Centrum voor Wiskunde en Informatica

REPORT RAPPORT

Polling models with and without switchover times

S.C. Borst, O.J. Boxma

Department of Operations Research, Statistics, and System Theory

Report BS-R9421 May 1994

Copyright © Stichting Mathematisch Centrum
P.O. Box 94079, 1090 GB Amsterdam (NL)
Kruislaan 413, 1098 SJ Amsterdam (NL)
Telephone +31 20 592 9333
Telefax +31 20 592 4199

Polling Models with and without Switchover Times

S.C. Borst, O.J. Boxma*
CWI
P.O. Box 94079, 1090 GB Amsterdam, The Netherlands

Abstract

We consider two different single-server cyclic polling models: (i) a model with zero switchover times, and (ii) a model with non-zero switchover times, in which the server keeps cycling when the system is empty.

For both models we relate the steady-state queue length distribution at a queue to the queue length distributions at server visit beginning and visit completion instants at that queue; in this way we obtain a short proof of the Fuhrmann-Cooper decomposition. For the large class of polling systems that allow a multitype branching process interpretation, we expose a strong relation between the queue length, as well as waiting time, distributions in the two models. This leads to a reduced computational complexity in determining the waiting time moments.

AMS Subject Classification (1991): 60K25, 68M20.
Keywords & Phrases: polling system, switchover times, queue lengths, waiting times, multitype branching processes, Fuhrmann-Cooper decomposition.

1 Introduction

Consider a single-server cyclic polling system: a single server S visits n queues $Q_1, \dots, Q_n$ in cyclic order, serving customers at Q_i according to some service discipline D_i, $i = 1, \dots, n$. Such polling models have many applications in, e.g., computer-communications, manufacturing and traffic. The abundant literature on polling systems (cf. Takagi [15]) contains the exact analysis of polling systems for a large number of service disciplines, like the exhaustive and gated disciplines.

Polling models are closely related to queueing models with vacations of the server; the intervisit time of the server w.r.t. some Q_i can be viewed as a vacation. Accordingly, the concept of queue length decomposition for queues with vacations (cf. Fuhrmann and Cooper [8]) has proven to be very fruitful for the analysis of polling models. It has also led to the concept of work decomposition in polling models (cf. Boxma [3]), that relates the amount of work in a polling system *with* switchover times to the amount of work in the same polling system but *without* switchover times. Hitherto, polling models with switchover times and polling models without switchover times had usually been treated separately, often via different approaches; the problem with simply letting the switchover times tend to zero in a polling model with non-zero switchover times is that the number of polling epochs in an idle period tends to

*The work of the second author was supported in part by the European Grant BRA-QMIPS of CEC DG XIII.

Report BS-R9421
ISSN 0924-0659
CWI
P.O. Box 94079, 1090 GB Amsterdam, The Netherlands

infinity, leading to degenerate distributions at such epochs (cf. [6], [11]). The relation between the polling models with and without switchover times has further been exposed in some recent papers of Cooper et al. [4], Fuhrmann [9] and Srinivasan et al. [14]; these authors consider waiting times and joint queue lengths instead of workloads. The main goal of this paper is to unify and generalize some of their results.

Firstly, in Section 2, we use a beautiful relation of Eisenberg [5] (see also [6]), that has received too little attention in the literature, to relate the Generating Functions (GF) of queue lengths at various instants in the polling system (visit beginnings and endings, service beginnings and endings). We observe that this relation, that was presented by Eisenberg in [5] for the case of non-zero switchover times, also holds for the case of zero switchover times, and we show how it almost instantaneously gives a simple proof of the above-mentioned Fuhrmann-Cooper decomposition for the queue lengths at the various queues of a polling system.

Eisenberg's relation leads to an expression for the joint queue length GF at service completion instants at some queue into the joint queue length GF's at the beginning and end of a visit of the server to that queue. The latter GF's can be easily related, and determined, for the important class of polling models in which each queue satisfies the following property:

Property 1.1
If S 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 $h_i(z_1, \dots, z_n)$, which can be any n-dimensional probability generating function.

Resing [12] (see also Fuhrmann [7]) has studied polling systems that satisfy this property; this includes the case of exhaustive or gated service at all queues, but it excludes the case of 1-limited service at any queue. He has shown that, for this class of polling systems, the joint queue length process at visit instants of a fixed queue is a so-called *multitype branching process* (MTBP) with immigration. The theory of multitype branching processes (cf. Athreya and Ney [1], Resing [13]) now leads to an expression for the generating function of the joint queue length process at visit beginning (polling) instants. In Section 3, for models that satisfy Property 1.1, we use a slightly adapted version of the results of Resing [12] to relate the joint queue length GF's at visit beginning and visit ending instants, and then to obtain those GF's. The results expose a close similarity between the cases with and without switchover times. In Section 4 we determine the steady-state marginal queue length GF at Q_i, both for the model with and the model without switchover times, and we relate the transforms for those two cases; similarly for the waiting time Laplace-Stieltjes Transform (LST) at Q_i. In Section 5 we describe how the relationship between the waiting times in both models can be exploited to reduce the complexity of numerical moment calculations.

2 A GENERAL POLLING MODEL

We study the following polling model. A single server S visits n infinite-buffer queues $Q_1, \dots, Q_n$ in cyclic order $Q_1, \dots, Q_n, Q_1, \dots$. Customers arrive at Q_i according to a Poisson process with rate λ_i; $\lambda := \sum_{i=1}^{n} \lambda_i$. A customer at Q_i, to be called a type-i customer, requires a service time $\mathbf{B}_i$ which is a random variable with distribution $B_i(\cdot)$ with mean β_i and LST

$\beta_i(\cdot)$. Define $\rho_i := \lambda_i\beta_i$ as the traffic intensity at Q_i. Denote by $\rho := \sum_{i=1}^{n}\rho_i$ the total traffic intensity. When S moves from Q_i to Q_{i+1}, this requires a switchover time $\mathbf{S}_i$ which is a random variable with distribution $S_i(\cdot)$ with mean s_i and LST $\sigma_i(\cdot)$. In this model with switchover times, S switches even when the system is empty. We also consider the variant of this model in which all switchover times are *zero*. In the latter model, when the system has become empty S makes a full cycle (viz. passes all the queues once) and subsequently stops right before Q_1. All this requires zero time. When the first new customer arrives, S cycles along the queues to that customer. The choice of Q_1 is arbitrary, but for the application of the MTBP theory in Section 3 it will be necessary to fix one position. There we shall discuss this issue in more detail.

We assume that all the usual independence assumptions hold between the service times, the switchover times and the interarrival times. We assume that the ergodicity conditions are fulfilled, and we restrict ourselves to results for the stationary situation.

Eisenberg [5] studies this model for the case with switchover times and with the exhaustive service discipline at all queues (while briefly discussing the case of gated service at all queues). He considers the following four quantities, with $\mathbf{N}$ denoting a vector of numbers of customers at $Q_1, \dots, Q_n$ and N a realization:

$S_{b_i}(t, N) :=$ # service beginnings at Q_i in $(0, t)$ in which the queue length state is N;
$S_{c_i}(t, N) :=$ # service completions at Q_i in $(0, t)$ in which the queue length state is N;
$V_{b_i}(t, N) :=$ # visit beginnings at Q_i in $(0, t)$ in which the queue length state is N;
$V_{c_i}(t, N) :=$ # visit completions at Q_i in $(0, t)$ in which the queue length state is N.

In the case of a service or visit completion the state is defined as what exists immediately after the departure of the customer.

Eisenberg [5] now makes the crucial observation that each time a visit beginning or a service completion occurs, this coincides with either a service beginning or a visit completion. Hence

$$V_{b_i}(t, N) + S_{c_i}(t, N) = S_{b_i}(t, N) + V_{c_i}(t, N). \tag{2.1}$$

We observe that (2.1) not only holds for the case with switchover times and exhaustive or gated service, but for any service discipline, and also for the case of zero switchover times. Define the following equilibrium state probabilities for this polling model:

$S_{b_{i,N}} := \Pr(\mathbf{N} = N,\ S \text{ is at } Q_i \mid \text{service beginning instant})$;
$S_{c_{i,N}} := \Pr(\mathbf{N} = N,\ S \text{ is at } Q_i \mid \text{service completion instant})$;
$V_{b_{i,N}} := \Pr(\mathbf{N} = N \mid \text{visit beginning at } Q_i)$;
$V_{c_{i,N}} := \Pr(\mathbf{N} = N \mid \text{visit completion at } Q_i)$.

Eisenberg [5] divides all four terms in (2.1) by the total number of service completions at all queues in $(0, t)$, and takes the limit for $t \to \infty$. He thus relates those four equilibrium state probabilities:

$$\gamma_i V_{b_{i,N}} + S_{c_{i,N}} = S_{b_{i,N}} + \gamma_i V_{c_{i,N}}. \tag{2.2}$$

Here γ_i is the long-term ratio of the number of visit completions at Q_i to the number of customers that are handled by the system; in this cyclic polling model $\gamma_i \equiv \gamma$. Written in terms of generating functions, this yields:

$$\gamma V_{b_i}(z) + S_{c_i}(z) = S_{b_i}(z) + \gamma V_{c_i}(z), \tag{2.3}$$

for $z = (z_1, \ldots, z_n)$, $|z_j| \leq 1$, $j = 1, \ldots, n$; here $V_{b_i}(z)$ and $V_{c_i}(z)$ denote the GF of the joint queue length distribution at visit beginnings and visit completions of Q_i, while $S_{b_i}(z)$ and $S_{c_i}(z)$ denote the GF of the joint distribution of queue length vector and server position at service beginnings and service completions.
Now Eisenberg observes that $S_{c_i}(z)$ and $S_{b_i}(z)$ are related via

$$S_{c_i}(z) = S_{b_i}(z)\beta_i(\sum_{j=1}^{n} \lambda_j(1 - z_j))/z_i, \tag{2.4}$$

for $|z_j| \leq 1$, $j = 1, \ldots, n$.
It follows from (2.3) and (2.4) that

$$S_{c_i}(z) = \frac{\gamma\beta_i(\sum\limits_{j=1}^{n} \lambda_j(1 - z_j))}{z_i - \beta_i(\sum\limits_{j=1}^{n} \lambda_j(1 - z_j))}[V_{b_i}(z) - V_{c_i}(z)]. \tag{2.5}$$

Eisenberg, considering the variant with switchover times and exhaustive service, subsequently expresses $V_{b_i}(z)$ in $V_{c_{i-1}}(z)$. For the moment we refrain from that (see Section 3), but we observe that formula (2.5) is generally valid for the class of polling systems described in the beginning of this section (with and without switchover times).
Taking in (2.5) $z = (1, \ldots, 1, y, 1, \ldots, 1)$, with y as i-th argument, and dividing by the probability λ_i/λ that an arbitrary service completion is at Q_i, gives the queue length GF at Q_i at a service completion instant at Q_i. A standard up- and down-crossing argument combined with PASTA shows that the queue length distribution at Q_i at its service completion instants, at its customer arrival instants and in steady state are all the same. Hence, with $\mathbf{N}_i$ the steady-state queue length at Q_i and with $\mathbf{X}_i$ and $\mathbf{Y}_i$ the steady-state queue lengths at Q_i at the beginning and end of a visit at that queue: for $|y| \leq 1$,

$$\mathrm{E}(y^{\mathbf{N}_i}) = \frac{\lambda}{\lambda_i}\frac{\gamma\beta_i(\lambda_i(1-y))}{y - \beta_i(\lambda_i(1-y))}[\mathrm{E}(y^{\mathbf{X}_i}) - \mathrm{E}(y^{\mathbf{Y}_i})]. \tag{2.6}$$

Note that $1/\gamma$ equals the mean number of customers served per cycle, hence also the mean number of customers that arrive per cycle: $1/\gamma = \lambda\mathbf{EC}$, with $\mathbf{EC}$ the mean length of one cycle of S along the queues (a cycle w.r.t. Q_i is defined as the period between the start of two successive visits to Q_i; it is easily seen that the mean cycle time is the same for all i). Because S spends on the average a fraction ρ_i of a cycle at Q_i, we can write:

$$\mathbf{EX}_i - \mathbf{EY}_i = \lambda_i(1 - \rho_i)\mathbf{EC} = \frac{\lambda_i(1 - \rho_i)}{\lambda\gamma}. \tag{2.7}$$

From (2.6) and (2.7), for $|y| \leq 1$:

$$\mathrm{E}(y^{\mathbf{N}_i}) = \frac{(1-\rho_i)(1-y)\beta_i(\lambda_i(1-y))}{\beta_i(\lambda_i(1-y)) - y}\frac{\mathrm{E}(y^{\mathbf{Y}_i}) - \mathrm{E}(y^{\mathbf{X}_i})}{(1-y)(\mathbf{EX}_i - \mathbf{EY}_i)}. \tag{2.8}$$

The first term in the rhs denotes the GF $\mathrm{E}(y^{\mathbf{N}_{i|M/G/1}})$ of the queue length distribution in the 'corresponding' isolated $M/G/1$ queue of Q_i with arrival rate λ_i and service time distribution $B_i(\cdot)$. Now consider the second term. Observe that $\mathbf{Y}_i$ not only denotes the queue length at Q_i at the end of a visit to that queue, but also the queue length at Q_i at the beginning of an intervisit period for that queue; while $\mathbf{X}_i$ denotes the queue length at Q_i at the *end* of such an intervisit period. With this interpretation $\mathbf{X}_i$ is in distribution equal to the sum of $\mathbf{Y}_i$ and the number of Poisson arrivals during the intervisit period. Introducing $\mathbf{Z}_i$, a stochastic variable with distribution the queue length distribution at an arbitrary instant in an intervisit period of Q_i, we can now write (cf. [2]):

$$\mathrm{E}(y^{\mathbf{Z}_i}) = \frac{\mathrm{E}(y^{\mathbf{Y}_i}) - \mathrm{E}(y^{\mathbf{X}_i})}{(1-y)(\mathrm{E}\mathbf{X}_i - \mathrm{E}\mathbf{Y}_i)}. \tag{2.9}$$

This relation appears in the polling literature for various special cases (e.g., for exhaustive vacation models, where $\mathbf{Y}_i \equiv 0$). It holds also for non-Poisson arrivals, when $\mathbf{Z}_i$ is defined as the queue length that is observed by an arbitrary customer who arrives at Q_i during an intervisit period.
(2.8) and (2.9) together yield the well-known Fuhrmann-Cooper queue length decomposition [8], applied to a queue in a polling model with or without switchover times: for $|y| \leq 1$,

$$\mathrm{E}(y^{\mathbf{N}_i}) = \mathrm{E}(y^{\mathbf{N}_{i|M/G/1}})\mathrm{E}(y^{\mathbf{Z}_i}). \tag{2.10}$$

Remark 2.1
Fuhrmann and Cooper [8] state five conditions under which their decomposition holds:
(i). Customers arrive at Q_i according to a Poisson process of rate λ_i.
(ii). All customers arriving at Q_i are eventually served.
(iii). Customers enter service in an order that is independent of their service times.
(iv). Service is non-preemptive.
(v). The rules that govern when the server begins and ends visit periods to Q_i do not anticipate future jumps of the Poisson arrival process at Q_i.
These assumptions indeed hold in our polling model, and are implicitly used in the derivation of (2.10). The above proof, with as key steps (2.1), (2.4) and (2.9), in fact also holds for vacation models without a polling context. Note that the relations $1/\gamma = \lambda\mathrm{E}\mathbf{C}$ and (2.7) hold generally for queues with some vacation (intervisit) mechanism. We refer to Keilson & Servi [10] for another short proof of the Fuhrmann-Cooper decomposition.

□

The waiting time LST at Q_i immediately follows from the queue length GF for $\mathbf{N}_i$, when we assume that within each queue customers are served in order of arrival. Denote by $\mathbf{R}_i$ and $\mathbf{W}_i$ respectively the sojourn time and the waiting time of an arbitrary type-i customer. Denote by $\mathbf{R}_{i|M/G/1}$ and $\mathbf{W}_{i|M/G/1}$ respectively the sojourn time and the waiting time of an arbitrary customer in the 'corresponding' isolated $M/G/1$ queue of Q_i. Application of the distributional form of Little's law, cf. Keilson & Servi [10], yields for $|y| \leq 1$:

$$\mathrm{E}(y^{\mathbf{N}_i}) = \mathrm{E}(\mathrm{e}^{-\lambda_i(1-y)\mathbf{R}_i}), \tag{2.11}$$

and hence, putting $\omega = \lambda_i(1-y)$,

$$\mathrm{E}(\mathrm{e}^{-\omega \mathbf{R}_i}) = \mathrm{E}(\mathrm{e}^{-\omega \mathbf{R}_{i|M/G/1}})\mathrm{E}((1-\omega/\lambda_i)^{\mathbf{Z}_i}). \tag{2.12}$$

Because the waiting time and service time of a customer at Q_i are independent, it follows that

$$\mathrm{E}(\mathrm{e}^{-\omega \mathbf{W}_i}) = \mathrm{E}(\mathrm{e}^{-\omega \mathbf{W}_{i|M/G/1}})\mathrm{E}((1-\omega/\lambda_i)^{\mathbf{Z}_i}). \tag{2.13}$$

In Section 4 we shall return to this relation, for the case of polling models that satisfy Property 1.1.

3 The joint queue length distribution at polling epochs

In the previous section we have seen that Eisenberg's results [5] yield simple relations between the generating function $S_{c_i}(z)$ of the joint queue length vector at service completion epochs (or $S_{b_i}(z)$, at service beginning epochs) and the generating functions $V_{b_i}(z)$ and $V_{c_i}(z)$ of the joint queue length vector at visit beginning and visit completion epochs. We now restrict ourselves to polling models for which the service discipline at each queue satisfies Property 1.1. Property 1.1 prescribes how each of the customers present at Q_i at the visit beginning is replaced by independent families of customers at its visit completion. This enables one to express $V_{c_i}(\cdot)$ nicely into $V_{b_i}(\cdot)$, and to finally determine each of the functions $V_{b_i}(\cdot)$ (after which the GF's $V_{c_i}(\cdot)$, $S_{c_i}(\cdot)$ and $S_{b_i}(\cdot)$ follow). In our analysis we follow Resing [12].

First a remark about the ergodicity conditions. In the sequel we assume that $\rho < 1$ and $s_i < \infty$ for all i. Resing [12] proves that for the subclass of so-called Bernoulli-type service disciplines these conditions together constitute sufficient ergodicity conditions. His proof is based on the observation that for these Bernoulli-type service disciplines the derivatives of $h_i(z_1,\dots,z_n)$ take the form $\frac{\partial h_i(z)}{\partial z_j}|_{z=1}= \lambda_j \alpha_i \frac{\beta_i}{1-\rho_i}$, $i \neq j$, $\frac{\partial h_i(z)}{\partial z_i}|_{z=1}= 1-\alpha_i$, with α_i some coefficient in $(0,1]$ determined by the parameters of Q_i. It may be easily verified however that the latter form of the derivatives applies for *any* non-idling service discipline that satisfies Property 1.1 with $h_i(z_1,\dots,z_n) \neq z_i$. As the proof in Resing [12] further does not rely on the specific form of α_i, we may conclude that $\rho < 1$ and $s_i < \infty$ for all i together constitute sufficient ergodicity conditions for *any* non-idling service discipline that satisfies Property 1.1 with $h_i(z_1,\dots,z_n) \neq z_i$.

Property 1.1 implies that

$$V_{c_i}(z) = V_{b_i}(z_1,\dots,z_{i-1},h_i(z),z_{i+1},\dots,z_n). \tag{3.1}$$

In the case of gated service $h_i(z)$ is simply the GF of the joint distribution of the numbers of arrivals at all queues during one service time at Q_i: $h_i(z) = \beta_i(\sum_{j=1}^{n} \lambda_j(1-z_j))$.

In the case of exhaustive service: $h_i(z) = \zeta_i(\sum_{j \neq i} \lambda_j(1-z_j))$, with $\zeta_i(\cdot)$ the LST of the length of the busy period in the 'corresponding' isolated $M/G/1$ queue of Q_i.

Next we relate $V_{b_i}(z)$ to $V_{c_{i-1}}(z)$.

In case of non-zero switchover times:

$$V_{b_i}(z) = V_{c_{i-1}}(z)\sigma_{i-1}(\sum_{j=1}^{n} \lambda_j(1-z_j)). \tag{3.2}$$

In case of zero switchover times (in the sequel we add a superscript 0 for that case, to distinguish its quantities from those for non-zero switchover times):

$$V^0_{b_i}(z) = V^0_{c_{i-1}}(z), \tag{3.3}$$

for $i = 2, \ldots, n$. The relation between $V^0_{b_1}(z)$ and $V^0_{c_n}(z)$ deserves special attention, because of our convention, stated at the beginning of Section 2, concerning the behavior of the server in an empty system. When all queues in the model without switchover times have become empty, S in our convention makes a full cycle and subsequently stops right before Q_1 (all this requires no time). When the first new customer arrives, S cycles along the queues to that customer. The consequence of this is that when the system is empty at the start of a visit to Q_1, then the next visit to Q_1 does not take place until a customer has arrived. We can write:

$$V^0_{b_1}(z) = V^0_{c_n}(z) - V^0_{b_1}(0)[1 - g^0(z)], \tag{3.4}$$

with

$$g^0(z) := \sum_{i=1}^{n} \frac{\lambda_i}{\lambda} z_i. \tag{3.5}$$

$g^0(\cdot)$ represents the 'immigration process' of the MTBP: it is the GF of the arrival process of customers during periods in which the system is empty.
Substituting (3.1) into (3.2), respectively (3.3) and (3.4), we can relate $V_{b_i}(\cdot)$ to $V_{b_{i-1}}(\cdot)$. We distinguish between the two cases of non-zero and zero switchover times. In both cases the following branching functions play a crucial role, thus establishing the link between both cases.
Define

$$f(z) := (f_1(z), \ldots, f_n(z)), \tag{3.6}$$

with

$$f_i(z) := h_i(z_1, \ldots, z_i, f_{i+1}(z), \ldots, f_n(z)) \tag{3.7}$$

for $| z_j | \leq 1$, $j = 1, \ldots, n$. This is the *offspring* GF, the GF of the joint distribution of the numbers of customers at the end of a cycle w.r.t. Q_1 that are *descendants* of a type-i customer. In this branching process setting, a descendant of some customer K is a customer that has arrived during the service time of K or of one of its descendants.
Define

$$f^{(0)}(z) := z,$$

$$f^{(k)}(z) := f(f^{(k-1)}(z)), \qquad k \geq 1,$$

for $|z_j| \leq 1$, $j = 1, \ldots, n$.

Case I: Zero switchover times
Substituting (3.1) into (3.3),

$$V^0_{b_i}(z) = V^0_{b_{i-1}}(z_1, \ldots, z_{i-2}, h_{i-1}(z), z_i, \ldots, z_n) \tag{3.8}$$

for $i = 2, \dots, n$. Starting from (3.4) and (3.1) for $i = n$, and subsequently using (3.8) for $i = n, n-1, \dots, 2$, one obtains

$$V_{b_1}^0(z) = V_{b_1}^0(f(z)) - V_{b_1}^0(0)[1 - g^0(z)]. \tag{3.9}$$

Iterating (3.9) yields

$$V_{b_1}^0(z) = 1 - V_{b_1}^0(0)\sum_{k=0}^{\infty}[1 - g^0(f^{(k)}(z))], \tag{3.10}$$

with

$$V_{b_1}^0(0) = \left[1 + \sum_{k=0}^{\infty}[1 - g^0(f^{(k)}(0))]\right]^{-1}, \tag{3.11}$$

the infinite sum being convergent when the ergodicity conditions are fulfilled.

Introduce, for $|z_j| \leq 1$, $j = 1, \dots, n$,

$$H(z) := \sum_{k=0}^{\infty}\sum_{i=1}^{n}\lambda_i(1 - f_i^{(k)}(z)). \tag{3.12}$$

Then we can write

$$V_{b_1}^0(z) = 1 - V_{b_1}^0(0)\sum_{k=0}^{\infty}\sum_{i=1}^{n}\frac{\lambda_i}{\lambda}(1 - f_i^{(k)}(z)) = 1 - V_{b_1}^0(0)H(z)/\lambda, \tag{3.13}$$

$$V_{b_1}^0(0) = [1 + H(0)/\lambda]^{-1}. \tag{3.14}$$

Remark 3.1
Although we shall sometimes find it convenient to concentrate on Q_1, it should be noted that our convention for the position of S in an empty system does not affect the waiting time and queue length distributions.
In fact our convention slightly differs from that of Resing [12], who assumes that in an empty system S immediately stops right *behind* Q_1 and hence takes $g^0(z) = \sum_{i=1}^{n}\frac{\lambda_i}{\lambda}f_i(z)$. Our convention enables us to simultaneously apply the MTBP theory and Eisenberg's approach.
□

Case II: Non-zero switchover times
Substituting (3.1) into (3.2),

$$V_{b_i}(z) = V_{b_{i-1}}(z_1, \dots, z_{i-2}, h_{i-1}(z), z_i, \dots, z_n)\sigma_{i-1}(\sum_{j=1}^{n}\lambda_j(1 - z_j)). \tag{3.15}$$

Applying (3.15) n times (which corresponds to following the server during one full cycle w.r.t. Q_1),

$$V_{b_1}(z) = V_{b_1}(f(z))g(z), \tag{3.16}$$

with

$$g(z) = \prod_{i=1}^{n} \sigma_i(\sum_{j=1}^{i} \lambda_j(1 - z_j) + \sum_{j=i+1}^{n} \lambda_j(1 - f_j(z))). \tag{3.17}$$

$g(\cdot)$ represents the 'immigration process' of this MTBP: it is the GF of the vector of all customers that either have arrived in the switchover periods of the past cycle (measured w.r.t. Q_1), or are descendants of such customers. Iterating (3.16) yields

$$V_{b_1}(z) = \prod_{k=0}^{\infty} g(f^{(k)}(z)) = \prod_{k=0}^{\infty}\prod_{i=1}^{n} \sigma_i(\sum_{j=1}^{i} \lambda_j(1 - f_j^{(k)}(z)) + \sum_{j=i+1}^{n} \lambda_j(1 - f_j^{(k+1)}(z))), \tag{3.18}$$

the infinite product being convergent when the ergodicity conditions are fulfilled.

From (3.13) and (3.18) we see that $V_{b_1}(z)$ as well as $V_{b_1}^0(z)$ is determined by $\sum_{j=1}^{n} \lambda_j(1-f_j^{(k)}(z))$. For *constant switchover times* the connection becomes even closer, as (3.18) reduces to a simpler expression, that we present for future reference:

$$V_{b_1}(z) = \exp[-\sum_{k=0}^{\infty}\sum_{i=1}^{n} s_i\{\sum_{j=1}^{i} \lambda_j(1 - f_j^{(k)}(z)) + \sum_{j=i+1}^{n} \lambda_j(1 - f_j^{(k+1)}(z))\}]. \tag{3.19}$$

Using (3.12), and introducing the mean total switchover time $s := \sum_{i=1}^{n} s_i$, we can rewrite (3.19) into

$$V_{b_1}(z) = \exp[-sH(z) + \sum_{j=1}^{n} \lambda_j(1 - z_j)\sum_{i=1}^{j-1} s_i]. \tag{3.20}$$

4 Marginal queue lengths and waiting times

In the previous section the queue length GF's $V_{b_i}(z)$ and $V_{b_i}^0(z)$ at visit beginning instants have been determined for the class of cyclic polling models in which Property 1.1 holds for all service disciplines. In Section 2 we already obtained a decomposition for the GF of the marginal queue length distribution at Q_i, and for the waiting time LST at Q_i, into a corresponding $M/G/1$ term and a term involving $\mathrm{E}(y^{\mathbf{X}_i})$ and $\mathrm{E}(y^{\mathbf{Y}_i})$ (via the GF $\mathrm{E}(y^{\mathbf{Z}_i})$). In particular, introducing

$\tilde{h}_i(y) := h_i(1, \ldots, 1, y, 1, \ldots, 1)$,
$\tilde{V}_{b_i}(y) := V_{b_i}(1, \ldots, 1, y, 1, \ldots, 1)$,
$\tilde{V}_{b_i}^0(y) := V_{b_i}^0(1, \ldots, 1, y, 1, \ldots, 1)$,

with y as i-th argument, it follows from (2.9) and (3.1) for the case of non-zero switchover times that

$$\mathrm{E}(y^{\mathbf{Z}_i}) = \frac{\tilde{V}_{b_i}(\tilde{h}_i(y)) - \tilde{V}_{b_i}(y)}{(1-y)\tilde{V}_{b_i}'(1)(1-\tilde{h}_i'(1))}; \tag{4.1}$$

the same result holds for the case of zero switchover times, replacing $\tilde{V}_{b_i}(\cdot)$ by $\tilde{V}_{b_i}^0(\cdot)$ in (4.1). Similarly indicating queue lengths, and waiting times, by a superscript 0 in the case of zero times, we find from (2.10) and (2.13):

$$\mathrm{E}(y^{\mathbf{N}_i}) = \mathrm{E}(y^{\mathbf{N}_i^0}) \frac{[\tilde{V}_{b_i}(\tilde{h}_i(y)) - \tilde{V}_{b_i}(y)]\tilde{V}_{b_i}^{0\prime}(1)}{[\tilde{V}_{b_i}^0(\tilde{h}_i(y)) - \tilde{V}_{b_i}^0(y)]\tilde{V}_{b_i}'(1)}, \tag{4.2}$$

$$\mathrm{E}(\mathrm{e}^{-\omega \mathbf{W}_i}) = \mathrm{E}(\mathrm{e}^{-\omega \mathbf{W}_i^0}) \frac{[\tilde{V}_{b_i}(\tilde{h}_i(1-\omega/\lambda_i)) - \tilde{V}_{b_i}(1-\omega/\lambda_i)]\tilde{V}_{b_i}^{0\prime}(1)}{[\tilde{V}_{b_i}^0(\tilde{h}_i(1-\omega/\lambda_i)) - \tilde{V}_{b_i}^0(1-\omega/\lambda_i)]\tilde{V}_{b_i}'(1)}. \tag{4.3}$$

For exhaustive service $\tilde{h}_i(\cdot) \equiv 1$; for gated service $\tilde{h}_i(y) = \beta_i(\lambda_i(1-y))$. Differentiating (2.13) and (4.3) once, putting $\omega = 0$,

$$\mathbf{EW}_i = \frac{\lambda_i \beta_i^{(2)}}{2(1-\lambda_i\beta_i)} + \frac{\tilde{V}_{b_i}''(1)}{2\tilde{V}_{b_i}'(1)\lambda_i}(1+\tilde{h}_i'(1)) - \frac{\tilde{h}_i''(1)}{2(1-\tilde{h}_i'(1))\lambda_i}, \tag{4.4}$$

and

$$\mathbf{EW}_i - \mathbf{EW}_i^0 = [\frac{\tilde{V}_{b_i}''(1)}{2\tilde{V}_{b_i}'(1)\lambda_i} - \frac{\tilde{V}_{b_i}^{0\prime\prime}(1)}{2\tilde{V}_{b_i}^{0\prime}(1)\lambda_i}](1+\tilde{h}_i'(1)). \tag{4.5}$$

Let us now (without loss of generality) concentrate on $\mathbf{W}_1$ and $\mathbf{W}_1^0$. Introducing

$\tilde{f}_i^{(k)}(y) := f_i^{(k)}(y, 1, \ldots, 1)$,

it follows from (3.13) that

$$\tilde{V}_{b_1}^0(y) = 1 - V_{b_1}^0(0)\tilde{H}(y)/\lambda, \tag{4.6}$$

with

$$\tilde{H}(y) := \sum_{k=0}^{\infty}\sum_{i=1}^{n} \lambda_i(1 - \tilde{f}_i^{(k)}(y)). \tag{4.7}$$

In particular, $\tilde{V}_{b_1}^{0\prime}(1) = -V_{b_1}^0(0)\tilde{H}'(1)/\lambda$. It follows from (3.20) and (4.7) that $\tilde{V}_{b_1}'(1) = -s\tilde{H}'(1)$, and hence

$$\tilde{V}_{b_1}'(1) = \tilde{V}_{b_1}^{0\prime}(1)s\lambda/V_{b_1}^0(0). \tag{4.8}$$

For exhaustive service, from (4.3), (4.6), (4.8),

Corollary 4.1

$$\mathrm{E}(\mathrm{e}^{-\omega \mathbf{W}_1}) = \mathrm{E}(\mathrm{e}^{-\omega \mathbf{W}_1^0}) \frac{1 - \tilde{V}_{b_1}(1-\omega/\lambda_1)}{s\tilde{H}(1-\omega/\lambda_1)}, \tag{4.9}$$

which is Theorem 1 in Srinivasan et al. [14].

For constant switchover times, it follows from (3.13) and (3.20) that

$$\tilde{V}_{b_1}(y) = \exp[-s\tilde{H}(y)], \tag{4.10}$$

and hence we have

Corollary 4.2

$$\mathrm{E}(\mathrm{e}^{-\omega \mathbf{W}_1}) = \mathrm{E}(\mathrm{e}^{-\omega \mathbf{W}_1^0})\frac{\exp[-s\tilde{H}(\tilde{h}_1(1-\omega/\lambda_1))] - \exp[-s\tilde{H}(1-\omega/\lambda_1)]}{s[\tilde{H}(1-\omega/\lambda_1) - \tilde{H}(\tilde{h}_1(1-\omega/\lambda_1))]}, \tag{4.11}$$

which generalizes Theorem 2 in Srinivasan et al. [14] (there the result is obtained for exhaustive service, with $\tilde{h}_1(\cdot) \equiv 1$).

Remark 4.1
Substituting $\sigma_i(\omega) = 1 - s_i\omega + o(s_i)$ for $s_i \to 0$, $i = 1, \dots, n$, in (3.18) yields that

$$\begin{aligned} \tilde{V}_{b_1}(y) &= 1 - \sum_{i=1}^{n} s_i \sum_{k=0}^{\infty} \left[\sum_{j=1}^{i} \lambda_j(1 - \tilde{f}_j^{(k)}(y)) + \sum_{j=i+1}^{n} \lambda_j(1 - \tilde{f}_j^{(k+1)}(y)) \right] + o(s) \\ &= 1 - s\tilde{H}(y) + o(s). \end{aligned}$$

Hence, for $s \to 0$ in (4.9), $\mathbf{W}_1$ approaches $\mathbf{W}_1^0$ in distribution. Using (4.3) and (3.13), the same statement follows for other service disciplines satisfying Property 1.1.

□

5 Computational aspects

In this last section we describe how the relationship between the waiting times in the models with and without switchover times may be exploited to reduce the complexity of moment calculations. For ease of presentation we focus on the first moment, but similar observations hold for the higher moments.

To determine $\mathrm{E}\mathbf{W}_1$ and $\mathrm{E}\mathbf{W}_1^0$ according to formulas (4.4) and (4.5), we need to compute the quantities $\tilde{V}'_{b_1}(1)$, $\tilde{V}''_{b_1}(1)$, $\tilde{V}^{0\prime}_{b_1}(1)$, and $\tilde{V}^{0\prime\prime}_{b_1}(1)$. The quantities $\tilde{h}'_1(1)$ and $\tilde{h}''_1(1)$ occurring in (4.4) and (4.5) may simply be determined from the service discipline at Q_1.

We first introduce some notation. Denote $\phi_i^{(k)} := \frac{d}{dy}\tilde{f}_i^{(k)}(y)_{|y=1}$, $\psi_i^{(k)} := \frac{d^2}{dy^2}\tilde{f}_i^{(k)}(y)_{|y=1}$, $i = 1, \dots, n$, $k \geq 0$. Define $\Phi_i := \sum\limits_{k=0}^{\infty} \phi_i^{(k)}$, $\Psi_i := \sum\limits_{k=0}^{\infty} \psi_i^{(k)}$, $i = 1, \dots, n$, $\Phi := \sum\limits_{i=1}^{n} \lambda_i \Phi_i$, $\Psi := \sum\limits_{i=1}^{n} \lambda_i \Psi_i$.

Differentiating (4.6),

$$\tilde{V}^{0\prime}_{b_1}(1) = \frac{V^0_{b_1}(0)}{\lambda} \sum_{k=0}^{\infty} \sum_{i=1}^{n} \lambda_i \phi_i^{(k)} = \frac{V^0_{b_1}(0)}{\lambda}\Phi; \tag{5.1}$$

$$\tilde{V}^{0\prime\prime}_{b_1}(1) = \frac{V^0_{b_1}(0)}{\lambda} \sum_{k=0}^{\infty} \sum_{i=1}^{n} \lambda_i \psi_i^{(k)} = \frac{V^0_{b_1}(0)}{\lambda}\Psi. \tag{5.2}$$

Taking $z = (y, 1, \dots, 1)$ in (3.18), differentiating w.r.t. to y,

$$\tilde{V}'_{b_1}(1) = s \sum_{k=0}^{\infty} \sum_{i=1}^{n} \lambda_i \phi_i^{(k)} = s\Phi; \tag{5.3}$$

$$\begin{aligned}
\tilde{V}''_{b_1}(1) &= \left(\tilde{V}'_{b_1}(1)\right)^2 + s\sum_{k=0}^{\infty}\sum_{i=1}^{n}\lambda_i\psi_i^{(k)} \\
&+ \sum_{k=0}^{\infty}\sum_{i=1}^{n}\left(s_i^{(2)} - s_i^2\right)\left(\sum_{j=1}^{i}\lambda_j\phi_j^{(k)} + \sum_{j=i+1}^{n}\lambda_j\phi_j^{(k+1)}\right)^2 \\
&= \left(\tilde{V}'_{b_1}(1)\right)^2 + s\Psi + \sum_{i=1}^{n}\left(s_i^{(2)} - s_i^2\right)\chi_i,
\end{aligned} \tag{5.4}$$

with $\chi_i = \sum_{k=0}^{\infty}\left(\sum_{j=1}^{i}\lambda_j\phi_j^{(k)} + \sum_{j=i+1}^{n}\lambda_j\phi_j^{(k+1)}\right)^2$.
The quantities Φ, Ψ, and χ_i, $i = 1,\dots,n$, may be computed as follows. Taking $z = (y,1,\dots,1)$ in (3.7),

$$\tilde{f}_i^{(k+1)}(y) = h_i(\tilde{f}_1^{(k)}(y),\dots,\tilde{f}_i^{(k)}(y),\tilde{f}_{i+1}^{(k+1)}(y),\dots,\tilde{f}_n^{(k+1)}(y)). \tag{5.5}$$

Define $\mathbf{T}_i$ as the visit time at Q_i generated by an arbitrary type-i customer present at the start of a visit to Q_i. Then $\frac{\partial}{\partial z_j}h_i(z)_{|z=1} = \lambda_j\mathrm{E}\mathbf{T}_i$, $\frac{\partial^2}{\partial z_j\partial z_l}h_i(z)_{|z=1} = \lambda_j\lambda_l\mathrm{E}(\mathbf{T}_i^2)$, $j \neq i$, $l \neq i$. Define $\mathbf{V}_i$ as the total visit time at Q_i in a cycle. From $\mathrm{E}\mathbf{V}_i = \mathrm{E}\mathbf{X}_i\mathrm{E}\mathbf{T}_i$, $\mathrm{E}\mathbf{X}_i = \mathrm{E}\mathbf{Y}_i + \lambda_i(1-\rho_i)\mathrm{E}\mathbf{C}$, $\mathrm{E}\mathbf{V}_i = \rho_i\mathrm{E}\mathbf{C}$, $\mathrm{E}\mathbf{Y}_i = \mathrm{E}\mathbf{X}_i\frac{\partial}{\partial z_i}h_i(z)_{|z=1}$, it immediately follows that $\frac{\partial}{\partial z_i}h_i(z)_{|z=1} = [\lambda_i - 1/\beta_i]\,\mathrm{E}\mathbf{T}_i + 1$.
Thus, differentiating (5.5),

$$\begin{aligned}
\phi_i^{(k+1)} &= \sum_{j=1}^{i}\frac{\partial}{\partial z_j}h_i(z)_{|z=1}\phi_j^{(k)} + \sum_{j=i+1}^{n}\frac{\partial}{\partial z_j}h_i(z)_{|z=1}\phi_j^{(k+1)} \\
&= \mathrm{E}\mathbf{T}_i\left[\sum_{j=1}^{i}\lambda_j\phi_j^{(k)} + \phi_i^{(k)}/\mathrm{E}\mathbf{T}_i - \phi_i^{(k)}/\beta_i + \sum_{j=i+1}^{n}\lambda_j\phi_j^{(k+1)}\right];
\end{aligned} \tag{5.6}$$

$$\begin{aligned}
\psi_i^{(k+1)} &= \sum_{j=1}^{i}\frac{\partial}{\partial z_j}h_i(z)_{|z=1}\psi_j^{(k)} + \sum_{j=i+1}^{n}\frac{\partial}{\partial z_j}h_i(z)_{|z=1}\psi_j^{(k+1)} \\
&+ \sum_{j=1}^{i}\sum_{l=1}^{i}\frac{\partial^2}{\partial z_j\partial z_l}h_i(z)_{|z=1}\phi_j^{(k)}\phi_l^{(k)} + 2\sum_{j=1}^{i}\sum_{l=i+1}^{n}\frac{\partial^2}{\partial z_j\partial z_l}h_i(z)_{|z=1}\phi_j^{(k)}\phi_l^{(k+1)} \\
&+ \sum_{j=i+1}^{n}\sum_{l=i+1}^{n}\frac{\partial^2}{\partial z_j\partial z_l}h_i(z)_{|z=1}\phi_j^{(k+1)}\phi_l^{(k+1)} \\
&= \mathrm{E}\mathbf{T}_i\left[\sum_{j=1}^{i}\lambda_j\psi_j^{(k)} + \psi_i^{(k)}/\mathrm{E}\mathbf{T}_i - \psi_i^{(k)}/\beta_i + \sum_{j=i+1}^{n}\lambda_j\psi_j^{(k+1)}\right] \\
&+ \mathrm{E}(\mathbf{T}_i^2)\left(\sum_{j=1}^{i}\lambda_j\phi_j^{(k)} + \sum_{j=i+1}^{n}\lambda_j\phi_j^{(k+1)}\right)^2 \\
&+ \phi_i^{(k)}\left[2\sum_{j=1}^{i-1}\tau_{ij}\phi_j^{(k)} + \tau_{ii}\phi_i^{(k)} + 2\sum_{j=i+1}^{n}\tau_{ij}\phi_j^{(k+1)}\right],
\end{aligned} \tag{5.7}$$

with $\tau_{ij} := \frac{\partial^2}{\partial z_i z_j} h_i(z)_{|z=1} - \lambda_i \lambda_j \mathrm{E}(\mathbf{T}_i^2)$.
Summing (5.6) and (5.7) over $k \geq 0$, we obtain

$$\frac{\Phi_i}{\beta_i} - \frac{\phi_i^{(0)}}{\mathrm{E}\mathbf{T}_i} + \sum_{j=i+1}^{n} \lambda_j \phi_j^{(0)} = \Phi, \tag{5.8}$$

and

$$\frac{\Psi_i}{\beta_i} - \frac{\psi_i^{(0)} + \mathrm{E}(\mathbf{T}_i^2)\chi_i + \xi_i}{\mathrm{E}\mathbf{T}_i} + \sum_{j=i+1}^{n} \lambda_j \psi_j^{(0)} = \Psi, \tag{5.9}$$

with $\xi_i = \sum_{k=0}^{\infty} \phi_i^{(k)} \left[2 \sum_{j=1}^{i-1} \tau_{ij} \phi_j^{(k)} + \tau_{ii} \phi_i^{(k)} + 2 \sum_{j=i+1}^{n} \tau_{ij} \phi_j^{(k+1)} \right]$.
Multiplying (5.8) and (5.9) by ρ_i, summing over $i = 1, \ldots, n$, using $\phi_1^{(0)} = 1$, $\phi_i^{(0)} = 0$, $i = 2, \ldots, n$, $\psi_i^{(0)} = 0$, $i = 1, \ldots, n$, we obtain

$$\Phi = \frac{\rho_1/\mathrm{E}\mathbf{T}_1}{1-\rho}, \tag{5.10}$$

and

$$\Psi = \frac{\sum_{i=1}^{n} \rho_i \left[\mathrm{E}(\mathbf{T}_i^2)\chi_i + \xi_i \right] / \mathrm{E}\mathbf{T}_i}{1-\rho}. \tag{5.11}$$

Note that Φ, and hence also $\tilde{V}'_{b_1}(1) = s\Phi$, the mean queue length at Q_1 at polling epochs, do not depend on the service disciplines at the other queues and only depend on the individual parameters of the other queues through ρ and s.
For exhaustive service, i.e., $\mathrm{E}\mathbf{T}_i = \beta_i/(1-\rho_i)$, $\mathrm{E}(\mathbf{T}_i^2) = \beta_i^{(2)}/(1-\rho_i)^3$, $\tau_{ij} = -\lambda_i \lambda_j \mathrm{E}(\mathbf{T}_i^2)$, the expressions for χ_i and ξ_i, $i = 1, \ldots, n$, reduce to those obtained by Srinivasan et al. [14].
For constant switchover times the last term in (5.4) drops out, so that according to formulas (4.5), (5.1)-(5.4), and (5.10) $\mathbf{EW}_1 - \mathbf{EW}_1^0 = (1 + \tilde{h}'_1(1))\mathrm{E}\mathbf{C}\beta_1/(2\mathrm{E}\mathbf{T}_1)$, which illustrates the fact that the relationship between the waiting times in the models with and without switchover times then takes a remarkably simple form, cf. (4.11). In particular, for exhaustive and gated service $\mathbf{EW}_1 - \mathbf{EW}_1^0 = (1-\rho_1)\mathrm{E}\mathbf{C}/2$ and $\mathbf{EW}_1 - \mathbf{EW}_1^0 = (1+\rho_1)\mathrm{E}\mathbf{C}/2$, respectively, as obtained in [9] and [4] by using different techniques.
The bulk of the computational effort is involved with the calculation of the quantities χ_i and ξ_i, $i = 1, \ldots, n$, which may be done completely *independent* of the switchover times. The coefficients $\phi_i^{(k)}$, $i = 1, \ldots, n$, $k \geq 0$, needed for these calculations, may be computed recursively from (5.6), supplemented with $\phi_1^{(0)} = 1$, $\phi_i^{(0)} = 0$, $i = 2, \ldots, n$. The number of elementary operations (additions, multiplications) involved is $O(n\log_\rho(\epsilon))$ with ϵ the level of accuracy desired. Once the quantities χ_i and ξ_i, $i = 1, \ldots, n$, have been calculated, $\mathbf{EW}_1$ may be computed for any value of the switchover times in $O(n)$ elementary operations.

Acknowledgement The authors are indebted to J.A.C. Resing for useful discussions, and to M.M. Srinivasan for showing them the preprint [14].

References

[1] Athreya, K.B., Ney, P.E. (1972). Branching Processes (Springer Verlag, Berlin).

[2] Borst, S.C. (1994). Polling systems with multiple coupled servers. CWI Report BS-R9408.

[3] Boxma, O.J. (1989). Workloads and waiting times in single-server systems with multiple customer classes. *Queueing Systems* **5**, 185-214.

[4] Cooper, R.B., Niu, S.-C., Srinivasan, M.M. (1992). A decomposition theorem for polling models: the switchover times are effectively additive. To appear in *Operations Research.*

[5] Eisenberg, M. (1972). Queues with periodic service and changeover time. *Oper. Res.* **20**, 440-451.

[6] Eisenberg, M. (1993). The polling system with a stopping server. Report AT & T Bell Laboratories, Holmdel, NY.

[7] Fuhrmann, S.W. (1981). Performance analysis of a class of cyclic schedules, Bell Laboratories technical memorandum 81-59531-1.

[8] Fuhrmann, S.W., Cooper, R.B. (1985). Stochastic decompositions in the $M/G/1$ queue with generalized vacations. *Oper. Res.* **33**, 1117-1129.

[9] Fuhrmann, S.W. (1992). A decomposition result for a class of polling models. *Queueing Systems* **11**, 109-120.

[10] Keilson, J., Servi, L.D. (1990). The distributional form of Little's law and the Fuhrmann-Cooper decomposition. *Oper. Res. Lett.* **9**, 239-247.

[11] Levy, H., Kleinrock, L. (1991). Polling systems with zero switch-over periods: a general method for analyzing the expected delay. *Perf. Eval.* **13**, 97-107.

[12] Resing, J.A.C. (1993). Polling systems and multitype branching processes. *Queueing Systems* **13**, 409-426.

[13] Resing, J.A.C. (1990). Asymptotic Results in Feedback Systems. PhD Thesis, Technical University Delft.

[14] Srinivasan, M.M., Niu, S.-C., Cooper, R.B. (1993). Relating polling models with nonzero and zero switchover times. Report University of Tennessee.

[15] Takagi, H. (1990). Queueing analysis of polling models: an update. In: H. Takagi (ed.), Stochastic Analysis of Computer and Communication Systems (North-Holland Publ. Cy., Amsterdam), pp. 267-318.