1991

J.A.C. Resing

Polling systems and multitype branching processes

Department of Operations Research, Statistics, and System Theory Report BS-R9128 December

CWI, nationaal instituut voor onderzoek op het gebied van wiskunde en informatica

CWI is the research institute of the Stichting Mathematisch Centrum, which was founded on February 11, 1946, as a non-profit institution aiming at the promotion of mathematics, computer science, and their applications. It is sponsored by the Dutch Government through the Netherlands organization for scientific research (NWO).

Polling Systems and Multitype Branching Processes

J.A.C. Resing
CWI
P.O.Box 4079, 1009 AB Amsterdam, The Netherlands

The joint queue length process in polling systems with and without switchover times is studied. If the service discipline in each queue satisfies a certain property it is shown that the joint queue length process at polling instants of a fixed queue is a multitype branching process (MTBP) with immigration. In the case of polling models with switchover times, it turns out that we are dealing with an MTBP with immigration in each state, whereas in the case of polling models without switchover times we are dealing with an MTBP with immigration in state zero. The theory of MTBP's leads to expressions for the generating function of the joint queue length process at polling instants. Sufficient conditions for ergodicity and moment calculations are also given.

1980 Mathematics Subject Classification: 60K25, 60J80.

Keywords and Phrases: Polling systems, queue length distribution, ergodicity conditions, branching processes, immigration.

1 Introduction

In this paper we consider continuous-time cyclic polling systems. The single server who successively visits a number of different stations may or may not experience switchover times when he switches between stations. We are particularly interested in the joint queue length process in the different queues.

Polling systems have been considered in numerous papers (see the survey of Takagi [18]). For example, a large number of service disciplines has been considered. Typical service disciplines are exhaustive service (per visit the server continues to serve all customers at a station until it empties), gated service (per visit the server serves only those customers at a station which are found there upon his visit), and 1-limited service (per visit at most one customer is served at a station).

When one overviews the literature, there is a striking distinction between polling models with service disciplines that allow a rather simple analysis (including gated and exhaustive service) and polling models with service disciplines (like 1-limited service) that defy any exact analysis except for some special cases like completely symmetrical queues or like the two-queue case. It turns out that an analytical method, called the buffer occupancy method, does work for the first group of models, and does not work for the second group.

This paper is concerned with the question: Why is there such a sharp distinction between the two groups of models? We show that the theory of multitype branching processes

Report BS-R9128
ISSN 0924-0659
CWI
P.O. Box 4079, 1009 AB Amsterdam, The Netherlands

(MTBP's) with immigration plays a key role in answering this question. We shall demonstrate that the number of customers at different queues on successive moments that the server reaches a fixed queue, say queue 1, is an MTBP with immigration for the first group of models and is not an MTBP for the second group of models. This observation, in combination with results about MTBP's, immediately leads to an expression for the generating function of the stationary joint distribution of the number of customers at different queues on these moments for polling models with, e.g., exhaustive and gated service. Furthermore, it leads to the same kind of result for some new polling models. So the main contribution of this paper is the unification and generalization of some known results for models with exhaustive and gated service to a general class of polling systems, by using the framework of MTBP's.

The rest of this paper is organized as follows. In Section 2 we describe the model and the service disciplines that are considered in this paper. Section 3 is devoted to MTBP's. We recall some known results for MTBP's and prove a theorem for MTBP's with immigration in state zero. Section 4 contains the main theorems in this paper. It is shown that in our model the number of customers at different queues on successive moments that a server reaches a fixed queue is an MTBP with immigration. This leads to the expression for the generating function of the stationary joint distribution of the number of customers at different queues on these moments. Section 5 is concerned with the question : What are the ergodicity conditions for our model ? In Section 6 we do some queue length moment calculations and the paper is finished in Section 7 with conclusions and a list of possible extensions.

2 Model description

We consider a system consisting of N infinite-buffer queues, $Q_1, \ldots, Q_N$, and a single server. The service time distribution at Q_j is $B_j(\cdot)$ with first moment β_j and with Laplace-Stieltjes transform (LST) $\beta_j(\cdot)$. The server moves among the queues in a cyclic order. We consider both the model with switchover times and the model without switchover times. In the model with switchover times, the server always switches, even when the system is empty. When leaving Q_j and before moving to the next queue, the server incurs a switchover period whose duration is a random variable S_j with first moment σ_j and LST $\sigma_j(\cdot)$. In the model without switchover times, when the system is empty the server stops switching and waits at a fixed queue, say Q_1, until there is a customer arrival. Customers arrive at the queues according to independent Poisson processes with rate λ_j, $j = 1, \ldots, N$. Furthermore, arrival processes, service times and switchover times are independent. The service disciplines that we consider in this paper satisfy the following property (see Fuhrmann [4]).

Property 1 *If the server arrives at Q_i to find k_i customers there, then during the course of the server's visit, each of these k_i customers will effectively be replaced in an i.i.d. manner by a random population having probability generating function (p.g.f.) $h_i(s_1, \ldots, s_N)$, which can be any N-dimensional p.g.f..*

The exhaustive and gated service discipline both satisfy this property. In these cases the functions $h_i(s_1, \ldots, s_N)$ are given by

$$h_i(s_1, \ldots, s_N) = \theta_i(\sum_{j \neq i} \lambda_j(1 - s_j)) \tag{1}$$

and

$$h_i(s_1,\ldots,s_N) = \beta_i(\sum_j \lambda_j(1-s_j)) \tag{2}$$

respectively, where $\theta_i(\cdot)$ denotes the LST of a busy period in an M/G/1 queue with arrival rate λ_i and service time distribution $B_i(\cdot)$.

Also the more general service disciplines binomial-exhaustive (see Levy [7]) and binomial-gated (see Levy [8]) satisfy property 1. The p.g.f.'s are given by

$$h_i(s_1,\ldots,s_N) = (1-p_i)s_i + p_i\theta_i(\sum_{j\neq i} \lambda_j(1-s_j)) \tag{3}$$

and

$$h_i(s_1,\ldots,s_N) = (1-p_i)s_i + p_i\beta_i(\sum_j \lambda_j(1-s_j)) \tag{4}$$

respectively. Here the parameter p_i is the probability that a customer, who is found in Q_i at the moment that the server reaches Q_i, is served.

Remark that we allow different service disciplines at different queues. Hence, also the mixture of exhaustive and gated service disciplines, as described in Takagi [17], is contained in our model.

In Section 4 we shall show that the numbers of customers at different queues on successive moments that the server reaches Q_1, for service disciplines that satisfy property 1, constitute an MTBP with immigration. Essential in property 1 is that all k_i customers are effectively replaced in an *i.i.d.* manner. For the 1-limited service discipline, if the server arrives at Q_i to find two customers there, one of the customers is replaced by a random population having p.g.f. $\beta_i(\sum_j \lambda_j(1-s_j))$, while the other is replaced by a random population having p.g.f. s_i. Hence the 1-limited service discipline does not satisfy property 1. Also the Bernoulli service discipline does not belong to the above described class of service policies. Recall that in the Bernoulli service discipline after each service which does not leave Q_i empty, the server serves another customer with probability p_i and moves to the next queue with probability $1-p_i$. For this discipline obvious dependencies arise between the random populations that effectively replace different customers.

The following service discipline, called *Bernoulli-type*, does satisfy property 1. When the server reaches Q_i and finds k_i customers, each of these customers is independently handled in the following way. The customer is served and customers arriving during the busy period generated by this customer are served according to a Bernoulli discipline with parameter p_i, $0 \leq p_i \leq 1$. So, after each service which does not finish the busy period, the next customer in the busy period is served with probability p_i and the server stops serving the busy period with probability $1-p_i$. As far as we know this service discipline is new. It both generalizes the gated service discipline ($p_i = 0$) and the exhaustive service discipline ($p_i = 1$).

LEMMA 1 *For the Bernoulli-type service discipline the p.g.f.* $h_i(s_1,\ldots,s_N)$ *is given by*

$$h_i(s_1,\ldots,s_N) = \Phi_{p_i,i}(\sum_{j\neq i} \lambda_j(1-s_j)) + \frac{(1-p_i)\beta_i(\sum_j \lambda_j(1-s_j))}{s_i - p_i\beta_i(\sum_j \lambda_j(1-s_j))}(s_i - \Phi_{p_i,i}(\sum_{j\neq i} \lambda_j(1-s_j))), \tag{5}$$

where $\Phi_{p_i,i}(s)$ is the unique solution of

$$\Phi_{p_i,i}(s) = p_i\beta_i(s + \lambda_i(1 - \Phi_{p_i,i}(s))).$$

Proof: Define $\psi_i(r,s) = E(e^{-rT_i}s^{X_i})$, where T_i is the length of a busy period generated by a single customer and X_i is the number of customers at the end of the busy period, in an M/G/1 vacation model with Bernoulli schedule with arrival intensity λ_i, service time distribution $B_i(\cdot)$ and Bernoulli parameter p_i. Then

$$h_i(s_1,\ldots,s_N) = \psi_i(\sum_{j\neq i}\lambda_j(1-s_j), s_i),$$

and the result follows from Theorem 1 of Ramaswamy and Servi [12].

3 Multitype branching processes

We start this section with recalling some terminology and stating some results about multitype branching processes (see Athreya and Ney [1]). Assume we have a finite number N of particle types. To define the particle production we need N generating functions, each in N variables,

$$f^{(i)}(s_1,\ldots,s_N) = \sum_{j_1,\ldots,j_N\geq 0} p^{(i)}(j_1,\ldots,j_N)s_1^{j_1}\cdots s_N^{j_N},\ i = 1,\ldots,N, \tag{6}$$

where $p^{(i)}(j_1,\ldots,j_N)$ is the probability that a type i particle produces j_1 particles of type 1, j_2 of type 2,..., j_N of type N, respectively. Let m_{ij} be the expected number of type j offspring of a single type i particle, i.e. $m_{ij} = \frac{\partial f^{(i)}}{\partial s_j}(1,\ldots,1)$.

An essential role is played by the mean matrix $M = (m_{ij} : i,j = 1,\ldots,N)$. The matrix M is called primitive, if there is an n such that all entries of the matrix M^n are strictly positive. As a consequence of the Perron-Frobenius theorem (see Seneta [14]), for a non-negative primitive matrix M there exists a positive real eigenvalue λ_{max} of M such that $|\lambda| < \lambda_{max}$ for all other eigenvalues λ of M.

As mentioned before, we shall prove in Section 4 that for our model (with, in particular, the service disciplines satisfying property 1) the number of customers at different queues on successive moments that the server reaches Q_1 is an MTBP with immigration. It turns out that, in the case with switchover times, we are dealing with an MTBP with immigration *in each state*. When there are no switchover times, we are dealing with an MTBP with immigration *in state zero*. In the following two subsections we pay attention to these two types of immigration.

3.1 Multitype branching processes with immigration in each state

Consider the multitype branching process with an independent immigration component in each state. So in addition to the generating functions $f^{(i)}(s_1,\ldots,s_N), i = 1,\ldots,N$, representing the offspring distributions, an additional generating function $g(s_1,\ldots,s_N)$ is given, representing the immigration distribution, i.e.

$$g(s_1,\ldots,s_N) = \sum_{j_1,\ldots,j_N\geq 0} q(j_1,\ldots,j_N)s_1^{j_1}\cdots s_N^{j_N}, \tag{7}$$

where $q(j_1, \ldots, j_N)$ is the probability that a group of immigrants consists of j_1 particles of type 1, j_2 of type 2, ..., j_N of type N, respectively.

Define the functions $f_n(s_1, \ldots, s_N)$ inductively by

$$\begin{cases} f_0(s_1, \ldots, s_N) &= (s_1, \ldots, s_N) \\ f_n(s_1, \ldots, s_N) &= (f^{(1)}(f_{n-1}(s_1, \ldots, s_N)), \ldots, f^{(N)}(f_{n-1}(s_1, \ldots, s_N))). \end{cases} \tag{8}$$

The following theorem is due to Quine [11].

THEOREM 1 *Let $Z_n = (Z_n^{(1)}, \ldots, Z_n^{(N)})$ be a multitype branching process with immigration in each state with offspring generating functions $f^{(i)}(s_1, \ldots, s_N)$, $i = 1, \ldots, N$ and immigration generating function $g(s_1, \ldots, s_N)$. Let the mean matrix M corresponding to the branching process be primitive and its maximal eigenvalue $\lambda_{max} < 1$. Assume the Markov chain Z_n is irreducible and aperiodic. Then a necessary and sufficient condition for the existence of a stationary distribution $\pi(j_1, \ldots, j_N)$ for the process Z_n is*

$$\sum_{\substack{j_1, \ldots, j_N \geq 0 \\ j_1 + \cdots + j_N > 0}} q(j_1, \ldots, j_N) \log(j_1 + \cdots + j_N) < \infty. \tag{9}$$

When this condition is satisfied, the generating function $P(s_1, \ldots, s_N)$ of the distribution $\pi(j_1, \ldots, j_N)$ satisfies

$$P(s_1, \ldots, s_N) = \prod_{n=0}^{\infty} g(f_n(s_1, \ldots, s_N)). \tag{10}$$

Proof: See Quine [11]. The formula (10) is derived by iteration of

$$P(s_1, \ldots, s_N) = g(s_1, \ldots, s_N) P(f_1(s_1, \ldots, s_N)). \tag{11}$$

We shall use (11) in Section 6 for moment calculations.

3.2 Multitype branching processes with immigration in state zero

In this subsection we consider the same process as in the previous subsection except that there is immigration only in state zero and not in every state. We shall prove (see also Resing [13]) the following multitype version of a theorem of Pakes[10]:

THEOREM 2 *Let $Z_n = (Z_n^{(1)}, \ldots, Z_n^{(N)})$ be a multitype branching process with immigration at state zero with offspring generating functions $f^{(i)}(s_1, \ldots, s_N)$, $i = 1, \ldots, N$ and immigration generating function $g(s_1, \ldots, s_N)$. Let the mean matrix M corresponding to the branching process be primitive and its maximal eigenvalue $\lambda_{max} < 1$. Assume the Markov chain Z_n is irreducible and aperiodic, and finally assume $Z_0 = (0, \ldots, 0)$. Then a necessary and sufficient condition for the existence of a stationary distribution $\pi(j_1, \ldots, j_N)$ for the process Z_n is*

$$\sum_{\substack{j_1, \ldots, j_N \geq 0 \\ j_1 + \cdots + j_N > 0}} q(j_1, \ldots, j_N) \log(j_1 + \cdots + j_N) < \infty. \tag{12}$$

When this condition is satisfied, the generating function $P(s_1,\ldots,s_N)$ *of the distribution* $\pi(j_1,\ldots,j_N)$ *satisfies*

$$P(s_1,\ldots,s_N) = 1 - \pi(0,\ldots,0)\sum_{n=0}^{\infty}(1 - g(f_n(s_1,\ldots,s_N))), \tag{13}$$

where

$$\pi(0,\ldots,0) = [1 + \sum_{n=0}^{\infty}(1 - g(f_n(0,\ldots,0)))]^{-1}. \tag{14}$$

Proof: Because of the assumption that Z_n is aperiodic and irreducible all states of the Markov chain are equivalent. Hence we restrict our attention to the state $(0,\ldots,0)$. Let T be the recurrence time of state zero. We need the following lemma.

LEMMA 2 *For* $n \geq 1$ *we have* $\Pr(T > n) = 1 - g(f_{n-1})$, *where* $f_{n-1} := f_{n-1}(0,\ldots,0)$.

Proof: Let Y_n be the multitype branching process with the same offspring generating function as Z_n. Furthermore $Y_0 = (0,\ldots,0)$ and only at time zero there is an immigration, with generating function $g(s_1,\ldots,s_N)$. Then the process Y_n has the same recurrence time of state zero as the process Z_n and the generating function of Y_n equals $g(f_{n-1}(s_1,\ldots,s_N))$. Hence $\Pr(T > n) = \Pr(Y_n \neq 0) = 1 - g(f_{n-1})$.

Now we can continue the proof of Theorem 2. We conclude that the Markov chain Z_n is recurrent from the fact that for multitype branching processes with $\lambda_{max} < 1$, we have $f_n(s_1,\ldots,s_N) \to (1,\ldots,1)$ (see Athreya and Ney [1]) and hence $\Pr(T > n) \to 0$.

The expected recurrence time of state zero equals

$$\sum_{n=1}^{\infty} n\Pr(T = n) = \sum_{n=0}^{\infty}\Pr(T > n) = 1 + \sum_{n=1}^{\infty}(1 - g(f_{n-1})), \tag{15}$$

and hence Z_n is positive recurrent with $\pi(0,\ldots,0) = [1+\sum_{n=0}^{\infty}(1-g(f_n))]^{-1}$ iff $\sum(1-g(f_n)) < \infty$. See Kaplan [6] for the proof that this condition is equivalent with condition (12). In fact Kaplan concludes that if (12) is satisfied $\sum(1 - g(f_n(s_1,\ldots,s_N))) < \infty$ for all $(s_1,\ldots,s_N)$ with $0 \leq s_i \leq 1$, $i = 1,\ldots,N$.

It only remains to prove equation (13). Define the transition probabilities

$$p_{i_1,\ldots,i_N;j_1,\ldots,j_N} := \Pr(Z_{n+1} = (j_1,\ldots,j_N) | Z_n = (i_1,\ldots,i_N))$$

and define

$$P_{i_1,\ldots,i_N}(s_1,\ldots,s_N) := \sum_{j_1,\ldots,j_N \geq 0} p_{i_1,\ldots,i_N;j_1,\ldots,j_N} s_1^{j_1}\cdots s_N^{j_N}.$$

Then

$$P_{i_1,\ldots,i_N}(s_1,\ldots,s_N) = g(s_1,\cdots,s_N)1[(i_1,\ldots,i_N) = (0,\ldots,0)]$$
$$+[f^{(1)}(s_1,\ldots,s_N)]^{i_1}\cdots[f^{(N)}(s_1,\ldots,s_N)]^{i_N}1[(i_1,\ldots,i_N) \neq (0,\ldots,0)].$$

Now we use

$$\pi_{j_1,\ldots,j_N} = \sum_{i_1,\ldots,i_N} \pi_{i_1,\ldots,i_N} p_{i_1,\ldots,i_N;j_1,\ldots,j_N}$$

to conclude

$$\begin{aligned} P(s_1,\ldots,s_N) &= \sum_{j_1,\ldots,j_N} \pi_{j_1,\ldots,j_N} s_1^{j_1}\cdots s_N^{j_N} \\ &= \sum_{j_1,\ldots,j_N}\sum_{i_1,\ldots,i_N} \pi_{i_1,\ldots,i_N} p_{i_1,\ldots,i_N;j_1,\ldots,j_N} s_1^{j_1}\cdots s_N^{j_N} \\ &= \sum_{i_1,\ldots,i_N} \pi_{i_1,\ldots,i_N} P_{i_1,\ldots,i_N}(s_1,\ldots,s_N) \\ &= P(f_1(s_1,\ldots,s_N)) + \pi(0,\ldots,0)[g(s_1,\ldots,s_N)-1]. \end{aligned} \tag{16}$$

Iteration of this equation, together with $f_n(s_1,\ldots,s_N) \to 1$ and $\sum(g(f_n(s_1,\ldots,s_N))-1) < \infty$, yields (13).

4 Polling systems and multitype branching processes

The following two theorems are the main results in the paper.

Define the time point t_n as the time point that the server reaches Q_1 for the n-th time.

Theorem 3 *Consider a polling system with switchover times S_j with LST $\sigma_j(\cdot)$. Assume that the service discipline at Q_j satisfies property 1 with p.g.f. $h_j(s_1,\ldots,s_N)$, $j=1,\ldots,N$. Then the numbers of customers in the different queues at time points t_n constitute a multitype branching process with immigration in each state, where the offspring generating functions $f^{(i)}(s_1,\cdots,s_N)$, $i=1,\ldots,N$, are given by*

$$f^{(i)}(s_1,\ldots,s_N) = h_i(s_1,\ldots,s_i,f^{(i+1)}(s_1,\ldots,s_N),\ldots,f^{(N)}(s_1,\ldots,s_N)) \tag{17}$$

and the immigration generating function $g(s_1,\ldots,s_N)$ is given by

$$g(s_1,\ldots,s_N) = \prod_{i=1}^{N} \sigma_i\Big(\sum_{k=1}^{i}\lambda_k(1-s_k) + \sum_{k=i+1}^{N}\lambda_k(1-f^{(k)}(s_1,\ldots,s_N))\Big). \tag{18}$$

Proof: Let t_n and t_{n+1} be two consecutive time points that the server reaches Q_1. Let c_A be a customer in the system at time t_n. The customers who arrive during the service of c_A, if c_A is served in (t_n,t_{n+1}), are called the first generation offspring of c_A. The customers, who arrive during the service of those customers of the first generation offspring who are served in (t_n,t_{n+1}), are called the second generation offspring of c_A, etc. The set of all customers who belong to the offspring of c_A, including c_A, is called the ancestral line of c_A. Those customers in the ancestral line of c_A who are still in the system at time t_{n+1} are called effective replacants of c_A.

The notion of ancestral lines is taken from Fuhrmann and Cooper [5]. Note however that we restrict the ancestral line to a finite interval (t_n,t_{n+1}). If, for example, c_A is not served in (t_n,t_{n+1}) then both the ancestral line and the set of effective replacants of c_A only consist of c_A.

Let c_B be a customer who arrives during a switching interval between t_n and t_{n+1}. Similar as above, we define the ancestral line and the effective replacants of c_B.

The total collection of customers in the different queues at time t_{n+1} consists of the effective replacants of customers in the system at time t_n and the effective replacants of customers who arrive during a switching interval. From the fact that all arrival processes are Poisson processes and the fact that all service disciplines satisfy property 1, it follows immediately that we are dealing with a multitype branching process with immigration in each state. Here the offspring of a type j customer corresponds to the effective replacants of a c_A customer from Q_j and the immigration corresponds to the effective replacants of all the c_B customers.

Next we shall calculate the offspring generating functions. By definition $f^{(N)}(s_1,\ldots,s_N) = h_N(s_1,\ldots,s_N)$. Assume we have calculated $f^{(k)}(s_1,\ldots,s_N)$ for $k = i+1,\ldots,N$. Then we will calculate $f^{(i)}(s_1,\ldots,s_N)$ by conditioning on the number of customers in the ancestral line present at the moment that the server leaves Q_i. With the notation

$$\begin{aligned}
p(\underline{i}) &= \Pr\{\text{ collection of effective replacants of a } c_A \text{ customer from } Q_i \text{ consists of } i_k \text{ customers in } Q_k,\ k=1,\ldots,N\},\\
q(\underline{j}) &= \Pr\{\text{ collection of customers in the ancestral line at time that server leaves } Q_i \text{ consists of } j_k \text{ customers in } Q_k,\ k=1,\ldots,N\},\\
p(\underline{i}|\underline{j}) &= \Pr\{\text{ collection of effective replacants of a } c_A \text{ customer from } Q_i \text{ consists of } i_k \text{ customers in } Q_k,\ k=1,\ldots,N \mid \\
&\qquad \text{collection of customers in the ancestral line at time that server leaves } Q_i \text{ consists of } j_k \text{ customers in } Q_k,\ k=1,\ldots,N\},
\end{aligned}$$

we have

$$\begin{aligned}
f^{(i)}(s_1,\ldots,s_N) &= \sum_{i_1,\ldots,i_N\geq 0} p(\underline{i}) s_1^{i_1}\ldots s_N^{i_N}\\
&= \sum_{i_1,\ldots,i_N\geq 0} s_1^{i_1}\ldots s_N^{i_N} \sum_{\substack{0\leq j_k\leq i_k,k=1,\ldots,i\\ 0\leq j_{k+1},k=i+1,\ldots,N}} q(\underline{j})p(\underline{i}|\underline{j})\\
&= \sum_{j_1,\ldots,j_N\geq 0} q(\underline{j}) s_1^{j_1}\ldots s_i^{j_i} \sum_{\substack{i_k\geq j_k,k=1,\ldots,i\\ i_k\geq 0,k=i+1,\ldots,N}} p(\underline{i}|\underline{j}) s_1^{i_1-j_1}\ldots s_i^{i_i-j_i} s_{i+1}^{i_{i+1}}\ldots s_N^{i_N}\\
&= \sum_{j_1,\ldots,j_N\geq 0} q(\underline{j}) s_1^{j_1}\ldots s_i^{j_i} [f^{(i+1)}(s_1,\ldots,s_N)]^{j_{i+1}}\cdots[f^{(N)}(s_1,\ldots,s_N)]^{j_N}\\
&= h_i(s_1,\ldots,s_i,f^{(i+1)}(s_1,\ldots,s_N),\ldots,f^{(N)}(s_1,\ldots,s_N)).
\end{aligned}$$

To calculate the immigration generating function $g(s_1,\ldots,s_N)$, let us first consider the generating function $g_i(s_1,\ldots,s_N)$ of the immigration consisting of effective replacants of those c_B customers who arrive during the switching interval from Q_i to Q_{i+1}.

Introduce the notation

$$\begin{aligned}
\tilde{p}(\underline{i}) &= \Pr\{\text{ collection of effective replacants of } c_B \text{ customers arriving during switch from } Q_i \text{ consists of } i_k \text{ customers in } Q_k,\ k=1,\ldots,N\},\\
\tilde{q}(\underline{j}) &= \Pr\{\text{ collection of arrivals during switch from}
\end{aligned}$$

Q_i consists of j_k customers in Q_k, $k = 1, \ldots, N\}$,

$\tilde{p}(\underline{i}|\underline{j})$ = Pr { collection of effective replacants of c_B customers arriving during switch from Q_i consists of i_k customers in Q_k, $k = 1, \ldots, N$ | collection of arrivals during switch from Q_i consists of j_k customers in Q_k, $k = 1, \ldots, N\}$.

Conditioning on the total number of arrivals to the different queues during this switching interval, we find

$$
\begin{aligned}
g_i(s_1, \ldots, s_N) &= \sum_{i_1, \ldots, i_N \geq 0} \tilde{p}(\underline{i}) s_1^{i_1} \ldots s_N^{i_N} \\
&= \sum_{i_1, \ldots, i_N \geq 0} s_1^{i_1} \ldots s_N^{i_N} \sum_{\substack{0 \leq j_k \leq i_k, k=1,\ldots,i \\ 0 \leq j_{k+1}, k=i+1,\ldots,N}} \tilde{q}(\underline{j}) \tilde{p}(\underline{i}|\underline{j}) \\
&= \sum_{i_1, \ldots, i_N \geq 0} s_1^{i_1} \ldots s_N^{i_N} \sum_{\substack{0 \leq j_k \leq i_k, k=1,\ldots,i \\ 0 \leq j_{k+1}, k=i+1,\ldots,N}} \int_0^\infty \frac{(\lambda_1 t)^{j_1}}{j_1!} e^{-\lambda_1 t} \ldots \frac{(\lambda_N t)^{j_N}}{j_N!} e^{-\lambda_N t} dS_i(t) \tilde{p}(\underline{i}|\underline{j}) \\
&= \sum_{j_1, \ldots, j_N \geq 0} \int_0^\infty \frac{(\lambda_1 t)^{j_1}}{j_1!} e^{-\lambda_1 t} \ldots \frac{(\lambda_N t)^{j_N}}{j_N!} e^{-\lambda_N t} dS_i(t) s_1^{j_1} \ldots s_i^{j_i} \\
&\quad [f^{(i+1)}(s_1, \ldots, s_N)]^{j_{i+1}} \ldots [f^{(N)}(s_1, \ldots, s_N)]^{j_N} \\
&= \sigma_i\Big(\sum_{k=1}^{i} \lambda_k (1 - s_k) + \sum_{k=i+1}^{N} \lambda_k (1 - f^{(k)}(s_1, \ldots, s_N))\Big).
\end{aligned}
$$

The total immigration generating function $g(s_1, \ldots, s_N)$ follows of course from

$$
g(s_1, \ldots, s_N) = \prod_{i=1}^{N} g_i(s_1, \ldots, s_N). \tag{19}
$$

This completes the proof of Theorem 3.

For the case without switchover times, define the time point t_n similar as above. When the system is empty at t_n, the server stops, waits until the first customer arrival after t_n and then starts serving this customer.

THEOREM 4 *Consider a polling system without switchover times. Assume that the service discipline at Q_j satisfies property 1 with p.g.f. $h_j(s_1, \ldots, s_N)$, $j = 1, \ldots, N$. Then the numbers of customers in the different queues at time points t_n constitute a multitype branching process with immigration in state zero, where the offspring generating functions $f^{(i)}(s_1, \cdots, s_N)$, $i = 1, \ldots, N$ are given by*

$$
f^{(i)}(s_1, \ldots, s_N) = h_i(s_1, \ldots, s_i, f^{(i+1)}(s_1, \ldots, s_N), \ldots, f^{(N)}(s_1, \ldots, s_N)) \tag{20}
$$

and the immigration generating function $g(s_1,\ldots,s_N)$ *is given by*

$$g(s_1,\ldots,s_N) = \sum_{j=1}^{N} \frac{\lambda_j}{\lambda} f^{(j)}(s_1,\ldots,s_N), \tag{21}$$

where $\lambda := \sum_{i=1}^{N} \lambda_i$.

Proof: The proof is similar to the proof of Theorem 3. However, there are no c_B customers, because there are no switching intervals. The only immigration is due to customers who arrive at an empty system. With probability λ_j/λ such a customer arrives at Q_j and hence the immigration generating function in state zero is given by

$$g(s_1,\ldots,s_N) = \sum_{j=1}^{N} \frac{\lambda_j}{\lambda} f^{(j)}(s_1,\ldots,s_N). \tag{22}$$

Combination of Theorem 1 and Theorem 3 in the case with switchover times and combination of Theorem 2 and Theorem 4 in the case without switchover times gives us explicit expressions for the stationary joint distribution of the number of customers in the different queues at moments that the server reaches Q_1.

5 Ergodicity

From Theorem 1 and Theorem 2 we conclude that, both in the case of MTBP's with immigration in each state and in the case of MTBP's with immigration in state zero, the conditions

1. $\lambda_{max} < 1$,
2. $$\sum_{\substack{j_1,\ldots,j_N \geq 0 \\ j_1+\cdots+j_N>0}} q(j_1,\ldots,j_N)\log(j_1+\cdots+j_N) < \infty,$$

are sufficient for ergodicity.

In the following we shall prove that for polling systems with Bernoulli-type service discipline (and hence in particular with gated or exhaustive service discipline) the condition $\lambda_{max} < 1$ is equivalent to the condition $\rho < 1$, where $\rho = \sum_{j=1}^{N} \lambda_j \beta_j$.

From the formula

$$f^{(i)}(s_1,\ldots,s_N) = h_i(s_1,\ldots,s_i, f^{(i+1)}(s_1,\ldots,s_N),\ldots,f^{(N)}(s_1,\ldots,s_N)),$$

one easily finds the relation (with $m_{ij} := \frac{\partial f^{(i)}}{\partial s_j}(1,\ldots,1)$ and $h_{ij} := \frac{\partial h_i}{\partial s_j}(1,\ldots,1)$)

$$m_{ij} = h_{ij} \cdot 1[j \leq i] + \sum_{k=i+1}^{N} h_{ik} m_{kj}.$$

Now we need the following three lemma's. Let $H = (h_{ij})$. In the sequel relations between vectors have to be read coordinatewise, i.e. for example for two vectors x and y, $x < y$ means $x_i < y_i$ for all $i = 1,\ldots,N$.

Lemma 3 *For the Bernoulli-type service discipline we have*

1. $\rho < 1 \Rightarrow H\beta < \beta$,
2. $\rho = 1 \Rightarrow H\beta = \beta$,
3. $\rho > 1 \Rightarrow H\beta > \beta$.

Proof: For the Bernoulli-type service we have, if $p_i \neq 0$, (cf. Lemma 1)

$$\begin{aligned} h_{ij} &= (1 - \Phi_{p_i,i}(0))\frac{\lambda_j \beta_i}{1 - p_i}, \quad i \neq j, \\ h_{ii} &= 1 - \frac{1 - \rho_i}{1 - p_i}(1 - \Phi_{p_i,i}(0)), \end{aligned}$$

where $\rho_i := \lambda_i \beta_i$. Hence

$$\begin{aligned} \sum_{j=1}^{N} h_{ij}\beta_j &= (\sum_{j \neq i} \rho_j (\frac{1 - \Phi_{p_i,i}(0)}{1 - p_i}) + 1 - \frac{1 - \rho_i}{1 - p_i}(1 - \Phi_{p_i,i}(0)))\beta_i \\ &= (1 + \frac{1 - \Phi_{p_i,i}(0)}{1 - p_i}(\rho - 1))\beta_i, \end{aligned}$$

and the lemma follows.

If $p_i = 0$, then

$$\begin{aligned} h_{ij} &= \frac{\lambda_j \beta_i}{1 - \rho_i}, \quad i \neq j, \\ h_{ii} &= 0, \end{aligned}$$

and hence

$$\sum_{j=1}^{N} h_{ij}\beta_j = \frac{\beta_i}{1 - \rho_i} \sum_{j \neq i} \rho_j, \tag{23}$$

and the lemma follows.

LEMMA 4 *For a vector $\beta > 0$, we have*

1. $H\beta < \beta \Rightarrow M\beta < \beta$,
2. $H\beta = \beta \Rightarrow M\beta = \beta$,
3. $H\beta > \beta \Rightarrow M\beta > \beta$.

Proof: We prove 1, the proofs of 2 and 3 are similar. From $\sum_{j=1}^{N} m_{Nj}\beta_j = \sum_{j=1}^{N} h_{Nj}\beta_j$ we conclude $\sum_{j=1}^{N} m_{Nj}\beta_j < \beta_N$. Furthermore we have

$$\sum_{j=1}^{N} m_{ij}\beta_j = \sum_{j=1}^{i} h_{ij}\beta_j + \sum_{k=i+1}^{N} h_{ik} \sum_{j=1}^{N} m_{kj}\beta_j \tag{24}$$

and hence $\sum_{j=1}^{N} m_{kj}\beta_j < \beta_k$ for $k = i+1, \ldots, N$ implies $\sum_{j=1}^{N} m_{ij}\beta_j < \beta_i$. The lemma now follows from an induction-like argument.

LEMMA 5 (SUBINVARIANCE THEOREM) *Let $\beta > 0$ be a vector and M be a non-negative irreducible matrix with Perron Frobenius eigenvalue λ_{max}. Then*

1. $M\beta < \beta \Rightarrow \lambda_{max} < 1$,
2. $M\beta = \beta \Rightarrow \lambda_{max} = 1$,
3. $M\beta > \beta \Rightarrow \lambda_{max} > 1$.

Proof: See Seneta [14], Theorem 1.6 and Exercise 1.17.

THEOREM 5 *For polling systems with Bernoulli-type service discipline we have $\lambda_{max} < 1 \Leftrightarrow \rho < 1$.*

Proof: Follows from a combination of the Lemma's 3, 4, and 5. The only difficulty that arises is that for the exhaustive service discipline ($p_i = 0$) the matrix M is not irreducible ($m_{iN} = 0$ for all i). Hence we can not apply the subinvariance theorem directly to M. However we can apply the subinvariance theorem to O, the $(N-1) \times (N-1)$ matrix with $O_{ij} = M_{ij}$ for $1 \leq i, j \leq N-1$, and it is easily checked that the maximal eigenvalue of O is equal to the maximal eigenvalue of M.

Let us now look at the second condition

$$\sum_{\substack{j_1,\ldots,j_N \geq 0 \\ j_1+\cdots+j_N>0}} q(j_1,\ldots,j_N)\log(j_1+\cdots+j_N) < \infty. \tag{25}$$

In the case of polling systems without switchover times we have

$$\sum_{j_1,\ldots,j_N \geq 0} q(j_1,\ldots,j_N)(j_1+\cdots+j_N) = \sum_{j=1}^{N} \frac{\lambda_j}{\lambda} \sum_{k=1}^{N} m_{jk} < \infty. \tag{26}$$

In the case of polling systems with switchover times we have

$$\sum_{j_1,\ldots,j_N \geq 0} q(j_1,\ldots,j_N)(j_1+\cdots+j_N) = \sum_{j=1}^{N} (\sum_{i=1}^{j} \lambda_i + \sum_{i=j+1}^{N} \lambda_i \sum_{k=1}^{N} m_{ik})\sigma_j < \infty. \tag{27}$$

Because

$$\sum_{\substack{j_1,\ldots,j_N \geq 0 \\ j_1+\cdots+j_N>0}} q(j_1,\ldots,j_N)\log(j_1+\cdots+j_N) < \sum_{j_1,\ldots,j_N \geq 0} q(j_1,\ldots,j_N)(j_1+\cdots+j_N), \tag{28}$$

condition 2 is always satisfied if we assume that all switchover times have finite first moment. In fact condition 2 weakens the assumption of a finite first moment of all switchover times.

6 Moment calculations

The formulas (11) and (16) can be used for moment calculations. For example for the first moments in the case of an MTBP with immigration in each state we find

$$\begin{pmatrix} EZ^{(1)} \\ \vdots \\ EZ^{(N)} \end{pmatrix} = (I-M)^{-1} \begin{pmatrix} \frac{\partial g}{\partial s_1}(1,\ldots,1) \\ \vdots \\ \frac{\partial g}{\partial s_N}(1,\ldots,1) \end{pmatrix} \tag{29}$$

and in the case of an MTBP with immigration in state zero we find

$$\begin{pmatrix} EZ^{(1)} \\ \vdots \\ EZ^{(N)} \end{pmatrix} = \pi(0,\ldots,0)(I-M)^{-1} \begin{pmatrix} \frac{\partial g}{\partial s_1}(1,\ldots,1) \\ \vdots \\ \frac{\partial g}{\partial s_N}(1,\ldots,1) \end{pmatrix}. \tag{30}$$

We can use these formulas to get an explicit expression for the expected queue lengths at moments that the server reaches Q_1 in polling systems with Bernoulli-type service discipline.

From

$$m_{ij} = h_{ij} \cdot 1[j \leq i] + \sum_{k=i+1}^{N} h_{ik} m_{kj},$$

we conclude

$$m_{ij} = \sum_{(i_1,\ldots,i_l) \in S_{ij}} h_{i_1 i_2} \cdots h_{i_{l-1} i_l},$$

with

$$S_{ij} = \{(i_1,\ldots,i_l) : l \geq 2, 1 \leq i_j \leq N, i_1 = i, i_l = j, i_1 < i_2 < \ldots < i_{l-1}, i_{l-1} \geq j\}.$$

Furthermore

$$g(s_1,\ldots,s_N) = \prod_{i=1}^{N} \sigma_i \Big(\sum_{k=1}^{i} \lambda_k (1-s_k) + \sum_{k=i+1}^{N} \lambda_k (1 - f^{(k)}(s_1,\ldots,s_N)) \Big) \tag{31}$$

in the case with switchover times, and

$$g(s_1,\ldots,s_N) = \sum_{i=1}^{N} \frac{\lambda_i}{\lambda} f^{(i)}(s_1,\ldots,s_N) \tag{32}$$

in the case without switchover times. Hence

$$\frac{\partial g}{\partial s_j}(1,\ldots,1) = \lambda_j \sum_{i=j}^{N} \sigma_i + \sum_{i=1}^{N} \sum_{k=i+1}^{N} \sigma_i \lambda_k m_{kj} \tag{33}$$

in the case with switchover times, and

$$\frac{\partial g}{\partial s_j}(1,\ldots,1) = \sum_{i=1}^{N} \frac{\lambda_i}{\lambda} m_{ij} \tag{34}$$

in the case without switchover times. All these formulas together give us explicit expressions for the first moments.

7 Conclusions and extensions

In this paper we show that the joint queue length process in the different queues of a polling system with Poisson arrivals, at time points that the server reaches a fixed queue, is a multitype branching process with immigration. This enables us to find an explicit expression for the generating function of the joint queue length process. The service discipline has to satisfy a certain property, recently introduced by Fuhrmann[4]. The analysis leads to known results for exhaustive and gated service disciplines, but also to new results for other service disciplines. For example, for binomial-exhaustive, binomial-gated and a mixture of gated and exhaustive service disciplines until now only expressions for moments were given.

The analysis presented in this paper can easily be extended to the following cases:

- models with Poisson batch arrivals (see Levy and Sidi [9]).
- models with customer branching and customer routing (see Sidi and Levy [15]).
- polling systems with fixed polling tables instead of cyclic polling systems (see Baker and Rubin [2]).
- models with globally gated service (see Boxma, Levy and Yechiali [3]).
- discrete time polling systems (see Takagi [16]).

Acknowledgement: The author likes to thank O.J. Boxma for valuable discussions related to this paper.

8 References

1. Athreya, K.B. and Ney, P.E. (1972). Branching Processes. Springer-Verlag, Berlin.
2. Baker, J.E. and Rubin, I. (1987). Polling with a general service order table. IEEE Trans. Commun. 35, 283-288.
3. Boxma, O.J., Levy, H. and Yechiali, U. (1990). Cyclic reservation schemes for efficient operation of multiple queue single server systems. To appear in Annals of Operations Research.
4. Fuhrmann, S.W. (1991). A decomposition result for a class of polling models. IBM Research Report, Zürich, May 1991.
5. Fuhrmann, S.W. and Cooper, R.B. (1985). Stochastic decompositions in the M/G/1 queue with generalized vacations. Oper. Res. 33, 1117-1129.
6. Kaplan, N. (1973). The multitype Galton-Watson process with immigration, Ann. Probab. 1, 947-953.
7. Levy, H. (1988). Optimization of polling systems: The fractional exhaustive service method. Technical report, Department of Computer Science, Tel Aviv University.
8. Levy, H. (1989). Analysis of cyclic polling systems with binomial gated service. In: Performance of Distributed and Parallel Systems, T.Hasegawa, H.Takagi, Y. Takahashi (eds.), Elsevier Science Publishers, North-Holland, 127-139.

9. Levy, H. and Sidi, M. (1989). Polling systems with correlated arrivals. IEEE Infocom 1989, 907-913.

10. Pakes, A.G. (1971). A branching process with a state dependent immigration component, Adv. Appl. Prob. 3, 301-314.

11. Quine, M.P. (1970). The multitype Galton-Watson process with immigration, J. Appl. Prob. 7, 411-422.

12. Ramaswamy, R. and Servi, L.D. (1988). The busy period of the M/G/1 vacation model with a Bernoulli schedule, Commun. Statist.-Stochastic Models 4, 507-521.

13. Resing, J.A.C. (1990). Asymptotic Results in Feedback Systems. Ph.D. Thesis, Technical University Delft.

14. Seneta, E. (1981). Non-negative Matrices and Markov Chains. Second Edition. Springer-Verlag, New York.

15. Sidi, M. and Levy, H. (1990). Customers routing in polling systems. In: Performance 1990, P.J.B.King, I.Mitrani and R.J.Pooley (eds.), North-Holland, Amsterdam, 319-331.

16. Takagi, H. (1986). Analysis of Polling Systems. The MIT Press, Cambridge, Massachusetts.

17. Takagi, H. (1989). Analysis of polling systems with a mixture of exhaustive and gated service disciplines. J. Oper. Res. Society Japan 32, 450-461.

18. Takagi, H. (1990). Queueing analysis of polling models: an update. Stochastic Analysis of Computer and Communications Systems, H. Takagi (ed.), Elsevier Science Pub., 267-318.