.6622	.6988
10	.7173	.6329	.7210
15	.7217	.6299	.7221
20	.7222	.6297	.7222
30	.7222	.6296	.7222

increasing in T, as one would expect. At a sufficiently large value of T (which is larger when the load is larger), both systems are essentially behaving as one in which exhaustive service occurs at both queues. This indicates that in considering the design of threshold policies it is probably in most cases sufficient to consider "small" values of T.

Remark 3. The results in the tables satisfy the conservation law as expressed in Formula (12).

5. Nonzero switchover times

In this section we consider the same model as in Section 2, but with nonzero switchover times between the queues. The switchover times from queue 1 to queue 2 are denoted by $S_k^{(1)}$; their distribution is denoted by $S_1(\cdot)$, with Laplace-Stieltjes transform $S_1^*(\cdot)$, and the mean switchover time by σ_1. Similarly for the switchover times from queue 2 to queue 1, with the index 1 replaced by 2. We make all the usual independence assumptions for the switchover times. We also assume the following. When the server is switching to queue 2, but the queue length at queue 1 reaches T before the server has arrived at queue 2, then the server will do one service in queue 2 before going back.

We must also specify the server behaviour in a completely empty system. For the moment we adopt the common assumption in the polling literature, i.e., the server keeps cycling in an empty system (but see the interesting recent studies [2,12]). The ergodicity conditions, which will be studied first, are obviously not affected by the server behaviour in an empty system.

5.1. Ergodicity conditions

In comparison to the model without switchover times, the conditions on the system parameters that imply ergodicity of the system are nontrivial when switchover times can occur. The reason for this is demonstrated in the outline of the proof of Theorem 4 below, which provides the ergodicity condition for the 2-queue model with threshold switching and switchover times. Let

$$\tau := \text{argmin}_{n\geqslant 1}\left\{ S_k^{(1)} + \sum_{k=1}^{n} B_k^{(2)} > t \mid t = \sum_{k=1}^{T} A_k^{(1)} \right\}, \tag{18}$$

where $S_k^{(1)}$ is a generic switchover time from queue 1 to queue 2, $A_k^{(1)}$ and $B_k^{(2)}$ are, for each k, generic interarrival times and service times for queue 1 and queue 2, respectively.

Theorem 4. *The system under consideration is ergodic if*

$$\mathsf{E}[\tau](1-\rho) > \lambda_2(\sigma_1+\sigma_2). \tag{19}$$

Proof. (Outline)
An outline is given that essentially demonstrates the physical meaning of the condition (19). Consider the case where there is a 'large' number of customers at queue 2. Trivially, if $\rho_1 \geqslant 1$ the system is not stable. If $\rho_1 < 1$, it is not difficult to see that queue 1 empties infinitely often. Thus, we may define a cycle as the length of time elapsed between two successive times at which queue 1 empties. The expected value of the cycle time is given by the following elementary calculation:

$$\sigma_1 + \mathsf{E}[\tau]\beta_2 + \sigma_2 + \frac{\lambda_1\beta_1(\sigma_1 + \mathsf{E}[\tau]\beta_2 + \sigma_2)}{1-\lambda_1\beta_1} = \frac{\sigma_1 + \mathsf{E}[\tau]\beta_2 + \sigma_2}{1-\lambda_1\beta_1}. \tag{20}$$

As the rate at which work arrives at queue 2 is $\lambda_2\beta_2$ over the entire cycle, and work is depleted at rate 1 over a period with expected length $\mathsf{E}[\tau]\beta_2$, ergodicity is thus ensured if

$$\mathsf{E}[\tau]\beta_2 > \lambda_2\beta_2\frac{\sigma_1 + \mathsf{E}[\tau]\beta_2 + \sigma_2}{1-\lambda_1\beta_1}. \tag{21}$$

Rearrangement of (21) yields the required ergodicity condition. □

Remark 5. The above argument may be made precise using tools developed in Dai [9]. In fact, using results of Meyn [20], we may also state that if $\mathsf{E}[\tau](1-\rho) < \lambda_2(\sigma_1+\sigma_2)$ then the system is transient.

Remark 6. Note that $\mathsf{E}[\tau]$, and hence the ergodicity condition, may depend on other than first moments of the underlying random variables. It is also interesting to observe that, for fixed ρ_1, a relatively large value of λ_1 will give rise to relatively many switchovers and to a relatively small value of $\mathsf{E}[\tau]$.

Remark 7. Note that the ergodicity condition (19) can also be written as $\lambda_2(\sigma_1+\sigma_2)/(1-\rho) < \mathsf{E}[\tau]$. In the ergodic case, and with a server that keeps cycling in an empty system, $(\sigma_1+\sigma_2)/(1-\rho)$ denotes the mean time between two successive visits to, say, queue 2. Hence (19) for this case states that the mean number of arrivals at queue 2 between two such visits should be less than $\mathsf{E}[\tau]$. This makes sense, as $\mathsf{E}[\tau]$ has the following interpretation: if queue 2 always has customers available, then $\mathsf{E}[\tau]$ of them can be served on average before the server is forced to switch back to queue 1.

5.2. The steady-state queue length distribution

It turns out that the analysis of the model with switchover times presents no new difficulties, once ergodicity has been deduced. In this section, the problem will be simply set up for $T = 1$ (the counterparts of equations (1) and (2) will be given), with equations for $\Pi_1(0, z_2)$ and $P_{2,0}(z_2)$ being stated. These will be combined to give an expression for $P_{2,0}(z_2)$. Finally, a matrix equation equivalent to (10) will be given for general T. The details for the model in full generality have been developed, but we have decided not to present them; the work is conceptually simple, albeit algebraically complex.

Consider the situation in which in an empty system the server idles at the most recently served queue (note that other situations are also easily handled). In this case, we may write:

$$\Pi_1(z_1, z_2) = \frac{B_1^*(z_1, z_2)}{z_1}[\Pi_1(z_1, z_2) - \Pi_1(0, z_2) + \Pi_1(0,0) r_1 z_1 + S_2^*(z_1, z_2)\Pi_2(z_1, z_2) + S_2^*(z_1, z_2) P_{2,0}(0) r_1 z_1], \tag{22}$$

$$\Pi_2(z_1, z_2) + P_{2,0}(z_2) = \frac{B_2^*(z_1, z_2)}{z_2}[P_{2,0}(z_2) - P_{2,0}(0) + P_{2,0}(0) r_2 z_2 + S_1^*(z_1, z_2)[\Pi_1(0, z_2) - \Pi_1(0,0)] + S_1^*(z_1, z_2)\Pi_1(0,0) r_2 z_2], \tag{23}$$

where

$$S_i^*(z_1, z_2) := S_i^*(\lambda_1(1 - z_1) + \lambda_2(1 - z_2)).$$

Following the same argument as in Section 2.2 (in particular, the function $\mu(z_2)$ is unchanged), we may write:

$$\Pi_1(0, z_2) = \Pi_1(0,0) r_1 \mu(z_2) + S_2^*(\mu(z_2), z_2) P_{2,0}(0) r_1 \mu(z_2) - S_2^*(\mu(z_2), z_2) P_{2,0}(z_2) + \frac{S_2^*(\mu(z_2), z_2) B_2^*(\mu(z_2), z_2)}{z_2}[P_{2,0}(z_2) - P_{2,0}(0) + P_{2,0}(0) r_2 z_2 + S_1^*(\mu(z_2), z_2)\Pi_1(0,0) r_2 z_2 + S_1^*(\mu(z_2), z_2)(\Pi_1(0, z_2) - \Pi_1(0,0))], \tag{24}$$

and

$$P_{2,0}(z_2) = \frac{1}{z_2}[P_{2,0}(0) r_2 z_2 c_i(z_2) + (P_{2,0}(z_2) - P_{2,0}(0)) c_0(z_2) + (\Pi_1(0, z_2) - \Pi_1(0,0)) d_0(z_2) c_0(z_2) + \Pi_1(0,0) r_2 z_2 d_0(z_2) c_0(z_2)], \tag{25}$$

where

$$d_j(z_2) := \int_0^\infty e^{-\lambda_1 t} \frac{(\lambda_1 t)^j}{j!} e^{-\lambda_2(1 - z_2)t} \, dS_1(t). \tag{26}$$

So, combining (24) and (25), we find

$$\frac{z - c_0(z) + d_0(z) c_0(z) S_2^*(\mu(z), z)[z - B_2^*(\mu(z), z))]}{z - S_1^*(\mu(z), z) S_2^*(\mu(z), z) B_2^*(\mu(z), z)} P_{2,0}(z) = F_0(z),$$

where $F_0(z)$ is a known function of z.

The above calculations may be performed for general T, where the following system of equations results (cf. (10)):

$$A(z)P(z) = F(z),$$

where

$$A_{ij}(z) = \begin{cases} z - c_0(z) + G_{ij}(z), & j = i, \\ c_{i-j}(z) + G_{ij}(z), & j < i, \\ G_{ij}(z), & j > i, \end{cases}$$

with

$$G_{ij}(z) := \frac{(\sum_{l=0}^{i} d_l(z) c_{i-l}(z)) S_2^*(\mu(z), z)[z - B_2^*(\mu(z), z))]\mu^j(z)}{z - S_1^*(\mu(z), z) S_2^*(\mu(z), z) B_2^*(\mu(z), z)}, \tag{27}$$

and $F(z)$ a vector containing known functions of z. Given that the system is ergodic (i.e., (19) holds), the matrix $A(z)$ once again has T zeros inside $|z| \leqslant 1$ and the elements $P_{2,i}(z)$ may be determined by solving the T equations that result.

5.3. Pseudoconservation laws

When there are *no* switchover times in a polling system, there is *conservation of work* which leads to *conservation laws* (cf. Formula (12)). In a polling model *with* switchover times, the principle of work conservation is replaced by the principle of *work decomposition* [3]: the amount of work in the system is in distribution equal to the sum of two independent quantities, viz. the amount of work in the corresponding system without switchover times and the amount of work at an arbitrary moment during a switchover. In its turn, this work decomposition result leads to a so-called *pseudoconservation law*—an exact expression for a weighted sum of the mean waiting times. When the server stops at some queue in the case of an empty system, this pseudoconservation law can be adapted, as is done in full generality in [2]. We restrict ourselves here to the somewhat easier case in which the server keeps cycling in an empty system. The pseudoconservation law then takes the following form [3]:

$$\rho_1 \mathrm{E}W_1^{(T)} + \rho_2 \mathrm{E}W_2^{(T)} = \rho \frac{\lambda_1 \beta_1^{(2)} + \lambda_2 \beta_2^{(2)}}{2(1-\rho)} + \rho \frac{s^{(2)}}{2(\sigma_1 + \sigma_2)} + \frac{\rho_1 \rho_2 (\sigma_1 + \sigma_2)}{1-\rho} + \mathrm{E}Z_1 + \mathrm{E}Z_2, \tag{28}$$

where $s^{(2)}$ denotes the second moment of $S_k^{(1)} + S_k^{(2)}$, and where Z_i denotes the amount of work left behind at queue i at the end of a server visit to that queue; this is the only quantity that depends on the service discipline at the queues. In the present system, with exhaustive service at queue 1, we have $Z_1 \equiv 0$; the only unknown in the right-hand side of (28), $\mathrm{E}Z_2$, can either be directly determined from the results in Section 5.2 for $\Pi_2(1, z_2)$, or from (28) and the results for the mean queue lengths cq. waiting times.

Acknowledgements

The authors gratefully acknowledge a discussion with S. Foss about the ergodicity conditions for the model with switchover times. The second author was supported by an ERCIM postdoctoral fellowship.

References

[1] Abate, J., and Whitt, W. (1992), "Numerical inversion of probability generating functions", *Operations Research Letters* 12, 245–251.

[2] Borst, S.C., "A pseudoconservation law for a polling system with a dormant server", in: Labetoulle, J., and Roberts J.W., eds., *The Fundamental Role of Teletraffic in the Evolution of Telecommunications Networks*, North-Holland Publ. Cy., Amsterdam, 729–742.

[3] Boxma, O.J., and Groenendijk, W.P. (1987), "Pseudoconservation laws in cyclic service systems", *Journal of Applied Probability*, 124, 949–964.
[4] Boxma, O.J., Koole, G., and Mitrani, I. (1995), "A two-queue polling model with a threshold service policy", in: Dowd, P., and Gelenbe, E., eds., *Proceedings MASCOTS '95*, IEEE Computer Society Press, Los Alamitos, CA, 84–89.
[5] Boxma, O.J., Koole, G., and Mitrani, I. (1995), "Polling models with threshold switching", in: Baccelli, F., Jean-Marie, A., and Mitrani, I., eds., *Quantitative Methods in Parallel Systems*, Springer Verlag, Berlin, 129–140.
[6] Choudhury, G.L., and Lucantoni, D.M. (1996), "Numerical computation of the moments of a probability distribution from its transform", *Operations Research* 44, 368–381.
[7] Cohen, J.W. (1982), *The Single Server Queue*, North-Holland Publ. Cy., Amsterdam.
[8] Cohen, J.W., and Down, D.G. (1995), "On the role of Rouché's theorem in queueing theory analysis", CWI Report BS-R9523, to appear in *Queueing Systems*.
[9] Dai, J.G. (1995), "On positive Harris recurrence of multiclass queueing networks: A unified approach via fluid limit models", *Ann. Appl. Probab.* 5, 49–77.
[10] Down, D.G., and Boxma, O.J. (1995), "A two-queue polling model with threshold switching", in: *Proceedings of the Nordic Teletraffic Seminar NTS-12*, Espoo, Finland.
[11] Duenyas, I., and Van Oyen, M.P. (1993), "Stochastic scheduling of parallel queues with set-up costs", Technical Report 93-09, Northwestern University.
[12] Eisenberg, M. (1994), "The polling system with a stopping server", *Queueing Systems* 18, 387–431.
[13] Gail, H.R., Hantler, S.L., and Taylor, B.A. (1996), "Spectral analysis of $M/G/1$ and $G/M/1$ type Markov chains", *Advances in Applied Probability* 28, 114–165.
[14] Gelenbe, E., and Mitrani, I. (1980), *Analysis and Synthesis of Computer Systems*, Academic Press, London.
[15] Haverkort, B., Idzenga, H.P., and Kim, B.G. (1994), "Performance evaluation of ATM cell scheduling policies using Stochastic Petri Nets", Report University of Twente, Tele-Informatics and Open Systems Group TIOS 94-19.
[16] Hofri, M., and Ross, K.W. (1987), "On the optimal control of two queues with server setup times and its analysis", *SIAM Journal on Computing* 16, 399–420.
[17] Koole, G.M. (1994), "Assigning a single server to inhomogeneous queues with switching costs", CWI Report BS-R9405, to appear in *Theoretical Computer Science*, 182 (1997).
[18] Lee, D.-S. (1993), "A two-queue model with exhaustive and limited service disciplines", Report C & C Research Laboratories, NEC USA Inc.
[19] Lee, D.-S., and Sengupta, B. (1993), "Queueing analysis of a threshold based priority scheme for ATM networks", *IEEE/ACM Transactions on Networking* 1, 709–717.
[20] Meyn, S.P. (1995), "Transience of multiclass queueing networks via fluid limit models", *Ann. Appl. Probab.* 5, 946–957.
[21] Mitrani, I., and Mitra, D. (1992), "A spectral expansion method for random walks on semi-infinite strips", in: Beauwens, R., de Groen, P., eds., *Iterative Methods in Linear Algebra*, Elsevier, Amsterdam.
[22] Neuts, M.F. (1979), "Queues solvable without Rouché's theorem", *Operations Research* 27, 767–781.
[23] Press, W.H., Flannery, B.P., and Teukolsky, S.A. (1988), *Numerical Recipes in C: The Art of Scientific Computing*, Cambridge.
[24] Reiman, M.I., and Wein, L.M. (1994), "Dynamic scheduling of a two-class queue with setups", working paper.
[25] Takács, L. (1968), "Two queues attended by a single server", *Operations Research* 16, 639–650.
[26] Yadin, M. (1970), "Queueing with alternating priorities, treated as random walk on the lattice in the plane", *Journal of Applied Probability* 7, 196–218.