Polling Models with Threshold Switching

O.J. Boxma[1] G.M. Koole[2] I. Mitrani[3]

[1]CWI, P.O. Box 94079, 1090 GB Amsterdam, The Netherlands;

Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands

[2]INRIA Sophia Antipolis, B.P. 93, 06902 Sophia Antipolis, France

[3]University of Newcastle, Newcastle upon Tyne NE1 7RU, England

Abstract

We consider a model of two $M/M/1$ queues, served by a single server. The service policy for this polling model is of threshold type: the server serves queue 1 exhaustively, and does not remain at an empty queue if the other one is non-empty. It switches from queue 2 to queue 1 when the size of the latter queue reaches some level T, either preemptively or non-preemptively. All switches are instantaneous.

We determine the joint queue length distribution, both using analytic techniques and using the power series algorithm.

1   Introduction

In this paper we consider a model of two $M/M/1$ queues, which are served by a single server. The service policy for this polling model is of the threshold type: the server serves queue 1 exhaustively, and does not remain at an empty queue if the other one is non-empty. It immediately switches from queue 2 to queue 1 when the size of the latter queue reaches some level T. Both the preemptive and the non-preemptive model are considered. All switches are instantaneous. The arrival rates are λ_1, λ_2, and the service rates μ_1, μ_2. The ergodicity condition is satisfied if we assume that the traffic load $\rho := \lambda_1/\mu_1 + \lambda_2/\mu_2 < 1$.

We are interested in the two-dimensional steady-state queue length process, 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.

The motivation for this work is two-fold. The first one is application-oriented. Polling models like the present one find wide applicability in computer-communications, manufacturing, road traffic, etc. In particular, in modern telecommunication networks 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 non-real-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 non-real-time traffic. Threshold-type service disciplines seem appropriate for this purpose; thus one would like to obtain insight into their performance. Some threshold-based priority systems have recently been proposed and analysed by Lee and Sengupta [11, 12]. In [11] 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 [12] a customer of each queue is served alternately unless the queue length of the real-time-traffic queue exceeds a certain threshold level; then only customers from that queue are served. This is similar to our model, in which queues are served *exhaustively* unless a threshold level is reached.

A second motivation for the present study is the interesting feature that the server behaviour is not only determined by the situation at the queue that is presently being visited, but also by the situation at the other queue. [4, 6, 8, 12, 14, 17] are among the few polling papers that take this possibility into consideration. Hofri and Ross [6] have studied the optimal switching policy for a two-queue polling model with holding costs and switchover times and costs. They have shown that, for the case of identical service time distributions and equal holding costs at both queues, the optimal policy w.r.t. the long-run discounted costs is to serve each queue exhaustively (see also Liu et al. [10] for a model with an arbitrary number of queues). Hofri and Ross furthermore conjecture that, once the server has exhausted a queue, it is optimal to switch to the other queue only if its queue length exceeds a certain level.

Koole [8] presents results on a model with exponential service times, but with unequal parameters. In this case the low priority queue should not always be served exhaustively. For all queue length combinations in a truncated state space he determines via dynamic programming whether the server should 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. More limited results, but for general service times, were obtained by Duenyas and Van Oyen [4]. For the same model, Reiman and Wein [14] arrive at a similar policy using heavy traffic analysis. Yadin [17] 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. His analysis is based on elegant random walk considerations and the method of collective marks.

For the preemptive model, we obtain the joint queue length distribution under the threshold policy in Section 2, via an analytic approach. In Section 3 we apply the same method to the non-preemptive model. This analytic approach

will generally break down when there are more than two queues. Hence, in Section 4, we develop a numerical approach, which is also applicable to the case of multiple queues. It is based on the power series algorithm (cf. [1, 9]). Some numerical results are presented to indicate the effect of the threshold level on the mean queue lengths.

Note This paper builds upon and extends [3], in which we restricted ourselves to the preemptive case.

2 Analytic Solution for the preemptive model

Let X_1 and X_2 be the numbers of jobs in queue 1 and queue 2, respectively. Also, let S be a random variable which is equal to 1 if the server is at queue 1, and to 2 if it is at queue 2. Define the steady-state probabilities

$$
\begin{aligned}
p_{ij} &= P(X_1 = i, X_2 = j, S = 1), \ i \geq 1, j \geq 0, \\
q_{ij} &= P(X_1 = i, X_2 = j, S = 2), \ 0 \leq i \leq T-1, j \geq 1, \\
r_{00} &= P(X_1 = 0, X_2 = 0).
\end{aligned}
$$

These probabilities satisfy the following balance equations:

$$
\begin{aligned}
p_{ij}(\lambda_1 + \lambda_2 + \mu_1) &= \lambda_1 p_{i-1,j}\delta(i > 1) + \lambda_2 p_{i,j-1}\delta(j > 0) \\
&\quad + \mu_1 p_{i+1,j} + \lambda_1 r_{00}\delta(i = 1, j = 0) + \lambda_1 q_{T-1,j}\delta(i = T, j > 0) \\
&\quad + \mu_2 q_{i1}\delta(i < T, j = 0), \ i \geq 1, j \geq 0, \quad (1)
\end{aligned}
$$

$$
\begin{aligned}
q_{ij}(\lambda_1 + \lambda_2 + \mu_2) &= \lambda_1 q_{i-1,j}\delta(i > 0) + \lambda_2 q_{i,j-1}\delta(j > 1) \\
&\quad + \mu_2 q_{i,j+1} + \lambda_2 r_{00}\delta(i = 0, j = 1) \\
&\quad + \mu_1 p_{1j}\delta(i = 0), \ 0 \leq i \leq T-1, j \geq 1, \quad (2)
\end{aligned}
$$

$$
r_{00}(\lambda_1 + \lambda_2) = \mu_1 p_{10} + \mu_2 q_{01}, \quad (3)
$$

where $\delta(A) = 1$ if the condition A holds, and 0 otherwise. In addition,

$$
r_{00} + \sum_{i=1}^{\infty}\sum_{j=0}^{\infty} p_{ij} + \sum_{i=0}^{T-1}\sum_{j=1}^{\infty} q_{ij} = 1. \quad (4)
$$

Introduce the generating functions

$$
\begin{aligned}
P(x, y) &= \sum_{i=1}^{\infty}\sum_{j=0}^{\infty} p_{ij} x^{i-1} y^j, \\
Q_i(y) &= \sum_{j=1}^{\infty} q_{ij} y^{j-1}, \ 0 \leq i \leq T-1,
\end{aligned}
$$

and define

$$K(x,y) = \lambda_1(1-x) + \lambda_2(1-y) + \mu_1(1 - \frac{1}{x}),$$

$$a(y) = \lambda_1 + \lambda_2(1-y) + \mu_2(1 - \frac{1}{y}).$$

Then the equations (1), (2) and (3) imply

$$P(x,y)K(x,y) = -\frac{\mu_1}{x}P(0,y) + \lambda_1 r_{00} + \lambda_1 x^{T-1}yQ_{T-1}(y)$$

$$+\mu_2 \sum_{i=1}^{T-1} x^{i-1}Q_i(0), \tag{5}$$

$$Q_0(y)a(y) = -\frac{\mu_2}{y}Q_0(0) + \lambda_2 r_{00} + \frac{\mu_1}{y}\Big(P(0,y) - P(0,0)\Big), \tag{6}$$

$$Q_i(y)a(y) = \lambda_1 Q_{i-1}(y) - \frac{\mu_2}{y}Q_i(0), \ 1 \le i \le T-1. \tag{7}$$

For every y in the interior of the unit disk, the kernel $K(x,y)$ has a unique zero, $x = \alpha(y)$, in the same region. It can be shown (see [15]), that that zero satisfies

$$\alpha(y) = E\big[e^{-\lambda_2(1-y)B_1}\big],$$

where B_1 is the busy period of queue 1 in isolation.

Since $P(x,y)$ is analytic in the polydisk $|x| \le 1$, $|y| \le 1$, the right-hand side of (5) must vanish for all zeros $x = \alpha(y)$ of $K(x,y)$ in $|y| \le 1$. This gives

$$P(0,y) = \frac{\lambda_1}{\mu_1}r_{00}\alpha(y) + \frac{\lambda_1}{\mu_1}\alpha^T(y)yQ_{T-1}(y)$$

$$+\frac{\mu_2}{\mu_1} \sum_{i=1}^{T-1} \alpha^i(y)Q_i(0). \tag{8}$$

Remark It should be noted that $\alpha(y)$ is the generating function of the number of arrivals at queue 2 during one busy period at queue 1. The appearance of this term is natural. For example, the ith term in the sum in (8) corresponds to the following situation. The server leaves queue 2 behind empty, finding i customers at queue 1. The next time that he leaves queue 1, after i 'busy periods', the number of customers that he will find at queue 2 has generating function $\alpha^i(y)$.

Successive substitutions of (7) yield:

$$Q_{T-1}(y) = \Big(\frac{\lambda_1}{a(y)}\Big)^{T-1}Q_0(y) - \frac{\mu_2}{ya(y)} \sum_{i=1}^{T-1} \Big(\frac{\lambda_1}{a(y)}\Big)^{T-1-i}Q_i(0).$$

4

Equation (8) can now be rewritten as

$$P(0, y) = \frac{\lambda_1}{\mu_1} r_{00} \alpha(y)$$

$$+ \frac{\lambda_1}{\mu_1} \alpha^T(y) y \left(\left(\frac{\lambda_1}{a(y)} \right)^{T-1} Q_0(y) - \frac{\mu_2}{ya(y)} \sum_{i=1}^{T-1} \left(\frac{\lambda_1}{a(y)} \right)^{T-1-i} Q_i(0) \right)$$

$$+ \frac{\mu_2}{\mu_1} \sum_{i=1}^{T-1} \alpha^i(y) Q_i(0). \tag{9}$$

By using (3), equation (6) can be simplified to

$$Q_0(y) a(y) = \frac{\mu_1}{y} P(0, y) + \left(\lambda_2 - \frac{\lambda_1 + \lambda_2}{y} \right) r_{00}. \tag{10}$$

Eliminating $P(0, y)$ from (9) and (10) allows us to write

$$u(y) Q_0(y) = \mu_2 \sum_{i=1}^{T-1} \alpha(y)^i [(ya(y))^{T-1} - (\lambda_1 \alpha(y))^{T-i} a(y)^{i-1} y^{T-1}] Q_i(0)$$

$$- r_{00} [\lambda_1 (1 - \alpha(y)) + \lambda_2 (1 - y)] (ya(y))^{T-1}, \tag{11}$$

where

$$u(y) := (ya(y))^T - (\lambda_1 y \alpha(y))^T. \tag{12}$$

It only remains now to determine the T unknown constants $Q_i(0)$, $i = 1, 2, \ldots, T-1$, and r_{00}. The last of these is given by $r_{00} = 1 - \rho$ (this follows from first principles, but can also be obtained from the normalizing equation). Additional equations for the other unknowns are obtained by noting that the right-hand side of (11) must vanish whenever $u(y) = 0$ and $|y| \leq 1$. When the ergodicity condition $\rho < 1$ is fulfilled, the function $u(y)$ has exactly T zeros in the unit disc, of which one is at $y = 1$ (this is seen directly), and the other $T-1$ are strictly in the interior. Indeed, consider the two functions in the right-hand side of (12), $v(y) := (ya(y))^T$ and $w(y) := (\lambda_1 y \alpha(y))^T$. These functions are analytic in the unit disc, and satisfy $|v(y)| \geq |w(y)|$ when $|y| = 1$. Moreover, that inequality is strict everywhere on the circle except at $y = 1$. At this last point, the derivative of $u(y)$ is given by

$$u'(1) = T \lambda_1^{T-1} \mu_2 \frac{1 - \rho}{1 - \lambda_1/\mu_1}.$$

Thus, when $\rho < 1$, this derivative is positive at $y = 1$. Hence, if the unit circle is deformed slightly in the region of unity, by making it pass through a point $y = 1 + \epsilon$ for some small ϵ, then on the modified contour there would be a strict inequality $|v(y)| > |w(y)|$. Applying Rouché's theorem to those two functions,

and noting that $v(y)$ has exactly one (real) zero of multiplicity T in the unit disc, establishes the desired proposition.

Let us now study the $T-1$ zeros inside the unit disc in some more detail. Let e_k, $k = 0, 1, \ldots, T-1$ be the T roots of unity of order T: $e_k = e^{2ki\pi/T}$. For $k = 1, 2, \ldots, T-1$, the equation $e_k y a(y) = \lambda_1 y \alpha(y)$ has a single root, y_k, strictly inside the unit disk. Indeed, that equation is equivalent to

$$\lambda_2 y^2 - \left[\lambda_1\left(1 - \frac{\alpha(y)}{e_k}\right) + \lambda_2 + \mu_2\right] y + \mu_2 = 0;$$

one can in fact apply Rouché's theorem to the two functions $f(y) = [\lambda_1(1 - \frac{\alpha(y)}{e_k}) + \lambda_2 + \mu_2]y$ and $g(y) = \lambda_2 y^2 + \mu_2$. It can also be seen that

$$y_k = E\left[e^{-\lambda_1(1 - \alpha(y_k)/e_k)B_2}\right],$$

where B_2 is the busy period of queue 2 in isolation.

Substituting $y = y_1, y_2, \ldots, y_{T-1}$ into the right-hand side of (11) and equating to 0 yields $T-1$ linear relations for the unknown $Q_i(0)$. Setting $y = y_0 = 1$ leads to an identity and does not supply any new information. It should be noted that the roots y_k are either real, or come in complex-conjugate pairs, $y_k = \overline{y_{T-k}}$. Since each such pair provides two real linear equations, there are exactly $T-1$ equations for the $T-1$ unknowns.

Having found the constants $Q_i(0)$, the generating functions $P(0, y)$, $Q_i(y)$ and $P(x, y)$ are completely determined, and can be used to calculate performance measures of interest.

We have not formally verified (apart from some special cases) that the equations for $Q_i(0)$ are independent of each other when $\rho < 1$. However, an intuitive argument for independence is provided by the remark that if that set of equations has more than one solution, then the Markov process would have more than one equilibrium distribution, which is impossible.

As an example, we have computed the expected queue length EX_1 using a symbolic manipulation package. It is readily seen that

$$EX_1 = \frac{d}{dx}xP(x, 1)\Big|_{x=1} + \sum_{i=0}^{T-1} iQ_i(1).$$

The only time consuming part is the numerical computation of the $Q_i(0)$'s, which takes a few seconds on a workstation. The results are equal to those in tables 1 to 3 (as found using the power series algorithm), and their imaginary parts are in the order of 10^{-10}, giving an indication of the accuracy of the final results.

Remark When $T = 1$, this model is equivalent to the M/M/1 queue with two customer classes and preemptive-resume priority for class 1. Indeed, our results

6

for $T = 1$ agree with those in Ch. 4 of Jaiswal [7]. Similarly, when $T = \infty$, the model reduces to the classical two-queue model with exhaustive service at both queues, cf. Takács [16].

3   Analytic Solution for the non-preemptive model

In some applications, it may be undesirable or impossible to interrupt a service which is in progress. It is then of interest to consider a non-preemptive threshold policy: the server switches from queue 2 to queue 1 only at service completion instants, if the current size of queue 1 is greater than or equal to T. The analysis proceeds along similar lines to that in Section 2.

The probabilities p_{ij} and q_{ij} are defined as before, except that q_{ij} can now have non-zero values for all $i \geq 0$ and $j \geq 1$. The new balance equations corresponding to (1) and (2) have the form:

$$
\begin{aligned}
p_{ij}(\lambda_1 + \lambda_2 + \mu_1) &= \lambda_1 p_{i-1,j}\delta(i > 1) + \lambda_2 p_{i,j-1}\delta(j > 0) \\
&\quad + \lambda_1 r_{00}\delta(i = 1, j = 0) + \mu_1 p_{i+1,j} + \mu_2 q_{i1}\delta(i < T, j = 0) \\
&\quad + \mu_2 q_{i,j+1}\delta(i \geq T), \ i \geq 1, j \geq 0,
\end{aligned}
$$
(13)

$$
\begin{aligned}
q_{ij}(\lambda_1 + \lambda_2 + \mu_2) &= \lambda_1 q_{i-1,j}\delta(i > 0) + \lambda_2 q_{i,j-1}\delta(j > 1) \\
&\quad + \lambda_2 r_{00}\delta(i = 0, j = 1) + \mu_2 q_{i,j+1}\delta(i < T) \\
&\quad + \mu_1 p_{1j}\delta(i = 0), \ i \geq 0, j \geq 1.
\end{aligned}
$$
(14)

Equation (3) remains unchanged, and the normalizing equation becomes

$$
r_{00} + \sum_{i=1}^{\infty}\sum_{j=0}^{\infty} p_{ij} + \sum_{i=0}^{\infty}\sum_{j=1}^{\infty} q_{ij} = 1 \ .
$$
(15)

In addition to the generating functions $P(x,y)$ and $Q_i(y)$ $(i = 1, 2, \ldots, T-1)$, introduced in Section 2, define

$$
Q(x,y) = \sum_{i=T}^{\infty}\sum_{j=1}^{\infty} q_{ij} x^{i-T} y^{j-1}.
$$

Then equations (13) are transformed into

$$
\begin{aligned}
P(x,y)K(x,y) &= -\frac{\mu_1}{x}P(0,y) + \lambda_1 r_{00} \\
&\quad + \mu_2 \sum_{i=1}^{T-1} x^{i-1} Q_i(0) + \mu_2 x^{T-1} Q(x,y),
\end{aligned}
$$
(16)

where $K(x,y)$ is the same kernel as in (5). Equations (14) imply

$$
Q(x,y)[\lambda_1(1-x) + \lambda_2(1-y) + \mu_2] = \lambda_1 Q_{T-1}(y).
$$
(17)

Equations (6) and (7) remain unchanged.

Setting, as before, $x = \alpha(y)$ in (16) and using the fact that the right-hand side must vanish, yields a relation similar to (8):

$$\mu_1 P(0, y) = \lambda_1 \alpha(y) r_{00} + \mu_2 \sum_{i=1}^{T-1} \alpha(y)^i Q_i(0) + \mu_2 \alpha(y)^T Q(\alpha(y), y). \qquad (18)$$

Next, substitute (17) into (18), express $Q_{T-1}(y)$ in terms of $Q_0(y)$ and eliminate $P(0, y)$ from the resulting expression and from (6). This leads to

$$Q_0(y)h(y) =$$
$$y^{T-1} \sum_{i=1}^{T-1} [\mu_2 a(y)^{T-1} \alpha(y)^i - \lambda_1^{T-i} y^{-1} \alpha(y)^T a(y)^{i-1} c(\alpha(y), y)] Q_i(0)$$
$$-r_{00} y^{T-1} a(y)^{T-1} [\lambda_1(1 - \alpha(y)) + \lambda_2(1 - y)], \qquad (19)$$

where

$$c(x, y) = \frac{\mu_2}{\lambda_1(1 - x) + \lambda_2(1 - y) + \mu_2},$$

and

$$h(y) = y^T a(y)^T - \lambda_1^T y^{T-1} \alpha(y)^T c(\alpha(y), y). \qquad (20)$$

Again it remains to determine the T unknown constants $Q_i(0)$, $i = 1, 2, \ldots, T-1$, and r_{00}. The last of these is still equal to $1 - \rho$, since the probability of an empty system does not depend on the scheduling strategy. Additional equations for the other unknowns are obtained by noting that the right-hand side of (19) vanishes whenever $h(y) = 0$ and $|y| \leq 1$.

When the ergodicity condition $\rho < 1$ is satisfied, the function $h(y)$ has exactly T zeros in the unit disc, of which one is at $y = 1$ (this is seen directly), and the other $T - 1$ are strictly in the interior. Indeed, consider the two functions in the right-hand side of (20), $f(y) := y^T a(y)^T$ and $g(y) := \lambda_1^T y^{T-1} \alpha(y)^T c(\alpha(y), y)$. These functions are analytic in the unit disc, and satisfy $|f(y)| \geq |g(y)|$ when $|y| = 1$. Moreover, that inequality is strict everywhere on the circle except at $y = 1$. At this last point, the derivative of $h(y)$ is given by

$$h'(1) = \lambda_1^{T-1}(\lambda_1 + T\mu_2)\frac{1 - \rho}{1 - \lambda_1/\mu_1}.$$

Thus, when $\rho < 1$, this derivative is positive at $y = 1$. Hence, if the unit circle is deformed slightly in the region of unity, by making it pass through a point $y = 1 + \epsilon$ for some small ϵ, then on the modified contour there would be a strict inequality $|f(y)| > |g(y)|$. Applying Rouché's theorem to those two functions, and noting that $f(y)$ has exactly one (real) zero of multiplicity T in the unit disc, establishes the desired proposition.

The $T - 1$ zeros of $h(y)$ in the interior of the unit disc provide the necessary equations for determining the unknown probabilities $Q_i(0)$, $i = 1, 2, \ldots, T - 1$.

Remark When $T = 1$, this model is equivalent to the M/M/1 queue with two customer classes and nonpreemptive priority for class 1. Indeed, our results for $T = 1$ agree with those of Miller [13]. Similarly, when $T = \infty$, the model reduces to the classical two-queue model with exhaustive service at both queues, cf. Takács [16].

4 Power series algorithm

The method of Sections 2 and 3 cannot be readily extended to the case of three or more queues. The power series algorithm is a recently developed method that enables the numerical solution of multi-dimensional Markov processes (cf. the survey [1]). Below we apply this method to our two-queue models, but its extension to more than two queues is immediate.

Introduce an artificial parameter ζ in the transition rates by replacing λ_1 by $\zeta\lambda_1$ and λ_2 by $\zeta\lambda_2$. The service parameters μ_1 and μ_2 remain unchanged. Formally we can write

$$
\begin{aligned}
p_{ij} &= \sum_{k=0}^{\infty} \tilde{p}_{kij}\zeta^{i+j+k}, \\
q_{ij} &= \sum_{k=0}^{\infty} \tilde{q}_{kij}\zeta^{i+j+k}, \\
r_{00} &= \sum_{k=0}^{\infty} \tilde{r}_{k00}\zeta^{k},
\end{aligned}
$$

because p_{ij} and q_{ij} are $\mathcal{O}(\zeta^{i+j})$ (as follows from the theory developed in [9]).

The power series algorithm approximates the stationary distribution of a given Markov process by computing its partial sums, the coefficients of which can be computed by a simple recursion assuming that ζ is introduced correctly in the transition rates. It follows from [9] that the above choice is a good one for the model at hand.

Consider first the preemptive model. By equating terms with equal power of ζ in (1) and (2) we get

$$
\begin{aligned}
\mu_1\tilde{p}_{kij} + (\lambda_1 + \lambda_2)\tilde{p}_{k-1,i,j}\delta(k > 0) = \\
\lambda_1\tilde{q}_{k,T-1,j}\delta(i = T, j > 0) + \lambda_1\tilde{p}_{k,i-1,j}\delta(i > 1) \\
+ \lambda_1\tilde{r}_{k00}\delta(i = 1, j = 0) + \lambda_2\tilde{p}_{k,i,j-1}\delta(j > 0) + \mu_1\tilde{p}_{k-1,i+1,j}\delta(k > 0) \\
+ \mu_2\tilde{q}_{k-1,i,1}\delta(i < T, j = 0, k > 0), \; i > 0, \quad (21)
\end{aligned}
$$

$$
\begin{aligned}
\mu_2\tilde{q}_{kij} + (\lambda_1 + \lambda_2)\tilde{q}_{k-1,i,j}\delta(k > 0) = \\
\lambda_1\tilde{q}_{k,i-1,j}(i > 0) + \lambda_2\tilde{q}_{k,i,j-1}\delta(j > 1) \\
+ \lambda_2\tilde{r}_{k00}\delta(i = 0, j = 1) + \mu_1\tilde{p}_{k-1,1,j}\delta(i = 0, k > 0) \\
+ \mu_2\tilde{q}_{k-1,i,j+1}\delta(k > 0), \; i < T, j > 0. \quad (22)
\end{aligned}
$$

From the normalizing equation it follows that $\tilde{r}_{000} = 1$ and

$$\tilde{r}_{l00} + \sum_{i+j+k=l} (\tilde{p}_{kij} + \tilde{q}_{kij}) = 0, \ l \geq 1.$$

Combining these results with (21) and (22) enables us to derive all coefficients recursively. Taking $\zeta = 1$ returns us to the original model. We restrict the computation to all coefficients with the power of ζ up to a number K, i.e., all coefficients with $k + i + j \leq K$. The resulting partial sums are used as approximations for the steady-state probabilities. Note that if $K \geq 1$, then the approximation of r_{00} is exact, as $r_{00} = 1 - \zeta\lambda_1/\mu_1 - \zeta\lambda_2/\mu_2 = 1 - \zeta\rho$.

Below are three tables containing the average queue lengths for queue 1 and queue 2, for various parameter combinations. The tables also contain the queue lengths for the non-preemptive case, denoted by EX_1' and EX_2', which can be computed in an analogous way. The ϵ-algorithm (see [1]) was applied, and the results are based on the terms with power of ζ smaller than 60.

T	EX_1	EX_2	EX_1'	EX_2'
1	0.5000	1.5000	0.6667	1.3333
2	0.6607	1.3393	0.7522	1.2478
3	0.7711	1.2289	0.8254	1.1746
4	0.8457	1.1543	0.8794	1.1206
5	0.8960	1.1040	0.9175	1.0825
10	1.0000	1.0000	1.0000	1.0000

Table 1: Average queue sizes for different thresholds;
$\lambda_1 = \lambda_2 = 1$, $\mu_1 = \mu_2 = 3$.

T	EX_1	EX_2	EX_1'	EX_2'
1	1.0000	5.0000	1.2222	4.6667
2	1.2587	4.6119	1.3999	4.4001
3	1.4671	4.2993	1.5669	4.1497
4	1.6343	4.0485	1.7084	3.9374
5	1.7686	3.8471	1.8252	3.7622
10	2.1372	3.3020	2.1543	3.2351
15	2.2692	3.0944	2.2759	3.0873
20	2.3326	3.0008	2.3328	3.0004
30	2.3333	3.0000	2.3333	3.0000

Table 2: Average queue sizes for different thresholds;
$\lambda_1 = \lambda_2 = 1$, $\mu_1 = 2$ and $\mu_2 = 3$.

T	EX_1	EX_2	EX_1'	EX_2'
1	0.5000	4.0000	0.8750	3.7500
2	0.8177	3.7882	1.0801	3.6133
3	1.0950	3.6033	1.2971	3.4686
4	1.3364	3.4424	1.4994	3.3337
5	1.5467	3.3022	1.6815	3.2124
10	2.2574	2.8236	2.3188	2.7912
15	2.6393	2.5911	2.6671	2.5679
20	2.7836	2.4955	2.8114	2.4669
30	3.0000	2.3333	3.0000	2.3333

Table 3: Average queue sizes for different thresholds;
$\lambda_1 = \lambda_2 = 1$, $\mu_1 = 3$ and $\mu_2 = 2$.

Remark The conservation law for this model states that (cf. Gelenbe & Mitrani [5], Ch. 6)

$$\frac{EX_1}{\mu_1} + \frac{EX_2}{\mu_2} = \frac{EX_1'}{\mu_1} + \frac{EX_2'}{\mu_2} = \frac{\lambda_1/\mu_1^2 + \lambda_2/\mu_2^2}{1 - \rho}.$$

The results in the tables satisfy this conservation law almost exactly in all cases.

Remark Based on the numerical results and our intuition, we conjecture that EX_1 and EX_1' are mononically increasing in T. We have not been able to prove this.

Comments on the analytic and power series solutions For a given threshold, T, the analytic approach involves the determination of $T - 1$ zeros in the interior of the unit disk, plus the solution of $T - 1$ simultaneous linear equations. The complexity of that solution is therefore on the order of $O(T^3)$. The accuracy of the results is limited only by the numerical precision of the software that is used. On the other hand, both the complexity and the accuracy of the power series algorithm depend on the number of terms in the expansion that are computed. There is no guarantee that any fixed number of terms will achieve a given accuracy. However, the experimental results are quite impressive: a relative error of less than 1% is achieved with 4 terms for T=1, 30 terms for $T = 5$ and 12 terms for $T = 30$.

Remark In a forthcoming paper [2], the case of general service time distributions at both queues, switchover times and nonpreemptive switching rule will be analyzed. In that study the joint queue length distribution at imbedded epochs of service completions is determined.

Acknowledgement

References

[1] J.P.C. Blanc (1993), Performance analysis and optimization with the power-series algorithm, in: Performance Evaluation of Computer and Communication Systems (eds. L. Donatiello and R.D. Nelson), Springer, New York, 53-80

[2] O.J. Boxma and D.G. Down, A two-queue polling model with threshold switching, in preparation

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

[4] I. Duenyas and M.P. Van Oyen (1993), Stochastic scheduling of parallel queues with set-up costs, Technical Report 93-09, Northwestern University

[5] E. Gelenbe, I. Mitrani (1980), Analysis and Synthesis of Computer Systems, Academic Press, London

[6] M. Hofri, K.W. Ross (1987), On the optimal control of two queues with server setup times and its analysis, SIAM Journal on Computing, **16**, 399-420

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

[8] G.M. Koole (1994), Assigning a single server to inhomogeneous queues with switching costs, CWI Report BS-R9405

[9] G.M. Koole (1994), On the power series algorithm, in: Performance Evaluation of Parallel and Distributed Systems, Part 1 (eds. O.J. Boxma and G. M. Koole), CWI Tract 105, Amsterdam, 139-155

[10] Z. Liu, P. Nain, D. Towsley (1992), On optimal polling policies, Queueing Systems, **11**, 59-83

[11] D-S. Lee (1993), A two-queue model with exhaustive and limited service disciplines, Report C & C Research Laboratories, NEC USA Inc.

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

[13] R.G. Miller (1960), Priority queues, Ann. Math. Statistics, **31**, 86-103

[14] M.I. Reiman, L.M. Wein (1994), Dynamic scheduling of a two-class queue with setups, Working paper

[15] L. Takács (1962), Introduction to the Theory of Queues, Oxford University Press, New York

[16] L. Takács (1968), Two queues attended by a single server, Operations Research, **16**, 639-650

[17] M. Yadin (1970), Queueing with alternating priorities, treated as random walk on the lattice in the plane, Journal of Applied Probability, **7**, 196-218