Polling systems

O.J. Boxma

CWI
P.O. Box 94079, 1090 GB Amsterdam, The Netherlands
Faculty of Economics, Tilburg University
P.O. Box 90153, 5000 LE Tilburg, The Netherlands

A polling system is a queueing system in which several queues are attended by a single server. Spurred by various important applications, the field of polling systems is going through a period of feverish activity. The first part of this paper surveys some of the main developments. The second part generalizes the theory of polling systems to the case in which the customer arrival process depends on the position of the server, and to the case in which customers travel from queue to queue.

1 Introduction

It has been a great pleasure to write this paper on the mathematical analysis of the single-server polling system in honour of a truly devoted server. In a sometimes almost literally painstaking way, Cor Baayen saw to it as director of SMC that both LAW and CWI, and also both its mathematics and computer science groups, were served in an equally fair manner. He has strongly stimulated research at the interface of mathematics and computer science. His far-reaching vision has been crucial in realizing the INSP support for CWI in the eighties, which in its turn made it possible to build up a research group on the mathematical analysis of the performance of computer systems.

Consider the following situation. A director of a research institute divides his attention among several activities: scientific, financial, personnel matters, representative activities. Suppose that he devotes his energy for a while (a 'session') to tasks of a scientific nature, then switches to finance, etc. During a session other new tasks of the same type, as well as of different type, may be generated; furthermore, a task may have to be reconsidered in future sessions ('feedback'). The director is interested in the evolution of his workload, the

numbers of tasks of all types, etc. These quantities clearly depend on the way in which the offered load fluctuates over time; but the director can also influence the process by a judicious choice of the order of his activities and of the time he reserves for a session. The framework in which these matters can be studied is that of single-server queueing models. More precisely, it is the framework of *polling models*.

A polling model is a queueing model in which customers (tasks) arrive at a set of queues $Q_1, ..., Q_N$ according to some stochastic arrival process, requiring some stochastic amount of service. A single server B visits the queues in a fixed order to provide service. We assume throughout the paper that it is the cyclic order $Q_1, ..., Q_N, Q_1, ...$ (cf. Fig. 1).

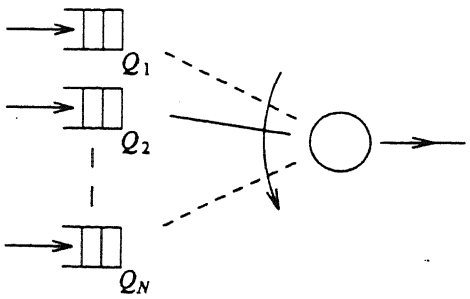

FIGURE 1. Queueing model of a polling system

When B visits Q_i and it is not empty, then B serves customers in a session at Q_i according to some service discipline. The most common service disciplines are:

- *1-limited*: serve just one customer (if at least one is present)
- *exhaustive*: serve customers until the queue is empty
- *gated*: serve precisely those customers that were already present at the start of the session

When Q_i is empty, or the session is completed, then B switches to Q_{i+1}. This may require some switchover time, which is represented by a stochastic variable.

The assumptions about the stochastic nature of the arrival process, service times and switchover times are introduced to represent the usually inherently random nature of customer behaviour, as well as a lack of detailed information. Moreover, a probability distribution for, say, service times may also represent an aggregate of in itself known, constant but distinct, service times of several types of customers. The purpose of the analysis of a polling model is to determine the performance of (several variants of) the underlying system, and

eventually to optimize system behaviour. Due to the stochasticity assumptions one can at most make probabilistic statements about the main performance measures of a polling model, like workload of the server, numbers of customers at the various queues, or their waiting times.
The analysis and optimization of polling systems has in recent years received an enormous amount of attention, and much progress has been made. It has also been one of the key research topics of the performance analysis group at CWI; cf. the PhD Theses of W.P. Groenendijk [10] and S.C. Borst [1]. Therefore it seems appropriate to briefly review the main developments, with some emphasis on contributions from the latter group. This review is presented in Section 2. In Section 3 we discuss a generalization of the standard polling model, in two directions that so far have received hardly any attention:
(i) The arrival rate of customers at the various queues may depend on the position of the server: information on which queue the server is presently visiting, and hence on which queue it will visit next, may influence the generation of new tasks.
(ii) Instead of leaving the system, customers may be routed to another (or the same) queue after having received a service. A customer's required service time at a queue may depend both on that queue and on the number of services it has already received.
We show how, for an important class of service disciplines, these generalizations can be analyzed in full detail. Crucial in this analysis is the application of the theory of multitype branching processes.
The above-mentioned features of feedback and customer information arise quite naturally in our director example; in the remainder of this section we mention several other applications of polling models.

Applications of polling models
Polling models arise in situations in which there are multiple customer classes sharing a common resource which is available to only one customer class at a time. The oldest polling model in the queueing literature concerns a patrolling repairman, who consecutively inspects a number of machines to check whether a breakdown has occurred and to restore such breakdowns [12]. In this example the server is the repairman, the queues are the machines, and the customers represent the breakdowns.
The application that gave polling models their name is a time-sharing computer system consisting of a number of terminals connected by multidrop lines to a central computer. The data transfer from the terminals to the computer (and back) is controlled via a 'polling scheme' in which the computer 'polls' the terminals, requesting their data, one terminal at a time. In this example the server represents the central computer, the queues are the terminals and the customers are the data.
The interest in polling models was strongly revived by the study of message transmission protocols in local area networks. Many communication systems provide a broadcast channel which is shared by all connected stations. When

two or more stations wish to transmit simultaneously, a conflict arises. The rules for either resolving or preventing such conflicts are referred to as 'multi-access protocols'. An important conflict-free protocol is the *token ring* protocol. In a token ring local area network, several stations (terminals, file servers, hosts, gateways, etc.) are connected to a common transmission medium in a ring topology. A special bit sequence called the *token* is passed from one station to the next; a station that 'possesses the token' is allowed to transmit a message. After completion of its transmission the station releases the token, giving the next station in turn an opportunity to transmit. This situation can be represented by a polling model with 1-limited service at each queue; the server is the token, the queues are the stations and the customers are the messages. Variants of the above-described token-passing mechanism give rise to related polling models, with e.g. exhaustive service at the queues. A queueing analysis of these polling models yields insight into the (dis)advantages of the various access protocols, and allows system designers to make performance predictions. We refer the reader to Takagi [18] and Grillo [9] for surveys on polling applications in respectively computer- and communication networks.
Other application areas of polling models include:

- robotics in manufacturing (a single machine processes several types of parts, incurring switchover times for changing tools)
- traffic signal control (the green light represents the availability of the server for a queue of vehicles)
- the operation of elevators (*multiple* servers are interesting here: is it better to have a concentration of elevators in a central area, or should they be dispersed over the building?)
- packet transfer protocols in B-ISDN (in such Broadband Integrated Services Digital Networks, channel access will be alternately granted to voice, video and data messages, all digitized into 53-byte packets)

The characteristic feature of all these applications is that the server is 'moving' between queues, implying that the priorities of the queues are dynamically (e.g., cyclically) changing. This sharply contrasts with classic *static* priority queueing models, where one type of customers always has priority over other customer types.

2 Analysis of polling systems

In this section we briefly review the exact analysis of the standard cyclic polling system. After a detailed model description we consecutively consider workloads, waiting times and queue lengths.

Model description
We here describe the standard cyclic polling model; in Section 3 we extend this

model in several ways. Customers arrive at N queues $Q_1, \ldots, Q_N$ with infinite waiting rooms according to N independent Poisson processes, with rates $\lambda_1, \ldots, \lambda_N$. Customers who arrive at Q_i are called type-i customers. Server B visits the queues in the cyclic order $Q_1, \ldots, Q_N, Q_1, \ldots$. Upon his visit to a queue, he serves one or more customers (if present) according to some service discipline like 1-limited, gated or exhaustive service (cf. Section 1). The service times of type-i customers are independent, identically distributed stochastic variables; their distribution is $B_i(\cdot)$, with first moment β_i, second moment $\beta_i^{(2)}$ and Laplace-Stieltjes Transform (LST) $\beta_i(\cdot)$. The switchover times of B between Q_i and Q_{i+1} are independent, identically distributed stochastic variables, with first moment s_i, second moment $s_i^{(2)}$ and LST $\sigma_i(\cdot)$. The total switchover time of B in one cycle has first and second moment s respectively $s^{(2)}$. We assume that the interarrival, service and switchover processes are mutually independent.

The offered traffic ρ_i at Q_i is defined as $\rho_i := \lambda_i \beta_i$, and the total offered traffic load is $\rho := \sum_{i=1}^{N} \rho_i$. Obviously $\rho < 1$ is a necessary condition for steady-state distributions of workloads, waiting times and queue lengths etc. to exist. When all switchover times are zero, this condition is also sufficient; otherwise the situation may be much more complicated, and in particular the service disciplines may influence the stability condition (e.g., in 1-limited service B is forced to spend time switching after each service). See Fricker and Jaïbi [8] for an extensive discussion of these stability issues. We assume in the sequel that steady-state distributions of all quantities under consideration exist.

The workload process

Consider first the case that all switchover times are zero. Then B is always working as long as there is at least one customer anywhere in the system. The amount of work in the system evolves in a way that does not depend on the order of service *of* the queues and *within* the queues, or on the service disciplines at the queues; this is the principle of *work conservation* (cf. Heyman and Sobel [13], p. 418). Hence, for any service discipline at the queues of the cyclic polling system, the amount of work is distributed as the amount of work in the 'corresponding single server queue' with FCFS (First Come First Served) order of service. Since the superposition of N independent Poisson processes is again a Poisson process, that 'corresponding single server queue' is an M/G/1 queue with arrival rate $\Lambda := \sum_{i=1}^{N} \lambda_i$ and with service time distribution $B(\cdot) := \sum_{i=1}^{N} (\lambda_i/\Lambda) B_i(\cdot)$.

Now consider the case that not all switchover times are zero. The principle of work conservation is clearly violated. However, it has been shown in [4] that a principle of *work decomposition* holds: the steady-state amount of work $\mathbf{V}_{with}$ in the polling system *with* switchover times is related to the steady-

state amount of work $\mathbf{V}_{without}$ in the 'corresponding polling system' *without* switchover times (hence in the above-mentioned 'corresponding M/G/1 queue') via

$$\mathbf{V}_{with} \stackrel{d}{=} \mathbf{V}_{without} + \mathbf{Y}, \tag{1}$$

where $\mathbf{Y}$ is the steady-state amount of work present in the system at an epoch in which B is not serving; $\stackrel{d}{=}$ denotes equality in distribution. Moreover, $\mathbf{V}_{without}$ and $\mathbf{Y}$ are independent. The distribution of $\mathbf{V}_{without}$ is known from M/G/1 theory. The distribution of $\mathbf{Y}$ can be determined in a number of cases, but with considerable effort. The mean $\mathrm{E}\mathbf{Y}$, on the other hand, is very easily determined for virtually any set of service disciplines at the various queues - which turns out to be most useful for deriving mean waiting times, as we'll see in formula (4) below.

REMARK 2.1
The proof of (1) as presented in [4] is based on three concepts which are sketchily indicated below.
(i) As long as B is serving, the amount of work evolves in exactly the same way as if B would be serving according to the LCFS (Last Come First Served) rule.
(ii) Characteristically for LCFS, an amount of work $\mathbf{Y}$ found by a customer C upon his arrival in a switchover period is not served until C has been served, plus all customers who arrive during C's service (C's offspring), plus all customers who arrive during those services, etc. (together - including himself - forming C's 'ancestral line').
(iii) The time period required to serve the ancestral line of C is distributed as the busy period in the above-mentioned 'corresponding M/G/1 queue'.
Since the principle of work conservation implies that *during* such a busy period the amount of work evolves in the same way, regardless whether service is FCFS or LCFS, combination of (i), (ii) and (iii) shows that the workload $\mathbf{V}_{with}$ is distributed as the superposition of $\mathbf{Y}$ and $\mathbf{V}_{without}$.
Another proof of (1), communicated to the author by B.T. Doshi, proceeds as follows. Assume for simplicity that the densities of the distributions of $\mathbf{V}_{with}$ and $\mathbf{Y}$ exist; denote them by $v(\cdot)$ and $y(\cdot)$, and denote their Laplace transforms by $\phi(\cdot)$ and $\eta(\cdot)$. Equating the downcrossing and upcrossing rates of level $x > 0$ gives:

$$v(x) - (1-\rho)y(x) = \Lambda \int_{0-}^{x} (1 - B(x-z))v(z)dz.$$

Combining this relation with $v(0) = (1-\rho)y(0)$ and taking Laplace transforms leads (with $\beta(\cdot)$ the LST of $B(\cdot)$) to:

$$\phi(\omega) - (1-\rho)\eta(\omega) = \Lambda \frac{1-\beta(\omega)}{\omega}\phi(\omega).$$

Hence

$$\phi(\omega) = \frac{(1-\rho)\omega}{\omega - \Lambda + \Lambda\beta(\omega)}\eta(\omega), \tag{2}$$

which proves the decomposition into two independent components: $\phi(\omega)$ is the product of the transform of the distribution of $\mathbf{V}_{without}$ (a well-known M/G/1 expression) and the transform $\eta(\omega)$ of the distribution of $\mathbf{Y}$. See [3] for a generalization of this principle of work decomposition, and for applications to various polling models with a *non-cyclic* visit pattern.

Waiting times

We restrict ourself here to *mean* waiting times. Denote the mean waiting time of type-i customers by $\mathrm{E}\mathbf{W}_i$, and the mean number of waiting type-i customers by $\mathrm{E}\mathbf{X}_i$. These quantities are related via Little's formula: $\mathrm{E}\mathbf{X}_i = \lambda_i \mathrm{E}\mathbf{W}_i$. It is easy to relate the mean workload in queueing models with Poisson arrivals to mean queue lengths, and hence to mean waiting times. Indeed, under mild restrictions that are fulfilled in the standard polling model described earlier in this section, we can write (cf. [3]):

$$\mathrm{E}\mathbf{V}_{with} = \sum_{i=1}^{N} \beta_i \mathrm{E}\mathbf{X}_i + \sum_{i=1}^{N} \rho_i \frac{\beta_i^{(2)}}{2\beta_i}. \tag{3}$$

Now take means in (1) and combine the resulting formula with (3). Application of Little's formula and $\mathrm{E}\mathbf{V}_{without} = \frac{\sum_{i=1}^{N} \lambda_i \beta_i^{(2)}}{2(1-\rho)}$ then yields the *pseudo-conservation law* [4]:

$$\sum_{i=1}^{N} \rho_i \mathrm{E}\mathbf{W}_i = \rho \frac{\sum_{i=1}^{N} \lambda_i \beta_i^{(2)}}{2(1-\rho)} + \mathrm{E}\mathbf{Y}. \tag{4}$$

Here (cf. the notation introduced in the model description)

$$\mathrm{E}\mathbf{Y} = \rho \frac{s^{(2)}}{2s} + \frac{s}{2(1-\rho)}[\rho^2 - \sum_{i=1}^{N} \rho_i^2] + \sum_{i=1}^{N} \mathrm{E}\mathbf{Z}_{ii}, \tag{5}$$

with $\mathbf{Z}_{ii}$ the amount of work left behind at Q_i by the departing server. $\mathrm{E}\mathbf{Z}_{ii}$, and hence $\mathrm{E}\mathbf{Y}$, can be explicitly determined for polling models with standard service disciplines like 1-limited, gated, or exhaustive. $\mathrm{E}\mathbf{Y} = 0$ for the case of zero switchover times, and then (4) reduces to the well-known *conservation law* [11]. The origin of the term conservation law is that the weighted sum $\sum_{i=1}^{N} \rho_i \mathrm{E}\mathbf{W}_i$ of the mean waiting times remains the same, regardless of any changes in the service disciplines at the various queues. In the case of switchover times this weighted sum *does* change when a service discipline is changed, but

only via a - usually simple - change in $\mathbf{EY}$.
The remarkably simple exact expression for $\sum_{i=1}^{N} \rho_i \mathbf{EW}_i$ has in the past few years turned out to be extremely useful for a variety of purposes: testing simulation results, the development of approximations for mean waiting times, and the optimization of server routing and server visit times.

Queue lengths
For the above-described N-queue cyclic polling model, with exhaustive service at all queues, Eisenberg [7] obtains the joint queue length PGF (Probability Generating Function) at epochs in which B reaches one of the queues. His solution method may also be used to handle the case of gated service at all queues. Furthermore, he also allows a fixed non-cyclic visit pattern. In a series of publications following Eisenberg's paper, an exact queue length analysis has been performed for several other N-queue polling models, with exhaustive or gated service, or mixtures and variants of these service disciplines; for an overview we refer to the survey of Takagi [19]. In contrast, polling models with limits on the number of customers to be served during a session, or on the session time, have mostly defied an exact analysis. The joint queue length distribution for the 2-queue model with 1-limited service at both queues can be obtained by transforming the problem into a Riemann- or Riemann-Hilbert boundary value problem (see, e.g., [6]), but for $N > 2$ it is not clear at all how the queue length problem can be attacked.
In an important paper, written at CWI, Resing [15] clarifies this sharp separation between 'easy' and 'hard' polling models. He considers a class of service disciplines with the following property:

Branching property
If there are k_i customers present at Q_i at the start of a visit, then during the course of the visit each of these k_i customers will effectively be replaced in an i.i.d. manner by a random population having some PGF $h_i(z_1, \ldots, z_N)$ which may be any N-dimensional PGF.

Resing demonstrates that, if the branching property holds at all queues, then the joint queue length process at successive moments that B reaches a fixed queue, say Q_1, is a *Multi-Type Branching Process* (MTBP) 'with immigration'. The theory of MTBP now yields stability conditions as well as an exact expression for the joint queue length PGF.
The 1-limited service discipline does not have the branching property. The gated and exhaustive disciplines, on the other hand, do possess this property, with respectively

$$h_i(z_1, \ldots, z_N) = \beta_i(\sum_{j=1}^{N} \lambda_j(1 - z_j)), \tag{6}$$

(note that this is the PGF of the joint distribution of the numbers of arrivals at the various queues during one service at Q_i), and

$$h_i(z_1,\ldots,z_N) = \theta_i(\sum_{j\neq i}\lambda_j(1-z_j)), \tag{7}$$

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 $B_i(\cdot)$.
In the next section we shall extend the queue length results for the polling model of the present section, with the branching property at all queues, to some more general polling models. Therefore we now go into more detail concerning the theory of MTBP with immigration and the results of Resing [15]. Consider a system with N particle types. Let $p^{(i)}(j_1,\ldots,j_N)$ denote the probability that a type-i particle 'produces' as offspring j_k particles of type k, $k = 1,\ldots,N$. The offspring PGF of $p^{(i)}(j_1,\ldots,j_N)$ is denoted by $f^{(i)}(z_1,\ldots,z_N)$, and the mean number of particles of type j produced by one type-i particle is denoted by m_{ij}. The matrix $M = (m_{ij})$ plays an essential role in the theory of MTBP. M is called primitive if there is an n such that all entries of the matrix M^n are strictly positive. The well-known Perron-Frobenius theorem implies that a nonnegative primitive matrix M has a positive real eigenvalue ν_{max} such that $\mid \nu \mid < \nu_{max}$ for all other eigenvalues ν of M.
Not only are particles produced by other particles; new particles can also enter the system via immigration (this corresponds to the arrival of customers during a period in which B is not serving). Let $q(j_1,\ldots,j_N)$ denote the probability that a group of immigrants consists of j_k particles of type k, $k = 1,\ldots,N$. Denote its PGF by $g(z_1,\ldots,z_N)$, and inductively define the functions $f_n(z_1,\ldots,z_N)$ by

$$f_0(z_1,\ldots,z_N) := (z_1,\ldots,z_N),$$

$$f_n(z_1,\ldots,z_N) := (f^{(1)}(f_{n-1}(z_1,\ldots,z_N)),\ldots,f^{(N)}(f_{n-1}(z_1,\ldots,z_N))).$$

Resing cites the following theorem, due to Quine [14]:

THEOREM 2.1
Let $\mathbf{Z}_n = (\mathbf{Z}_n^{(1)},\ldots,\mathbf{Z}_n^{(N)})$ be an MTBP with immigration in each state, with offspring PGF $f^{(i)}(z_1,\ldots,z_N)$, $i = 1,\ldots,N$, and immigration PGF $g(z_1,\ldots,z_N)$. Let the mean matrix M corresponding to the branching process be primitive and its maximal eigenvalue $\nu_{max} < 1$. Assume the Markov chain $\mathbf{Z}_n$ is irreducible and aperiodic. The stationary distribution $\pi(j_1,\ldots,j_N)$ of the process $\mathbf{Z}_n$ exists iff

$$\sum_{j_1+\ldots+j_N>0} q(j_1,\ldots,j_N)\ \log(j_1+\ldots+j_N) < \infty. \tag{8}$$

When this condition is satisfied, the PGF $P(z_1,\ldots,z_N)$ of the distribution

$\pi(j_1, \ldots, j_N)$ is given by

$$P(z_1, \ldots, z_N) = \prod_{n=0}^{\infty} g(f_n(z_1, \ldots, z_N)). \tag{9}$$

Resing proves the following theorem ([15], Theorem 3):

THEOREM 2.2
Assume that the service discipline at each queue Q_i of the cyclic polling model satisfies the branching property with PGF $h_i(z_1, \ldots, z_N)$, $i = 1, \ldots, N$. Then the numbers of customers in the queues at successive time points that B reaches Q_1 constitute an MTBP with immigration in each state, where the offspring PGF's $f^{(i)}(z_1, \ldots, z_N)$ are given by

$$f^{(i)}(z_1, \ldots, z_N) = h_i(z_1, \ldots, z_i, f^{(i+1)}(z_1, \ldots, z_N), \ldots, f^{(N)}(z_1, \ldots, z_N)), \tag{10}$$

and the immigration PGF $g(z_1, \ldots, z_N)$ is given by

$$g(z_1, \ldots, z_N) = \prod_{i=1}^{N} \sigma_i \left(\sum_{k=1}^{i} \lambda_k (1 - z_k) + \sum_{k=i+1}^{N} \lambda_k (1 - f^{(k)}(z_1, \ldots, z_N)) \right). \tag{11}$$

REMARK 2.2
The proof of Theorem 2.2 is established by considering the evolution of the joint queue length process between two successive time points, say t_n and t_{n+1}, that B reaches Q_1. Let c_A be a customer in the system at t_n. Define the *ancestral line* of c_A as c_A plus the set of all g_1 customers who have arrived during the service of c_A, plus the set of all g_2 customers who have arrived during the service of those g_1 customers, plus ... . Define the *effective replacements* of c_A as those customers from the ancestral line of c_A who are still present at t_{n+1}. If c_A is not served in this cycle, the effective replacements of c_A consist of only c_A itself.
In a similar way the effective replacements of a customer c_B who arrives during a switchover interval between t_n and t_{n+1} are defined.
The total collection of customers in the various queues at t_{n+1} consists of the effective replacements of all customers present at t_n plus the effective replacements of all customers who have arrived during a switchover interval between t_n and t_{n+1}. The fact that all arrival processes are Poisson processes, combined with the fact that all service disciplines satisfy the *branching property*, implies that the joint queue length process at successive epochs when B reaches Q_1 constitutes an MTBP with immigration. The offspring, in the sense of the MTBP, of one type-j customer is the set of effective replacements of that customer, and the immigration corresponds to the effective replacements of all arrivals during the switchover periods in one cycle. In particular, $f^{(N)}(z_1, \ldots, z_N) = h_N(z_1, \ldots, z_N)$, but $f^{(N-1)}(z_1, \ldots, z_N) = h_{N-1}(z_1, \ldots, z_{N-1}, f^{(N)}(z_1, \ldots, z_N))$. The latter formula reflects the fact that

type-N arrivals during a type-$(N-1)$ service may still generate their own offspring during the cycle. To arrive at the nested structure of the last PGF, the following property is used: The PGF of $\mathbf{A}_1 + \ldots + \mathbf{A}_\mathbf{K}$, with $\mathbf{A}_1, \mathbf{A}_2, \ldots$ and $\mathbf{K}$ independent nonnegative integer-valued stochastic variables with PGF $A(\cdot)$ respectively $K(\cdot)$, is given by

$$\sum_{n=0}^{\infty}\sum_{j=0}^{\infty}\Pr(\mathbf{K}=j)\Pr(\mathbf{A}_1+\ldots+\mathbf{A}_j=n)z^n$$

$$=\sum_{j=0}^{\infty}\Pr(\mathbf{K}=j)A(z)^j = K(A(z)).$$

REMARK 2.3
It follows from the above two theorems that the PGF of the joint queue length process at moments that B reaches Q_1 is given by the infinite-product expression (9). Let us explain and illustrate this result by considering the 2-queue case. Denote by $P_i(z_1,z_2)$ ($G_i(z_1,z_2)$) the PGF of the joint queue length distribution when B reaches (leaves) Q_i; so $P_1(z_1,z_2)$ is the PGF we are looking for. Then we have the following four relations.

$$P_1(z_1,z_2) = \sigma_2(\lambda_1(1-z_1)+\lambda_2(1-z_2))G_2(z_1,z_2), \tag{12}$$

$$G_2(z_1,z_2) = P_2(z_1,h_2(z_1,z_2)),$$

$$P_2(z_1,z_2) = \sigma_1(\lambda_1(1-z_1)+\lambda_2(1-z_2))G_1(z_1,z_2),$$

$$G_1(z_1,z_2) = P_1(h_1(z_1,z_2),z_2).$$

Here we have used the memoryless property of the Poisson arrival processes, and the nested structure outlined above for the sum of a random number of stochastic variables, as well as the following property of PGF's:
The PGF of the sum of two independent stochastic variables is the product of their PGF's.
Combination of the four relations in (12) yields:

$$P_1(z_1,z_2) = \sigma_1(z_1,h_2(z_1,z_2))\sigma_2(z_1,z_2)P_1(h_1(z_1,h_2(z_1,z_2)),h_2(z_1,z_2)).$$

Remembering the definitions of the immigration PGF $g(z_1,z_2)$ and the offspring PGF's $f^{(i)}(z_1,z_2)$, we can rewrite this into

$$P_1(z_1,z_2) = g(z_1,z_2)P_1(f_1(z_1,z_2)). \tag{13}$$

Iteration of this functional equation leads to the infinite-product expression (9), with $N=2$.

REMARK 2.4
Polling models *with* and *without* switchover times are usually treated separately

in the literature, often via different approaches; the difficulty with simply letting the switchover times tend to zero in a polling model with switchover times is that the number of visits in an idle period tends to infinity, leading to degenerate distributions at such visit epochs. However, the following way out is possible. Let us assume that B in an empty system rests at, say, Q_1. For this situation Resing [15] shows, for the class of polling models with the branching property, that the joint queue length process at successive moments that B visits Q_1 is again an MTBP - but now with immigration only in state zero. In [2] it is subsequently shown how the identical offspring PGF's of the MTBP's corresponding to the polling model *with* respectively *without* switchover times give rise to a strong relation between their respective joint queue length processes (see also [17]).

3 Polling systems with smart or persistent customers

In this section we shall generalize the polling model of Section 2 in two directions: polling models with arrival rates that depend on the server position ('smart customers') and polling models with feedback and customer routing ('persistent customers'). For each of these directions we outline (because of space restrictions without detailed proofs) how the model can be analyzed completely when the service discipline at each queue satisfies the branching property.

3.1 Smart customers

In some polling applications, knowledge about the server position may influence the arrival rates of the customer types. In the director's example, the knowledge that the director will next turn to personnel matters may generate some new personnel tasks, while there is less hurry in creating tasks of another nature. Let us model this as follows, making a few adaptations in the model described in the previous section. The arrival process of customers at Q_i, when B is at Q_j, is Poisson with rate λ_{ij}; the arrival process of customers at Q_i, when B is switching from Q_j to Q_{j+1}, is Poisson with rate μ_{ij}. When the service discipline at each queue satisfies the branching property, with PGF $h_i(z_1,\ldots,z_N)$ at Q_i, then it is easy to check that the joint queue length process at successive moments that B visits, say, Q_1 is an MTBP with immigration. The immigration PGF is given by (cf. (11)):

$$g(z_1,\ldots,z_N)=\prod_{i=1}^{N}\sigma_i\left(\sum_{k=1}^{i}\mu_{ki}(1-z_k)+\sum_{k=i+1}^{N}\mu_{ki}(1-f^{(k)}(z_1,\ldots,z_N))\right). \tag{14}$$

In the case of gated service at Q_i the offspring PGF is (cf. (6)):

$$h_i(z_1,\ldots,z_N)=\beta_i(\sum_{j=1}^{N}\lambda_{ji}(1-z_j)), \tag{15}$$

and in the case of exhaustive service at Q_i the offspring PGF is (cf. (7)):

$$h_i(z_1, \ldots, z_N) = \theta_i(\sum_{j \neq i} \lambda_{ji}(1 - z_j)). \tag{16}$$

The reasoning presented in Remark 2.3 should make it clear that the present model again gives rise to a functional equation of the type (13), iteration of which leads to an infinite-product expression for $P_1(z_1, \ldots, z_N)$ like (9). The PGF of the joint queue length distribution at the end of a switchover from Q_i to Q_{i+1} is simply expressed in the PGF at the beginning of that switchover (the end of a visit to Q_i), and the latter PGF can be expressed in the PGF of the joint queue length distribution at the beginning of that visit to Q_i by substitution of the offspring PGF $h_i(.)$ at the i-th position in the PGF.

Several interesting special cases deserve further attention. E.g., $\lambda_{ij} = \Lambda p_{ij}$ and $\mu_{ij} = \Lambda q_{ij}$ with $p_{ij}, q_{ij} \geq 0$ and $\sum_{i=1}^{N} p_{ij} = \sum_{i=1}^{N} q_{ij} = 1$ for all j corresponds to a fixed total arrival rate Λ. If the service discipline at each queue is gated (hence when B visits Q_i, he will only serve customers that were already present at the start of the session), the smartest thing for an arriving customer to do is to go to the *next* queue: $\lambda_{i+1,i} = \mu_{i+1,i} = \Lambda$, and $\lambda_{ij} = \mu_{ij} = 0$ for all $i \neq j+1$. The most foolish behaviour, on the other hand, is represented by $\lambda_{i,i} = \mu_{i,i} = \Lambda$, and $\lambda_{ij} = \mu_{ij} = 0$ for all $j \neq i$. The former choice clearly minimizes the waiting time of each individual arriving customer. Let us now moreover assume that $B_i(\cdot) \equiv B(\cdot)$. Then the above choice also minimizes, in the sense of stochastic ordering, the workload of the server. This may be proven using coupling methods; see [5] for the more restricted fully symmetric case.
In the case of identical service time distributions and fixed total arrival rate Λ, the work decomposition (1) still holds (check the level crossing argument presented in Remark 2.1), and $\mathbf{EY}$ can easily be calculated. But if not all service time distributions are the same, or the total arrival rate is not constant, then the whole work decomposition concept breaks down. Some reflection will make it clear that when switchover times are zero, even the concept of work *conservation* is destroyed.

3.2 Feedback and customer routing

In the director's example, a completed task may have to be reconsidered in future sessions. This feature can be incorporated in the model of Section 2 in the following way. A newly arriving customer at Q_i (Poisson with arrival rate λ_i) is called a type-$(i,1)$ customer. After completion of its service, it moves to Q_k with probability $p_{ik}^{(1)}$, becoming a type-$(k,2)$ customer, and it leaves the system with probability $p_{i0}^{(1)}$. More generally, a type-(i,j) customer denotes a customer at Q_i who has to be served for the j-th time; after having received service, it moves to Q_k with probability $p_{ik}^{(j)}$, and it leaves the system with

probability $p_{i0}^{(j)}$. A type-(i,j) customer requires a service time at Q_i with distribution $B_{ij}(\cdot)$, with LST $\beta_{ij}(\cdot)$. We assume that $p_{i0}^{(L)} = 1$ for all i, i.e., each customer requires at most L services.
Customer routing has hardly been studied in the context of polling, although several applications in token ring networks, robotics and computer systems exist; cf. Sidi et al. [16]. The latter paper analyzes the case of fixed transition probabilities p_{ij} of customers from Q_i to Q_j, with fixed service time distribution $B_i(\cdot)$ at Q_i and exhaustive or gated service at all queues.
In this section we study the NL-dimensional queue length process $\mathbf{X} = (\mathbf{X}_{11}, \ldots, \mathbf{X}_{1L}; \ldots; \mathbf{X}_{N1}, \ldots, \mathbf{X}_{NL})$, where $\mathbf{X}_{ij}$ denotes the number of customers of type-(i,j) at a moment at which B reaches Q_1.
The *branching property* of Section 2 has to be adapted in the sense that one has to distinguish L PGF's $h_{ij}(z_1, \ldots, z_N)$, $j = 1, \ldots, L$, in Q_i.
It is easily seen that $\mathbf{X}$ is an MTBP with immigration in each state. For the general case, determination of the offspring PGF's and the immigration PGF is somewhat involved. For example, one has to take the possibility into account that a customer is fed back to the same queue; and in the case of exhaustive service, such a customer may then receive more than one service during the same session. We shall refrain from formulating and proving the generalization of Theorem 2.2 here in its full generality. Instead, we illustrate the structure of the MTBP by considering a two-queue example with gated service at both queues. Similar to Remark 2.3, we denote by $P_i(z_{11}, \ldots, z_{2L})$ ($G_i(z_{11}, \ldots, z_{2L})$) the PGF of the joint queue length distribution when B reaches (leaves) Q_i. We have the following four relations:

$$P_1(z_{11}, \ldots, z_{2L}) = \sigma_2(\lambda_1(1 - z_{11}) + \lambda_2(1 - z_{21}))G_2(z_{11}, \ldots, z_{2L}), \tag{17}$$

$$G_2(z_{11}, \ldots, z_{2L}) = P_2(z_{11}, \ldots, z_{1L}; y_{21}, \ldots, y_{2L}),$$

$$P_2(z_{11}, \ldots, z_{2L}) = \sigma_1(\lambda_1(1 - z_{11}) + \lambda_2(1 - z_{21}))G_1(z_{11}, \ldots, z_{2L}),$$

$$G_1(z_{11}, \ldots, z_{2L}) = P_1(y_{11}, \ldots, y_{1L}; z_{21}, \ldots, z_{2L}).$$

Here, for $i = 1, 2$, $j = 1, \ldots, L$,

$$y_{ij} := \beta_{ij}(\lambda_1(1 - z_{11}) + \lambda_2(1 - z_{21}))\ [p_{i0}^{(j)} + p_{i1}^{(j)} z_{1,j+1} + p_{i2}^{(j)} z_{2,j+1}].$$

Note that $\beta_{ij}(\lambda_1(1-z_{11})+\lambda_2(1-z_{21}))$ is the PGF of the numbers of new arrivals at the various queues during a type-(i,j) service, and that $p_{i0}^{(j)} + p_{i1}^{(j)} z_{1,j+1} + p_{i2}^{(j)} z_{2,j+1}$ is the PGF of the numbers of type-$(k, j+1)$ customers, $k = 1, 2$, generated by the feedback of one type-(i,j) customer.
Combination of the four relations in (17) leads to a recursion for $P_1(z_{11}, \ldots, z_{2L})$, of similar form as (13), which can be solved iteratively.

REMARK 3.1
We thus obtain the PGF of the joint queue length distribution at time points in which B reaches Q_1. But the four relations in (17) then also yield the PGF's

of the joint queue length distributions at time points in which B leaves Q_1, reaches Q_2 and leaves Q_2. The PGF of the joint steady-state queue length distribution may also be determined from these results, once the service order at the queues is specified (e.g., serve type-$(i, j+1)$ before type-(i, j) customers).

REMARK 3.2
The case of a *single* queue with feedback, contained in the present model, is also interesting in itself. We can obtain the joint queue length distribution of the numbers of customers that are present for the first,..., L-th time, at the time points at which B starts a new session.

REMARK 3.3
Several variants and generalizations can also be handled in the framework of an MTBP. For example, one can allow zero switchover times between sessions, obtaining an MTBP with immigration only in state zero. Furthermore, instead of assuming $p_{i0}^{(L)} = 1$, we may also assume that $p_{ik}^{(j)} = p_{ik}$ and $B_{ij}(\cdot) \equiv B_i(\cdot)$ for all $j \geq L$, $k = 0, 1, \ldots, L$. The resulting MTBP still has a finite number of NL variables. This generalizes the model of Sidi et al. [16] in various ways.
We may generalize our model even further, while retaining the MTBP structure. For example, we can allow 'smart customers' *in combination with* feedback and routing; and we can also allow the possibility that a served customer not just feeds back, but branches into several customers: a task of type-(i, j) that has been handled by the director may simultaneously give rise to tasks $(k_1, j+1)$ and $(k_2, j+1)$. While these possibilities may make the job of a director rather complicated, they do not fundamentally complicate the analysis of his workload.

ACKNOWLEDGMENT
The author gratefully acknowledges several useful discussions with Sem Borst and Jacques Resing.

REFERENCES
1. S.C. Borst (1994). *Polling Systems.* PhD Thesis, Tilburg University.
2. S.C. Borst, O.J. Boxma (1994). Polling models with and without switchover times. *Report CWI, BS-R9421.*
3. O.J. Boxma (1989). Workloads and waiting times in single-server systems with multiple customer classes. *Queueing Systems* 5, 185-214.
4. O.J. Boxma, W.P. Groenendijk (1987). Pseudo-conservation laws in cyclic service systems. *J. Appl. Probab.* 24, 949-964.
5. O.J. Boxma, M. Kelbert (1994). Stochastic bounds for a polling system. *Annals of Oper. Res.* 48, 295-310.
6. J.W. Cohen, O.J. Boxma (1983). *Boundary Value Problems in Queueing System Analysis* (North-Holland Publ. Cy., Amsterdam).

7. M. Eisenberg (1972). Queues with periodic service and changeover times. *Oper. Res.* 20, 440-451.
8. C. Fricker, R. Jaïbi (1994). Monotonicity and stability of periodic polling models. *Queueing Systems* 15, 211-238.
9. D. Grillo (1990). Polling mechanism models in communication systems - some application examples. In: *Stochastic Analysis of Computer and Communication Systems*, ed. H. Takagi (North-Holland, Amsterdam), 639-658.
10. W.P. Groenendijk (1990). *Conservation Laws in Polling Systems.* PhD Thesis, University of Utrecht.
11. L. Kleinrock (1964). *Communication Nets - Stochastic Message Flow and Delay* (Dover, New York).
12. C. Mack, T. Murphy, N.L. Webb (1957). The efficiency of N machines unidirectionally patrolled by one operative when walking times and repair times are constants. *J. Roy. Statist. Soc. B* 19, 166-172.
13. D.P. Heyman, M.J. Sobel (1982). *Stochastic Models in Operations Research*, Vol. I (McGraw-Hill Book Company, New York).
14. M.P. Quine (1970). The multitype Galton-Watson process with immigration. *J. Appl. Probab.* 7, 411-422.
15. J.A.C. Resing (1993). Polling systems and multitype branching processes. *Queueing Systems* 13, 409-426.
16. M. Sidi, H. Levy, S.W. Fuhrmann (1992). A queueing network with a single cyclically roving server. *Queueing Systems* 11, 121-144.
17. M.M. Srinivasan, S.-C. Niu, R.B. Cooper (1993). Relating polling models with nonzero and zero switchover times. Report Univ. of Tennessee; to appear in *Queueing Systems.*
18. H. Takagi (1991). Application of polling models to computer networks. *Comp. Netw. ISDN Syst.* 22, 193-211.
19. H. Takagi (1990). Queueing analysis of polling models: An update. In: *Stochastic Analysis of Computer and Communication Systems*, ed. H. Takagi (North-Holland, Amsterdam), 267-318.