ZOR – Mathematical Methods of Operations Research (1994) 40:289–303

Waiting Times in a Two-Queue Model with Exhaustive and Bernoulli Service

Jan A. WESTSTRATE

Shell Nederland Informatieverwerking, Shell Nederland B. V., Postbus 5835, 2280 HV Rijswijk, The Netherlands

Rob D. VAN DER MEI

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

Abstract: This paper deals with waiting times in a two-queue polling system in which one queue is served according to the Bernoulli service discipline and the other one attains exhaustive service. Exact results are derived for the LST's of the waiting time distributions via an iteration scheme. Based on those results the mean waiting times are expressed in the system parameters.

Keywords: Polling systems, Bernoulli service, waiting times.

1 Introduction

A polling system consists of a number of queues, attended to by a single server. A larger number of queueing theoretic studies about polling systems has been published with the analysis focussing on characterizing the system performance (cf. [18] for an overview). The vast majority of those studies considers polling systems with service policies commonly used in industry: the exhaustive, the gated and the limited service strategies. The main disadvantage of those traditional systems is the inability to exercise control and to affect their design by optimizing a performance measure such as the mean waiting time of an arbitrary customer in the system or the mean amount of unfinished work in the system.

As computer and telecommunication systems become more complicated and the processing power of micro processors becomes less inexpensive, the advantage of more sophisticated polling systems becomes apparent. Recently, new service policies have been introduced, such as the fractional service policies (cf. [13, 14]), the Bernoulli-type policies (cf. [16]) and the Bernoulli service policy (cf. [12]). In this paper we will focus on the Bernoulli service strategy.

The Bernoulli service strategy with parameter p $(0 \leq p \leq 1)$ is described as follows. When the server arrives at a queue at least one customer is served if the queue is not empty; otherwise, the server immediately starts to move to the next queue. After each service which does not leave the queue empty, with probability

0340–9422/94/40:3/289–303 $2.50

$1 - p$ another customer is served at that queue; otherwise, the server proceeds to the next queue. Note that $p = 0$ and $p = 1$ correspond to the exhaustive and the 1-limited service strategy, respectively; thus, the Bernoulli service policy generalizes both the classical 1-limited policy and the exhaustive policy. An advantage of the Bernoulli service policy is that the parameter p allows both flexible modeling and system optimization.

Fuhrmann [10] classifies service policies into two classes, depending on whether or not the policies satisfy a certain property (cf. Section 5 below), and he shows that models in which all service policies satisfy this property are relatively easy to analyze. For such models Resing [16] gives an exact expression for the joint queue length distribution at polling instants. One may verify that the exhaustive, the gated, the Bernouli-type and the fractional service policies in [13, 14] satisfy this property, whereas the limited policies and the Bernoulli policy generally do *not*. For polling systems in which not all service strategies satisfy Fuhrmann's property, exact results are very scarce and are restricted to two-queue or fully symmetric models. Boxma and Groenendijk [5] and Cohen [7] used the theory of Riemann-Hilbert boundary value problems to determine the waiting times at both queues for two-queue models with either 1-limited or semi-exhaustive service at both queues. The reader is referred to [3] for further discussions of the application of the technique of boundary value problems to two-queue models and to [1, 2, 12, 15, 17, 19, 21] for references on queueing systems with Bernoulli service policies.

Our motivation is two-fold. Firstly, we have a mathematical interest in the analysis of a generalization of the most basic service disciplines, the exhaustive and the 1-limited service disciplines. Secondly, we would like to use the insight and exact results to be developed in the present study, for deriving and testing waiting time approximations in polling systems with Bernoulli service.

This paper concerns the waiting times in a polling system with two queues in which one queue has a Bernoulli service policy with parameter p $(0 \leq p \leq 1)$ and the other one a Bernoulli service policy with a parameter equal to zero, the exhaustive service policy. We shall indicate this system as the two-queue Exhaustive/Bernoulli (p) system. For this system exact expressions for the Laplace-Stieltjes Transforms (LST) of the waiting time distributions are derived via an iteration procedure; so, we need not solve a Riemann-Hilbert boundary value problem.

The paper is organized as follows. In Section 2 the model description is given. In Section 3 we determine the generating function of the joint equilibrium queue length distributions at polling instants. To this end, we first derive recurrence relations between the generating functions of the joint equilibrium queue length distributions at polling instants at both of the queues. Then these recurrence relations are used to give explicit expressions for the generating functions of the steady-state queue lengths at polling instants via an iteration scheme. In Section 4 these results are used to obtain the mean waiting times at both queues in the system parameters. In Section 5 we discuss some topics for further research.

2 Model Description

A single server S serves two queues Q_1 and Q_2 in cyclic order. Both queues have an infinite buffer capacity. Type-i customers arrive at Q_i according to a Poisson process with rate λ_i, $i = 1, 2$. The service times at Q_i are independent, identically distributed (i.i.d.) stochastic variables with LST $B_i\{.\}$, first moment β_i and second moment $\beta_i^{(2)}$, $i = 1, 2$.

The offered load ρ_i to Q_i is defined by

$$\rho_i := \lambda_i \beta_i \ , \qquad i = 1, 2 \ , \tag{1}$$

and the total offered load, ρ, to the system is given by:

$$\rho := \rho_1 + \rho_2 \ . \tag{2}$$

The service strategy at Q_1 is the exhaustive service policy; Q_2 is served according to the Bernoulli service strategy with parameter p ($0 \le p \le 1$). The times needed by the server to move from Q_1 to Q_2 are i.i.d. random variables with LST $S_{12}\{.\}$, mean s_{12} and second moment $s_{12}^{(2)}$; the switch-over times to move from Q_2 to Q_1 have parameters $S_{21}\{.\}$, s_{21} and $s_{21}^{(2)}$, respectively. The first and second moment of the total switchover time during a cycle are denoted by

$$s := s_{12} + s_{21} \tag{3}$$

and

$$s^{(2)} := s_{12}^{(2)} + 2s_{12}s_{21} + s_{21}^{(2)} \ . \tag{4}$$

All stochastic processes are assumed to be mutually independent.

Necessary and sufficient conditions for the stability of polling systems have been derived in [9]. For the present model this condition reads:

$$\rho + \lambda_2 p s < 1 \ . \tag{5}$$

Throughout it is assumed that the stability condition (5) is satisfied.

For convenience, we introduce the following notation. If $X\{(1 - z_1)\lambda_1 + (1 - z_2)\lambda_2\}$ is the LST with argument $(1 - z_1)\lambda_1 + (1 - z_2)\lambda_2$ of a certain stochastic variable, then we define for $|z_1| \le 1$, $|z_2| \le 1$:

$$X\{z_1, z_2\} := X\{(1 - z_1)\lambda_1 + (1 - z_2)\lambda_2\}, \quad \text{and}$$

$$XY\{z_1, z_2\} := X\{z_1, z_2\} Y\{z_1, z_2\}$$

3 Derivation of the Generating Functions of the Queue Lengths at Polling Instants

In this section we determine the generating functions of the joint equilibrium queue length distributions at so-called polling instants of Q_1 and Q_2, i.e., the instants at which S arrives at the queues. To this end, we first derive recurrence relations between the generating functions; secondly, these recurrence relations are used to obtain explicit expressions for the generating functions of the steady-state queue lengths at polling instants.

3.1 *Determination of Recurrence Relations Between the Generating Functions*

Let $x_n^{(i)}$ denote the number of type-i customers in the system at the n-th polling instant, $i = 1, 2$, $n = 1, 2, \ldots$. Then the joint queue length process at Q_1 and Q_2 at successive polling epochs, $M := \{(x_n^{(1)}, x_n^{(2)}), n = 1, 2, \ldots\}$, forms an irreducible and positive recurrent bivariate Markov process.

By convention, visits to Q_1 correspond to the cases in which n is odd; Q_2 is visited when n is even, $n = 1, 2, \ldots$. Hence, M is clearly periodic. However, the embedded Markov processes $M_1 = \{(x_{2n+1}^{(1)}, x_{2n+1}^{(2)}), n = 0, 1, \ldots\}$ and $M_2 = \{(x_{2n+2}^{(1)}, x_{2n+2}^{(2)}), n = 0, 1, \ldots\}$ are irreducible, positive recurrent and aperiodic Markov processes with stationary transition probabilities, so that their limiting distributions exist. They shall be determined in the following. Define for $|z_1| \leq 1$, $|z_2| \leq 1$:

$$F^{(n)}(z_1, z_2) := E\{z_1^{x_n^{(1)}} z_2^{x_n^{(2)}}\}, \qquad n = 1, 2, \ldots . \tag{6}$$

A study of the one step transition probabilities of the Markov chain M yields recurrence relations for the generating functions of the queue lengths at polling instants, $F^{(n)}(z_1, z_2)$, $|z_1| \leq 1$, $|z_2| \leq 1$, $n = 1, 2, \ldots$. For the derivation of those relations we need some additional definitions and a theorem concerning the joint distribution of the length of a busy period and the number of customers at the end of that busy period in an $M/G/1$ queue with vacations and a Bernoulli service discipline. Define for such a queue:

$S_j(t, k) :=$ the joint cumulative probability distribution of the length of a busy period and the queue length at the end of that busy period, conditioned on the fact that the busy period starts with j customers, $t \geq 0, j, k = 0, 1, \ldots$.

Note that for all $t \geq 0$ $S_0(t, 0) := 1$ and $S_0(t, k) := 0$ for $k > 0$. We also define the joint LST and generating function:

$$\sigma_j(r, s) := \sum_{k=0}^{\infty} s^k \int_{t=0}^{\infty} e^{-rt} d_t S_j(t, k), \; |s| \leq 1, \, Re\, r \geq 0, j = 0, 1, \ldots \; . \tag{7}$$

Using Theorem I in [15], we can write for $|z_1| \leq 1, |z_2| \leq 1, p \in [0, 1]$:

$$\sigma_j((1 - z_1)\lambda_1, z_2) = \Omega_p(z_1, z_2)z_2^j + [1 - \Omega_p(z_1, z_2)]\mu_2(z_1, p)^j, \quad j = 0, 1, \ldots , \tag{8}$$

with

$$\Omega_p(z_1, z_2) := \frac{pB_2\{z_1, z_2\}}{z_2 - (1 - p)B_2\{z_1, z_2\}} , \tag{9}$$

and for $|z_1| \leq 1$, $\mu_2(z_1, p)$ the unique solution of

$$z_2 = (1 - p)B_2\{z_1, z_2\} , \qquad |z_2| \leq 1, p \in [0, 1] \; . \tag{10}$$

Moreover, for $|z_2| \leq 1$ we introduce $\mu_1(z_2)$ as the unique solution of:

$$z_1 = B_1\{z_1, z_2\} , \qquad |z_1| \leq 1 \; . \tag{11}$$

Remarks: $\sigma_j((1 - z_1)\lambda_1, z_2)$ is the joint generating funtion of the number of type-1 arrivals during a busy period of the Bernoulli queue and the number of type-2 customers in the system at the end of that busy period, conditioned on the fact that the busy period starts with j (type-2) customers, $j = 0, 1, \ldots$.

$\mu_2(z_1, p)$ is the joint generating function of the number of type-1 arrivals and the number of customers served during a busy period of an ordinary $M/G/1$ queue with the same traffic characteristics as Q_2; this result can be obtained by means of the so-called method of collective marks (cf. [6] (p. 340)).

$\mu_1(z_2)$ is the generating function of the number of type-2 arrivals during an ordinary $M/G/1$ busy period in a queue with the same traffic characteristics as Q_1.

The existence of a unique root in (10).is demonstrated in [6] (Appendix 6). It is also shown there that if $\rho_2 < 1$ then $|\mu_2(z_1, p)| < 1$ for $p \in [0, 1]$, $|z_1| \leq 1$, except if $p = 0$ and simultaneously $z_1 = 1$; in the latter case $\mu_2(z_1, p) = 1$.

We are now nearly ready to present the derivation of the recurrence relations between the generating functions (6). We have obtained those results by a tedious, but straightforward, calculation using indicator functions, but we prefer to present them in another more intuitive way. Before we do so, we define

$y_n^{(i)} :=$ the number of type-i arrivals during the n-th visit of the server to a queue, $i = 1, 2, \quad n = 1, 2, \dots$.

Let us first consider a case in which n is odd; then the $(n+1)$-th polling epoch marks the beginning of a visit to Q_2. Then, because of the exhaustive service discipline at Q_1, the only type-1 customers present at Q_1, $x_{n+1}^{(1)}$, are those who arrived during the switch-over period between Q_1 and Q_2. Moreover, the set of the type-2 customers at Q_2, $x_{n+1}^{(2)}$, is composed of:

(i) the type-2 customers present at the n-th polling epoch, $x_n^{(2)}$; (ii) the type-2 customers who arrived during the subsequent visit of S to Q_1, the n-th visit of the server, and (iii) the type-2 customers who arrived during the switch-over period between Q_1 and Q_2.

Using these observations we can write for $|z_1| \leq 1$, $|z_2| \leq 1$:

$$E\{z_1^{x_{n+1}^{(1)}} z_2^{x_{n+1}^{(2)}}\} = S_{12}\{z_1, z_2\} E\{z_2^{y_n^{(2)}} z_2^{x_n^{(2)}}\} , \qquad n = 1, 3, 5, \dots . \tag{12}$$

If the server arrives at Q_1 and finds i type-1 customers present then we can view the visit period to Q_1 as a sequence of i.i.d. $M/G/1$ busy periods, (cf. [6] (p. 250), [16]). If we denote by P_k the k-th busy period in the sequence of i busy periods we can write for $|z_2| \leq 1$:

$$\begin{aligned} E\{z_2^{y_n^{(2)}} z_2^{x_n^{(2)}}\} &= \sum_{i=0}^{\infty} E\{z_2^{x_n^{(2)}}(x_n^{(1)} = i)\} [E\{e^{-(1-z_2)\lambda_2 P_1}\}]^i \\ &= \sum_{i=0}^{\infty} E\{z_2^{x_n^{(2)}}(x_n^{(1)} = i)\} \mu_1(z_2)^i \\ &= E\{\{\mu_1(z_2)\}^{x_n^{(1)}} z_2^{x_n^{(2)}}\} , \qquad n = 1, 3, 5, \dots . \end{aligned} \tag{13}$$

Combining (12) and (13) and using definition (6) gives for $|z_1| \leq 1$, $|z_2| \leq 1$:

$$F^{(n+1)}(z_1, z_2) = S_{12}\{z_1, z_2\} F^{(n)}(\mu_1(z_2), z_2) , \qquad n = 1, 3, 5, \dots . \tag{14}$$

Similarly, consider the case that n is even; then the $(n+1)$-th polling epoch marks the beginning of a visit to Q_1. The set of type-1 customers at Q_1, $x_{n+1}^{(1)}$, is composed of: (i) the type-1 customers present at the n-th polling instant, $x_n^{(1)}$, (ii) the type-1 arrivals during the visit of the server to Q_2, the n-th visit, and (iii) the type-1 arrivals during the switchover period between Q_2 and Q_1. Moreover, the set of type-2 customers at Q_2 is composed of: (i) the type-2 customers present at the end of the previous visit, which we shall denote by $u_n^{(2)}$, and (ii) the type-2 arrivals during the switchover period between Q_2 and Q_1.

Using these observations we can write for $|z_1| \leq 1, |z_2| \leq 1$:

$$E\{z_1^{x_{n+1}^{(1)}} z_2^{x_{n+1}^{(2)}}\} = S_{21}\{z_1, z_2\} E\{z_1^{y_n^{(1)}} z_1^{x_n^{(1)}} z_2^{u_n^{(2)}}\} , \qquad n = 2, 4, 6, \ldots , \tag{15}$$

where the second factor in the right hand side of (15) can be rewritten for $|z_1| \leq 1, |z_2| \leq 1$ as follows:

$$\sum_{i=0}^{\infty} \sum_{j=0}^{\infty} z_1^i \sigma_j((1 - z_1)\lambda_1, z_2) Pr\{x_n^{(1)} = i; x_n^{(2)} = j\}$$

$$= \Omega_p(z_1, z_2) \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} z_1^i [z_2^j - \mu_2(z_1, p)^j] Pr\{x_n^{(1)} = i; x_n^{(2)} = j\}$$

$$+ \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} z_1^i \mu_2(z_1, p)^j Pr\{x_n^{(1)} = i; x_n^{(2)} = j\}$$

$$= \Omega_p(z_1, z_2)[F^{(n)}(z_1, z_2) - F^{(n)}(z_1, \mu_2(z_1, p))]$$

$$+ F^{(n)}(z_1, \mu_2(z_1, p)) , \qquad n = 2, 4, 6, \ldots . \tag{16}$$

Combining (15) and (16) gives for $|z_1| \leq 1, |z_2| \leq 1$:

$$F^{(n+1)}(z_1, z_2) = S_{21}\{z_1, z_2\} F^{(n)}(z_1, \mu_2(z_1, p)) + S_{21}\{z_1, z_2\} \Omega_p(z_1, z_2)$$
$$\times [F^{(n)}(z_1, z_2) - F^{(n)}(z_1, \mu_2(z_1, p))] , \qquad n = 2, 4, 6, \ldots . \tag{17}$$

Define the limiting distributions for the embedded Markov processes M_1 and M_2 as follows: for $|z_1| \leq 1, |z_2| \leq 1$:

$$F_1(z_1, z_2) := \lim_{m \to \infty} F^{(2m+1)}(z_1, z_2), \; F_2(z_1, z_2) := \lim_{m \to \infty} F^{(2m+2)}(z_1, z_2) . \tag{18}$$

Using the recurrence relations (14) and (17) together with definition (18) we can relate $F_1(z_1, z_2)$, $F_2(z_1, z_2)$ as follows for $|z_1| \leq 1, |z_2| \leq 1, p \in [0, 1]$:

$$F_1(z_1, z_2) = S_{21}\{z_1, z_2\}\Omega_p(z_1, z_2)F_2(z_1, z_2) + S_{21}\{z_1, z_2\}[1 - \Omega_p(z_1, z_2)]F_2(z_1, \mu_2(z_1, p)) \ ; \tag{19}$$

$$F_2(z_1, z_2) = S_{12}\{z_1, z_2\}F_1(\mu_1(z_2), z_2) \ . \tag{20}$$

3.2 *Determination of Explicit Expressions for $F_1(z_1, z_2)$ and $F_2(z_1, z_2)$*

In this section explicit expressions for $F_1(z_1, z_2)$ and $F_2(z_1, z_2)$ will be derived from (19) and (20) via an iterative procedure.

Taking $z_1 = \mu_1(z_2)$ in relation (17) and combining the result with (18) gives for $|z_1| \leq 1, |z_2| \leq 1$ and $p \in [0, 1]$:

$$\begin{aligned} F_2(z_1, z_2) = {} & S_{12}\{z_1, z_2\}S_{21}\{\mu_1(z_2), z_2\} \times \Omega_p(\mu_1(z_2), z_2)F_2(\mu_1(z_2), z_2 \\ & + S_{12}\{z_1, z_2\}S_{21}\{\mu_1(z_2), z_2\} \\ & \times [1 - \Omega_p(\mu_1(z_2), z_2)]F_2(\mu_1(z_2), \mu_2(\mu_1(z_2), p)) \ . \end{aligned} \tag{21}$$

We will now solve equaton (21) by means of an iterative procedure. To this end, we define for $|z| \leq 1$ and $p \in [0, 1]$:

$$\begin{aligned} &\delta_p^{(0)}(z) := z; \ \delta_p(z) = \delta_p^{(1)}(z) := \mu_2(\mu_1(z), p) \ ; \\ &\delta_p^{(n)}(z) := \delta_p^{(1)}(\delta_p^{(n-1)}(z)) \ , \qquad n = 1, 2, \ldots \ . \end{aligned} \tag{22}$$

Taking $z_1 = \mu_1(z_2)$ in (21) and replacing z_2 by $\delta_p(z)$ gives for $|z| \leq 1$ and $p \in [0, 1]$:

$$\begin{aligned} F_2(\mu_1(\delta_p(z)), \delta_p(z)) = {} & F_2(\mu_1(\delta_p(z)), \delta_p^{(2)}(z)) \\ & \times \left\{ \frac{S_{12}S_{21}\{\mu_1(\delta_p(z)), \delta_p(z)\}[1 - \Omega_p(\mu_1(\delta_p(z)), \delta_p(z))]}{1 - S_{12}S_{21}\{\mu_1(\delta_p(z)), \delta_p(z)\}\Omega_p(\mu_1(\delta_p(z)), \delta_p(z))} \right\} \ . \end{aligned} \tag{23}$$

Moreover, taking $z_2 = \mu_2(z_1, p)$ in (21) and replacing z_1 by $\mu_1(z)$ in the resulting equation gives for $|z| \leq 1$ and $p \in [0, 1]$:

$$F_2(\mu_1(z), \delta_p(z)) = S_{12}\{\mu_1(z), \delta_p(z)\}S_{21}\{\mu_1(\delta_p(z)), \delta_p(z)\}$$
$$\times \{\Omega_p(\mu_1(\delta_p(z)), \delta_p(z))F_2(\mu_1(\delta_p(z)), \delta_p(z))$$
$$+ [1 - \Omega_p(\mu_1(\delta_p(z)), \delta_p(z))]F_2(\mu_1(\delta_p(z)), \delta_p^{(2)}(z))\} \ . \tag{24}$$

Combining (23) and (24) leads for $|z| \leq 1$ and $p \in [0, 1]$ to:

$$F_2(\mu_1(z), \delta_p(z)) = D_p(\delta_p(z))F_2(\mu_1(\delta_p(z)), \delta_p^{(2)}(z)) \ , \tag{25}$$

with

$$D_p(\delta_p(z)) = \frac{S_{12}\{\mu_1(z), \delta_p(z)\}S_{21}\{\mu_1(\delta_p(z)), \delta_p(z)\}[1 - \Omega_p(\mu_1(\delta_p(z)), \delta_p(z))]}{1 - S_{12}S_{21}\{\mu_1(\delta_p(z)), \delta_p(z)\}\Omega_p(\mu_1(\delta_p(z)), \delta_p(z))} \ . \tag{26}$$

Then, if we repeatedly replace z by $\delta_p(z)$ in (25), n times in succession, we have for $|z| \leq 1$ and $p \in [0, 1]$:

$$F_2(\mu_1(z), \delta_p(z)) = \left[\prod_{k=1}^{n} D_p(\delta_p^{(k)}(z))\right]F_2(\mu_1(\delta_p^{(n)}(z)), \delta_p^{(n+1)}(z)) \ . \tag{27}$$

It is shown in [21] that for $|z| \leq 1$ and $p \in [0, 1]$ there exist $a \in (0, 1]$ and $b \in (0, \infty)$ such that:

$$\lim_{n\to\infty} \delta_p^{(n)}(z) = a; \prod_{k=1}^{\infty} D_p(\delta_p^{(k)}(z)) = b \ , \tag{28}$$

and that, using the continuity of $F_2(u, v)$ in u and v,

$$\lim_{n\to\infty} F_2(\mu_1(\delta_p^{(n)}(z)), \delta_p^{(n+1)}(z)) = F_2(\mu_1(a), a) = \frac{1 - \dfrac{\lambda_2 ps}{1-\rho}}{\prod_{k=1}^{\infty} D_p(\delta_p^{(k)}(1))} \tag{29}$$

Finally, after some algebraïc manipulations, using (27), (20), (cf. [21] for a complete derivation), the generating functions of the queue lengths at polling instants can be obtained as follows for $|z_1| \leq 1$, $|z_2| \leq 1$, $p \in [0, 1]$:

$$F_1(z_1, z_2) = \left[1 - \frac{\lambda_2 ps}{1-\rho}\right] S_{21}\{z_1, z_2\}$$

$$\times \left\{ \Omega_p(z_1, z_2) S_{12}\{z_1, z_2\} G_p(z_2) \prod_{k=1}^{\infty} \frac{D_p(\delta_p^{(k)}(z_2))}{D_p(\delta_p^{(k)}(1))} \right.$$

$$+ [1 - \Omega_p(z_1, z_2)] S_{12}\{z_1, \mu_2(z_1, p)\} G_p(\mu_2(z_1, p))$$

$$\left. \times \prod_{k=1}^{\infty} \frac{D_p(\delta_p^{(k)}(\mu_2(z_1, p)))}{D_p(\delta_p^{(k)}(1))} \right\}, \tag{30}$$

and

$$F_2(z_1, z_2) = \left[1 - \frac{\lambda_2 ps}{1-\rho}\right] S_{12}\{z_1, z_2\} G_p(z_2) \prod_{k=1}^{\infty} \frac{D_p(\delta_p^{(k)}(z_2))}{D_p(\delta_p^{(k)}(1))}, \tag{31}$$

where for $|z| \leq 1$ and $p \in [0, 1]$:

$$G_p(z) := \frac{S_{21}\{\mu_1(z), z\}[1 - \Omega_p(\mu_1(z), z)]}{1 - S_{12}S_{21}\{\mu_1(z), z\}\Omega_p(\mu_1(z), z)}. \tag{32}$$

Remark: One may verify that in the special case $p = 0$ or $p = 1$, expressions (30) and (31) are identical to those derived in [8] and [11] (Section 6.3) for the Exhaustive/Exhaustive and the Exhaustive/1-limited case, respectively. These verifications are discussed in more detail in [21].

4 The Waiting Times

This section is concerned with the waiting times at the queues. Firstly, the waiting time distribution at the exhaustive queue, Q_1, is obtained; subsequently, we will consider the waiting time distribution at the Bernoulli queue, Q_2. In both cases we first give the LST of the waiting time distribution at that particular queue expressed in the generating functions derived in the previous section, and subsequently we calculate the mean waiting time.

To this end, define for $i = 1, 2$,

$W_i :=$ the waiting time of a type-i customer.

Let us first consider the waiting time at Q_1. The generating function of the queue lengths at polling instants of Q_1, the queue with the exhaustive service strategy, and the waiting time, W_1, of a type-1 customer are related as follows (cf. [20]):

$$E\{e^{-(1-z)\lambda_1 W_1}\} = \frac{1-\lambda_1\beta_1}{\left[\dfrac{\partial}{\partial z}F_1(z,1)\right]_{z=1}} \frac{1-F_1(z,1)}{B_1\{z,1\}-z}, \qquad |z| \le 1 . \tag{33}$$

Taking the derivative of (33) and evaluating in $z = 1$ yields:

$$\lambda_1 EW_1 = \frac{\left[\dfrac{\partial}{\partial z}\dfrac{1-F_1(z,1)}{1-z}\right]_{z=1}}{\left[\dfrac{\partial}{\partial y}F_1(y,1)\right]_{y=1}} + \frac{\lambda_1^2\beta_1^{(2)}}{2(1-\rho_1)} . \tag{34}$$

To get an expression for the mean waiting time at Q_1, we first expand, using equation (30), the function $(1 - F_1(z, 1))/(1 - z)$ in a power series in the neighbourhood of $z = 1$. Noting that for $p \in [0, 1]$ we have $F_2(1, \mu_2(1, p)) = 1 - \lambda_2 ps/(1 - \rho)$ (cf. (31)), we find after a lengthy but straightforward calculation the following expression for the mean waiting time at Q_1:

$$\begin{aligned} EW_1 = {} & \frac{\lambda_1\beta_1^{(2)} + \lambda_2\beta_2^{(2)}}{2(1-\rho_1)} + \frac{1-\rho}{1-\rho_1}\left\{\frac{s^{(2)}}{2s} + \frac{1-p}{p}\frac{\lambda_2\beta_2^2}{1-\rho} + \frac{\beta_2}{p}\right. \\ & \left. - \left[1 - \frac{\lambda_2 ps}{1-\rho}\right]\frac{s_{21}}{s}\frac{\beta_2}{p} - \left[1 - \frac{\lambda_2 ps}{1-\rho}\right]\frac{\beta_2}{\lambda_1 ps}\left[\frac{\partial}{\partial z}H_1(\mu_2(z,p))\right]_{z=1}\right\}, \end{aligned} \tag{35}$$

with

$$H_1(\mu_2(z,p)) := \prod_{k=0}^{\infty} \frac{D_p(\delta_p^{(k)}(\mu_2(z,p)))}{D_p(\delta_p^{(k)}(1))} . \tag{36}$$

In order to derive an expression for the waiting time at the Bernoulli queue, Q_2, we use the relation between the generating function of the queue lengths at polling instants of Q_2 and the LST of the distribution of the waiting time at Q_2 (cf. [19]). For $|z| \le 1$, $p \in [0, 1]$:

$$E\{e^{-(1-z)\lambda_2 W_2}\} = \frac{p}{1 - F_2(1, \mu_2(1, p))} \frac{F_2(1, z) - F_2(1, \mu_2(1, p))}{z - (1 - p)B_2\{1, z\}} . \tag{37}$$

Using $F_2(1, \mu_2(1, p)) = 1 - \lambda_2 ps/(1 - \rho)$ (cf. (31)), we can rewrite (37) as follows for $|z| \leq 1$, $p \in [0, 1]$:

$$E\{e^{-(1-z)\lambda_2 W_2}\} = \frac{1 - \rho}{\lambda_2 s} \frac{F_2(1, z) - 1}{z - (1 - p)B_2\{1, z\}} + \frac{p}{z - (1 - p)B_2\{1, z\}} . \tag{38}$$

Taking the derivative of (38) and evaluating it in $z = 1$ gives for $p \in [0, 1]$:

$$\lambda_2 EW_2 = \frac{1 - \rho}{\lambda_2 s} \frac{\left[\frac{\partial}{\partial z} F_2(1, z)\right]_{z=1}}{p} - \frac{1 - (1 - p)\rho_2}{p} . \tag{39}$$

To express EW_2 in the system parameters, we expand $F_2(1, z)$ (cf. (31)) in a power series in the neighbourhood of $z = 1$. After some further calculations we get for $p \in [0, 1]$:

$$\begin{aligned} EW_2 = {} & \frac{\lambda_1 \beta_1^{(2)} + \lambda_2 \beta_2^{(2)}}{2(1 - \rho_1)(1 - \rho)} \frac{1}{1 - \frac{\lambda_2 ps}{1 - \rho}} + \frac{1}{1 - \rho_1} \left[\frac{s^{(2)}}{2s} + \beta_2\right] \frac{1}{1 - \frac{\lambda_2 ps}{1 - \rho}} \\ & + \left\{s_{12} + \frac{s_{21}}{1 - \rho_1}\right\} \frac{1}{\frac{\lambda_2 ps}{1 - \rho}} + \frac{1}{\lambda_2} \left[\frac{\partial}{\partial z} H_2(z)\right]_{z=1} \frac{1}{\frac{\lambda_2 ps}{1 - \rho}} - \frac{1 - (1 - p)\rho_2}{\lambda_2 p} , \end{aligned} \tag{40}$$

with

$$H_2(z) := \prod_{k=1}^{\infty} \frac{D_p(\delta_p^{(k)}(z))}{D_p(\delta_p^{(k)}(1))} . \tag{41}$$

Remarks: By applying the chain rule to $D_p(\delta_p^{(k-1)}(\mu_2(\mu_1(z), p)))$ and noting that $\mu_1(1) = 1$ we get the following relation between the infinite products in (36) and (41):

$$\left[\frac{\partial}{\partial z} H_2(z)\right]_{z=1} = \frac{\lambda_2 \beta_1}{1 - \rho_1} \left[\frac{\partial}{\partial z} H_1(\mu_2(z, p))\right]_{z=1} . \tag{42}$$

A pseudo-conservation law for the present model has been derived in [4, 19]:

$$\rho_1 EW_1 + \rho_2\left[1 - \frac{\lambda_2 ps}{1-\rho}\right]EW_2 = \rho\frac{\lambda_1\beta_1^{(2)} + \lambda_2\beta_2^{(2)}}{2(1-\rho)} + \rho\frac{s^{(2)}}{2s}$$

$$+ \frac{s}{1-\rho}\rho_1\rho_2 + \frac{s}{1-\rho}\rho_2^2 p\ . \tag{43}$$

Using relation (42) we find after a tedious but straightforward calculation that the expressions for the mean waiting times in (35) and (40) satisfy this pseudo-conservation law.

5 Discussion

Recently, Resing [16] and Fuhrmann [10] have considered cyclic polling systems in which the service disciplines satisfy the following 'additivity property': if S arrives at a queue to find k customers there, then during the course of the server's visit, each of these k customers will be effectively replaced in an i.i.d. manner by a random population. For polling systems with an arbitrary number of queues and in which each of the service disciplines satisfies this property, it is shown in [16] that the joint queue length process constitutes a multi-type branching process (MTBP) with immigration. The p.g.f. of the joint queue length distribution at polling instants at a particular, but arbitrary, queue can be expressed in terms of the joint queue length distribution at the polling instant at the same queue in the previous cycle. The theory of MTBP's then leads to an iterative procedure to obtain the joint queue length distribution.

Although the Bernoulli service discipline does not satisfy the additivity property (except for the case $p = 0$), for the Exhaustive/Bernoulli (p) model analyzed in this paper the joint queue length distribution at polling instants is also obtained via an iterative procedure. This procedure is based on expression (25), in which the joint queue length distributions at two successive polling instants at Q_2 are related. The terms $\delta_p^{(1)}(z)$ defined in (22) can be interpreted as follows. Let n denote the number of customers served during a busy period of an ordinary $M/G/1$ queue. Moreover, let $v^2(P^{(1)})$ denote the number of type-2 arrivals during a busy period of Q_1 and let $v^1(P^{(2)}(n))$ denote the number of type-1 arrivals during a busy period at an ordinary $M/G/1$ queue with the same traffic characteristics as Q_2, at which busy period n customers are being served. Then we can write for $|z| \leq 1$ and $p \in [0, 1]$:

$$\delta_p^{(1)}(z) = E\{(1-p)^n z^{\sum_{i=1}^{v^1(P^{(2)}(n))} v^2(P^{(1)})_i}\} \ ; \tag{44}$$

the exponent of z denotes the number of type-2 arrivals during a sequence of $v^1(P^{(2)}(n))$ busy periods at Q_1. In the case $p = 0$ (exhaustive) the factor $(1-p)^n$ equals one, so that (44) expresses the p.g.f. of the number of effective replacants of a customer served at Q_2 just like in the MTBP set-up of [16]. In the analysis presented in this paper, a key role is played by relation (8). The second term in (8) would lead to a simple iterative procedure as in [16], whereas the first term (which disappears when $p = 0$) disturbs such a simple iterative solution. Nevertheless, the structure of the first term is such that the present model is still solvable by means of the more complicated iterative procedure presented in the previous sections. As a consequence, the Bernoulli service discipline seems to possess a sort of 'pseudo-additivity' property, as opposed to, e.g., the limited service disciplines.

It might be interesting to pursue this further. Moreover, it would be interesting to investigate to what extent the iterative approach of the MTBP theory can be applied to more general polling systems.

Acknowledgements: The authors are indebted to Onno Boxma and Hans Blanc for stimulating discussions and reading earlier drafts of this paper. They also wish to thank the associate editor and the referees for their useful comments.

References

[1] Blanc JPC, van der Mei RD (1992) Optimization of polling systems with Bernoulli schedules. Report FEW 563, To appear in Performance Evaluation. Department of Economics, Tilburg University The Netherlands

[2] Blanc JPC, van der Mei RD (1992) Optimization of polling systems by means of gradient methods and the power-series algorithm. Report FEW 575, Department of Economics, Tilburg University The Netherlands

[3] Boxma OJ (1986) Models of two queues: A few new views. In: Teletraffic Analysis and Computer Performance Evaluation, Boxma OJ, Cohen JW, Tijms HC (eds) North-Holland, Amsterdam, The Netherlands 75–98

[4] Boxma OJ (1989) Workloads and waiting times in single-server systems with multiple customer classes. Queueing Systems 5:185–214

[5] Boxma OJ, Groenendijk WP (1988) Two queues with alternating service and switching times. In: Queueing Theory and its Applications – Liber Amicorum for J.W. Cohen, Boxma OJ, Syski R (eds) North-Holland, Amsterdam, The Netherlands 261–282

[6] Cohen JW (1982) The single server queue (North-Holland, Amsterdam, The Netherlands; 2nd ed)

[7] Cohen JW (1988) A two-queue model with semi-exhaustive alternating service. In: Performance '87, Courtois PJ, Latouche G (eds) North-Holland, Amsterdam, The Netherlands 19–37

[8] Eisenberg M (1972) Queues with periodic service and changeover time. Operations Research 20:440–451

[9] Fricker C, Jaïbi R (1992) Monotonicity and stability of periodic polling models, Report FEW 559, Department of Economics, Tilburg University The Netherlands. To appear in Queueing Systems

[10] Fuhrmann SW (1992) A decomposition result for a class of polling models. Queueing Systems 11:109–120

[11] Groenendijk WP (1990) Conservation laws in polling systems. PhD Dissertation, University of Utrecht The Netherlands

[12] Keilson J, Servi LD (1986) Oscillating random walk models for $GI/G/1$ vacation systems with Bernoulli schedules. Journal of Applied Probability 23:790–802

[13] Levy H (1991) Binomial-gate service: A method for effective operation and optimization of polling systems. IEEE Transactions on Communications 39:1341–1350

[14] Levy H (1988) Optimization of polling systems: The fractional exhaustive service method. Dept of Comp Sc Tel Aviv University Israel

[15] Ramaswamy R, Servi LD (1988) The busy period of the $M/G/1$ vacation model with a Bernoulli schedule. Comm Stat-Stoch Models 4:507–521

[16] Resing JAC (1993) Polling systems and multi-type branching processes. Queueing Systems 13:409–426

[17] Servi LD (1986) Average delay approximations of $M/G/1$ cyclic service queues with Bernoulli schedules. IEEE Sel Areas Comm 4:813–822

[18] Takagi H (1990) Queueing analysis of polling models. In: Stochastic Analysis of Computer Communication Systems, H. Takagi (ed) North-Holland, Amsterdam, The Netherlands 267–318

[19] Tedijanto TE (1992) Exact results for the cyclic-service queue with a Bernoulli schedule. Performance Evaluation 15:89–97

[20] Watson KS (1985) Performance evaluation of cyclic service strategies – a survey. In: Performance '84, E Gelenbe (ed) North-Holland, Amsterdam, The Netherlands, 521–533

[21] Weststrate JA (1990) Waiting times in a two-queue model with exhaustive and Bernoulli service. Report FEW 437, Department of Economics, Tilburg University, The Netherlands

Received: February 1991
Revised version received: September 1993