Polling: Past, Present and Perspective

S.C. Borst*† s.c.borst@tue.nl
O.J. Boxma* o.j.boxma@tue.nl

September 11, 2018

Abstract

This is a survey on polling systems, focussing on the basic single-server multi-queue polling system in which the server visits the queues in cyclic order. The main goals of the paper are: (i) to discuss a number of the key methodologies in analyzing polling models; (ii) to give an overview of recent polling developments; and (iii) to present a number of challenging open problems.
Note: Invited paper, to appear in **TOP**.

1 Introduction

This paper is devoted to polling systems. The basic polling system is a queueing model in which customers arrive at n queues according to independent Poisson processes, and in which a single server visits those n queues in cyclic order to serve the customers. When $n = 1$, this system reduces to the classical $M/G/1$ queue. For general n, the basic polling system may be viewed as an $M/G/1$ queue with n customer classes and dynamically changing priority – in contrast to queueing models with multiple customer classes which have fixed priority levels. In many applications, the switchover times of the server, when moving from one queue to another, are nonnegligible and should be included in the model.

Applications of polling systems abound, because a service facility that can serve the needs of n different types of customers is such a natural setting in every-day life. Indeed, polling systems have been used to model a plethora of congestion situations, like (i) a patrolling repairman with n types of repair jobs, (ii) a machine producing n types of products on demand, (iii) protocols in computer-communication systems, allocating resources to n stations, job types or traffic sources, and (iv) a signalized road traffic intersection with n different traffic streams. These and other application areas have given rise to a huge range of variants and extensions of the basic polling system. Several overviews of the applicability of polling systems have been published, cf. Grillo [102], Levy and Sidi [129], Takagi [152] and Boon, Van der Mei and Winands [24]. We therefore refrain from an extensive discussion of polling applications. When it comes to polling surveys, one should of course mention that, until 2000, Takagi maintained a quite complete bibliography on polling models, which included more than 700 publications [154, 155]. A more recent survey is Vishnevskii and Semenova [159].

The main goals of the present paper are threefold. Firstly, to discuss a number of the key methodologies in analyzing polling models. Secondly, to give an overview of recent polling developments. Finally, to present a number of challenging open problems, which hopefully promote the interest of the reader in this fascinating field.

*Department of Mathematics and Computer Science, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands
†Nokia Bell Labs, Murray Hill, NJ, USA.

As a disclaimer, we would like to emphasize that we do not aim for completeness. Since the publication of the survey [155], several hundreds of polling papers have appeared. When discussing recent developments, we mainly focus on contributions which we believe to be methodologically important or which give rise to interesting open problems – and undoubtedly there is a bias towards publications which are in some way related to the authors.

Polling models are closely related to queueing models with vacations. One could naively model one queue of a polling model as a queue in isolation, in which the intervisit time (composed of switchover times and visit times at the other queues, i.e., the time periods the server spends at a queue) is viewed as a server vacation. Unfortunately, the intervisit times depend on the visit times in an intricate way. In this paper we do not give much consideration to queues with vacations; we refer the reader to the surveys of Doshi [71, 72] and the books of Takagi [153] and Tian and Zhang [156].

The remainder of the paper is organized in the following way. Section 2 presents a detailed model description. Section 3 reviews some properties and results of very general validity, including the so-called pseudoconservation law. Section 4 focuses on waiting times and (mainly) joint queue length distributions, for the important class of disciplines which satisfy a so-called branching property. Section 5 is devoted to polling models which do not satisfy that property.

The next few sections consider some special topics: polling models with arrival processes that generalize the above-mentioned Poisson processes (Section 6), scheduling in polling models (Section 7) and two types of asymptotics: many-queue asymptotics and heavy-traffic asymptotics. Section 9 contains a collection of interesting isolated polling models and results. Finally, Section 10 presents some suggestions for further research.

2 Model description

We are interested in situations in which a service facility offers services to n classes of customers, in some prescribed order. We present the model description via 10 assumptions. Some of these assumptions will be relaxed in later sections.

Assumption 1. The service facility has a single server; and that server works at unit speed when it is working.

Assumption 2. The number of customer classes, n, is finite.

Assumption 3. Customers in the various classes arrive at the service facility according to n independent Poisson arrival processes, with intensity λ_i for class i, joining a queue Q_i, $i = 1, 2, \ldots, n$. Customers of class i have service requirements which are independent, identically distributed (i.i.d.) random variables, generically denoted by $\mathbf{B}_i$, with distribution $B_i(\cdot)$ and Laplace-Stieltjes transform (LST) $\beta_i(\cdot)$, $i = 1, 2, \ldots, n$. Service requirements of customers of different classes are also independent of each other, and of the arrival processes.

Assumption 4. Each queue has an infinite buffer capacity. Furthermore, all customers have infinite patience; hence no customer is lost.

Assumption 5. The *routing policy* of the server is cyclic: the server successively visits the queues in order $Q_1, Q_2, \ldots, Q_n, Q_1, Q_2, \ldots, Q_n$ etc. Another option that we will briefly touch upon is a polling table, i.e., a fixed visit pattern which is cyclically repeated (like star polling with Q_1 as center of the star: $Q_1, Q_2, Q_1, Q_3, \ldots, Q_1, Q_n$). Yet another option is random polling, in which the server visits the queues according to a probabilistic visit scheme. Markovian polling refers to the case in which the transitions between queues follow a Markov chain.

Assumption 6. The *service policy*, describing the behavior of the server while visiting a queue, can be one of many policies which have been considered in the literature. The most popular ones are the following: (i) *exhaustive*: the server keeps serving a queue until it has become empty; (ii) *gated*: the server keeps serving a queue until all those customers have been served that were already present when the server arrived at that queue; (iii) *k-limited*: the server keeps

working at a queue until a predefined number of k customers has been served, or the queue has become empty – whichever occurs first. Other policies include *decrementing* service: the server serves a queue until the number in that queue has decreased to one less than the number present upon arrival of the server; *time-limited* service: the server serves customers at Q_i until a time limit T_i has been reached, or until the queue has become empty – whichever occurs first; and binomial-gated: the server restricts service to the customers present upon its arrival, but each of those is only served with a fixed probability p_i (in Q_i, $i = 1, 2, \ldots, n$). Another well-studied policy is *Globally gated*: when the server arrives at Q_1 at some time t_1, it starts a cycle of the n queues in which it only serves the customers that are already present at t_1.

Finally, we assume that a server does not stay at an empty queue if other queues are not empty (non-idling assumption); however, in Subsection 9.2 we briefly consider an idling service policy.

Assumption 7. The *service order* within each queue is First-Come First-Served (FCFS). This assumption was almost universally made in the polling literature until the work of Wierman *et al.* [163]. In Section 7 we will discuss non-FCFS service orders.

Assumption 8. The times to switch from Q_i, $i = 1, 2, \ldots, n$, to the next queue are assumed to be i.i.d. random variables, generically denoted by $\mathbf{S}_i$, with distribution $S_i(\cdot)$ and LST $\sigma_i(\cdot)$. All switchover times are assumed to be independent of each other and of the interarrival and service times. When the switchover times between successive queues are all zero, a special situation arises. If the system has become empty after a visit to, say, Q_i in the case of zero switchover times, then the server is assumed to visit queues $Q_{i+1}, \ldots, Q_n$ (which now takes zero time) and stay in front of Q_1 (see Section 4). In the case of non-zero switchover times, the server is assumed to keep switching in an empty system.

Assumption 9. As soon as a customer has been served, it leaves the system. At some places we briefly mention the case of customer routing; a served customer might rejoin the same queue, or join another one.

Assumption 10. The total traffic load is such that the key stochastic processes (queue lengths, waiting times) reach steady state. A necessary condition for this is that the total offered load $\rho := \sum_{i=1}^{n} \rho_i < 1$; here $\rho_i := \lambda_i \mathrm{E}\mathbf{B}_i$ is the mean offered load at Q_i per time unit, $i = 1, 2, \ldots, n$. When all switchover times are zero, this condition is also sufficient. Otherwise the situation may be much more complicated, and in particular the service policies may influence the stability condition; e.g., in 1-limited service, the server is forced to spend a switchover time after each service. See Fricker and Jaïbi [93] for an extensive discussion of these stability issues. We refer to Foss and Chernova [86, 87], Foss, Chernova and Kovalevskii [88], Foss and Last [90, 91], and Kovalevskii, Topchii and Foss [119] for stability results for various polling systems (not necessarily satisfying all of the above assumptions), along with related dominance theorems and fluid limits.

When a polling system satisfies all 10 assumptions, we denote it by PS.

3 General results

In this section we discuss a number of results which hold for basically all PS, i.e., polling systems that satisfy Assumptions 1–10 of Section 2. These are cycle-time and visit-time results (Subsection 3.1), workload decompositions (Subsection 3.2), pseudo-conservation laws for mean waiting times (Subsection 3.3), Eisenberg's relations between queue lengths at visit beginnings, visit completions, service beginnings and service completions (Subsection 3.4) and a general relation between the joint queue length distribution at an arbitrary epoch and the joint queue length distributions at visit beginnings and visit completions (Subsection 3.5).

3.1 Mean cycle and visit times

In a polling model of type PS, let us define the cycle time $\mathbf{C}_i$ of Q_i as the time between two successive visit beginnings of the server to Q_i. If the mean total switchover time in a polling model of type PS is positive, i.e., $s := \sum_{i=1}^{n} \mathrm{E}\mathbf{S}_i > 0$, then the mean cycle time for Q_i satisfies the following balance equation:

$$\mathrm{E}\mathbf{C}_i - s = \rho \mathrm{E}\mathbf{C}_i, \quad i = 1, 2, \ldots, n.$$

Indeed the lefthand side gives the mean length of time the server is working during an arbitrary cycle of Q_i, and the righthand side gives the mean amount of work arriving in PS during an arbitrary cycle $\mathbf{C}_i$. In steady state these two quantities should be equal. Hence we find:

$$\mathrm{E}\mathbf{C}_i = \frac{s}{1-\rho}, \quad i = 1, 2, \ldots, n. \tag{1}$$

Apparently each queue has the same mean cycle time $\mathrm{E}\mathbf{C}$. It is important to notice, though, that the *distributions* of the cycle times of different queues, and even the variances, may *not* the same (unless the system is completely symmetric).

The balance argument used above also immediately implies that the mean visit time $\mathrm{E}\mathbf{V}_i$ of Q_i is given by

$$\mathrm{E}\mathbf{V}_i = \rho_i \mathrm{E}\mathbf{C}_i = \frac{\rho_i s}{1-\rho}, \quad i = 1, 2, \ldots, n. \tag{2}$$

In a system with zero switchover times, viz., $s = 0$, Formulas (1) and (2) still hold if the server is assumed to keep cycling when the system has become empty (indeed, in an empty system there will be an infinite number of zero-length cycles); however, these formulas are meaningless then.

3.2 Workload decompositions

Again consider the polling system PS, and assume in addition that all switchover times are zero. The server is then always working as long as there are customers in the system (cf. Assumption 6). Since the server is working at unit speed when it is working (Assumption 1), a sample path consideration reveals that the amount of work in the system evolves in a way that does not depend on the order of service of the queues, or within the queues, and neither on the service policies at the queues. This is the *principle of work conservation* (cf. Heyman and Sobel [105], p. 418). In particular, the amount of work evolves exactly as in an $M/G/1$ queue in which the arrival rate is $\Lambda := \sum_{i=1}^{n} \lambda_i$ and in which the service time distribution is $\sum_{i=1}^{n} \frac{\lambda_i}{\Lambda} B_i(\cdot)$. We denote this queue by the 'corresponding $M/G/1$ queue'.

If the switchover times are positive, then the principle of work conservation is violated: the server is sometimes switching (not working) although there is work present in the system. It was proven in [37] that, for a cyclic polling system PS, a *principle of work decomposition* holds: the steady-state amount of work $\mathbf{V}_{with}$ in PS with switchover times is, in distribution, the sum of the steady-state amount of work $\mathbf{V}_{without}$ in the corresponding PS without switchover times (hence the corresponding $M/G/1$ queue) plus the steady-state amount of work $\mathbf{Y}$ present in the system at an epoch in which the server is not working:

$$\mathbf{V}_{with} \stackrel{d}{=} \mathbf{V}_{without} + \mathbf{Y}, \tag{3}$$

and $\mathbf{V}_{without}$ and $\mathbf{Y}$ are independent. This decomposition result was generalized in [33] to a large class of single-server queues with multiple customer classes and various forms of work interruptions. These decompositions fit in a line of decomposition results for queueing models with vacations/interruptions which goes back to the ground-breaking paper of Fuhrmann and

Cooper [96] which concentrates on *queue length* decompositions. It should be noticed that queue lengths are much more sensitive to distributional assumptions than workload, and hence the conditions for queue length decompositions to hold are also more stringent than those for workload decompositions. Most of the decomposition proofs rely on sample path considerations, and on the fact that the workload evolves exactly the same under FCFS and Last-Come First-Served (LCFS), and on the exploitation of nice properties of the LCFS Preemptive-Resume discipline. See also the insightful discussion in Ivanovs and Kella [110], and a workload decomposition for polling models with multi-dimensional Lévy input in Boxma and Kella [40].

3.3 Pseudoconservation laws

For the PS model one can express the mean workload $\mathrm{E}\mathbf{V}_{with}$ into the mean numbers $\mathrm{E}\mathbf{N}_i$ of waiting customers at the various queues of PS, and hence, via Little's formula, into the mean waiting times $\mathrm{E}\mathbf{W}_i$. This is sometimes referred to as Brumelle's formula [53]:

$$\mathrm{E}\mathbf{V}_{with} = \sum_{i=1}^{n} \mathrm{E}\mathbf{B}_i \mathrm{E}\mathbf{N}_i + \sum_{i=1}^{n} \rho_i \frac{\mathrm{E}\mathbf{B}_i^2}{2\mathrm{E}\mathbf{B}_i} = \sum_{i=1}^{n} \rho_i \mathrm{E}\mathbf{W}_i + \frac{1}{2}\sum_{i=1}^{n} \lambda_i \mathrm{E}\mathbf{B}_i^2. \tag{4}$$

Indeed, $\mathrm{E}\mathbf{B}_i\mathrm{E}\mathbf{N}_i$ is the mean amount of work of waiting customers at Q_i (we use here the fact that service at each queue is non-preemptive; hence we have to exclude a discipline like time-limited), and $\rho_i \frac{\mathrm{E}\mathbf{B}_i^2}{2\mathrm{E}\mathbf{B}_i}$ is the product of the probability that Q_i is being served at an arbitrary epoch, and the mean length of the residual service time of a customer at Q_i.

Using (3) and the fact that, in the case of zero switchover times, one has (using a well-known result for the 'corresponding $M/G/1$ queue'):

$$\mathrm{E}\mathbf{V}_{without} = \sum_{i=1}^{n} \frac{\lambda_i \mathrm{E}\mathbf{B}_i^2}{2(1-\rho)}, \tag{5}$$

the following so-called *pseudo-conservation law* (PCL) for the mean waiting times is obtained [37]:

$$\sum_{i=1}^{n} \rho_i \mathrm{E}\mathbf{W}_i = \rho \sum_{i=1}^{n} \frac{\lambda_i \mathrm{E}\mathbf{B}_i^2}{2(1-\rho)} + \mathrm{E}\mathbf{Y}. \tag{6}$$

In [37], $\mathrm{E}\mathbf{Y}$ is subsequently split in three terms:

$$\mathrm{E}\mathbf{Y} = \rho \frac{s^{(2)}}{2s} + \frac{s}{2(1-\rho)}[\rho^2 - \sum_{i=1}^{n} \rho_i^2] + \sum_{i=1}^{n} \mathrm{E}\mathbf{Z}_{ii}, \tag{7}$$

where s and $s^{(2)}$ are the mean and second moment of the total switchover time in one cycle of the server. The three terms reflect the influence of the presence of switchover times. All three terms have an easy probabilistic interpretation. Focussing on the contributions from Q_i, one has $\mathrm{E}\mathbf{Z}_{ii}$ in the last term, which denotes the mean amount of work left behind by the server in Q_i after a visit to that queue. In the first term one has a contribution $\rho_i \frac{s^{(2)}}{2s}$, which is the mean amount of work which has arrived in Q_i (after the server visit to Q_i) during the past part of the total switchover time in a cycle. Finally, the contribution of Q_i to the second term of (7), $\rho_i \sum_{j=i+1}^{n} \frac{\rho_j s}{1-\rho}$, is the mean total workload which has arrived in Q_i during the visit times at $Q_{i+1}, \ldots, Q_n$ of the server (cf. (2)).

The term $\mathrm{E}\mathbf{Z}_{ii}$ is the only term that depends on the service policy at the queues (and in fact only on the service policy at that particular queue). For many service policies, it is easy to determine $\mathrm{E}\mathbf{Z}_{ii}$. For exhaustive service, it equals zero, and for gated service $\mathrm{E}\mathbf{Z}_{ii} = \rho_i^2 \frac{s}{1-\rho}$; indeed, $\rho_i \mathrm{E}\mathbf{V}_i$ arrives on average at Q_i per visit, and $\mathrm{E}\mathbf{V}_i = \frac{\rho_i s}{1-\rho}$ according to (2).

The PCL has been generalized in several directions, including batch Poisson arrivals, polling tables and Markovian polling. The simplicity, quite general validity and robustness of the PCL make it suitable for several purposes. These include the development of approximations for mean waiting times and/or a check of such approximations and optimizations as will be discussed in Subsections 9.2 and 9.3.

3.4 Eisenberg's relation

In this subsection, following [29], we discuss a beautiful relation of Eisenberg [73], which in our view would have deserved greater attention in the polling literature. Eisenberg relates the probability generating functions of queue lengths at various instants: visit beginnings and endings, and service beginnings and endings. Eisenberg [73] studies a polling model with non-zero switchover times and 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 $Q_1, \ldots, Q_n$ and N a realization:

$\mathbf{S}_{b_i}(t, N)$:= number of service beginnings at Q_i in $(0, t)$ for which $\mathbf{N} = N$;
$\mathbf{S}_{c_i}(t, N)$:= number of service completions at Q_i in $(0, t)$ for which $\mathbf{N} = N$;
$\mathbf{V}_{b_i}(t, N)$:= number of visit beginnings at Q_i in $(0, t)$ for which $\mathbf{N} = N$;
$\mathbf{V}_{c_i}(t, N)$:= number of visit completions at Q_i in $(0, t)$ for which $\mathbf{N} = N$.

In the case of a service or visit completion, the state is defined as what exists immediately after the departure of the customer.
Eisenberg [73] 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,

$$\mathbf{V}_{b_i}(t, N) + \mathbf{S}_{c_i}(t, N) = \mathbf{S}_{b_i}(t, N) + \mathbf{V}_{c_i}(t, N). \tag{8}$$

As observed in [29], (8) not only holds for the case of non-zero 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:

$\tilde{S}_{b_i}(N) := \Pr(\mathbf{N} = N,\ S \text{ is at } Q_i \mid \text{service beginning instant})$;
$\tilde{S}_{c_i}(N) := \Pr(\mathbf{N} = N,\ S \text{ is at } Q_i \mid \text{service completion instant})$;
$\tilde{V}_{b_i}(N) := \Pr(\mathbf{N} = N \mid \text{visit beginning at } Q_i)$;
$\tilde{V}_{c_i}(N) := \Pr(\mathbf{N} = N \mid \text{visit completion at } Q_i)$.

Eisenberg [73] divides all four terms in (8) 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:

$$\gamma_i \tilde{V}_{b_i}(N) + \tilde{S}_{c_i}(N) = \tilde{S}_{b_i}(N) + \gamma_i \tilde{V}_{c_i}(N).$$

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 $\gamma_i \equiv \gamma$, $i = 1, \ldots, n$. Written in terms of PGF's (probability generating functions),

$$\gamma V_{b_i}(z) + S_{c_i}(z) = S_{b_i}(z) + \gamma V_{c_i}(z), \tag{9}$$

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

$$S_{c_i}(z) = S_{b_i}(z)\beta_i(\sum_{j=1}^{n} \lambda_j(1 - z_j))/z_i, \tag{10}$$

for $| z_j | \leq 1$, $j = 1, \ldots, n$.
It follows from (9) and (10) that

$$S_{c_i}(z) = \frac{\gamma\beta_i(\sum_{j=1}^{n} \lambda_j(1 - z_j))}{z_i - \beta_i(\sum_{j=1}^{n} \lambda_j(1 - z_j))}[V_{b_i}(z) - V_{c_i}(z)]. \tag{11}$$

Eisenberg, considering the variant with switchover times and exhaustive service, subsequently expresses $V_{b_i}(z)$ into $V_{c_{i-1}}(z)$. For the moment we refrain from that (see (15) below), but we observe that Formula (11) is generally valid for the polling systems PS described in Section 2 (with and without switchover times).

Taking $z = (1, \ldots, 1, y, 1, \ldots, 1)$ in (11), 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 PGF at Q_i at a service completion instant at Q_i. PASTA, in combination with a standard up- and down-crossing argument, 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 (or, equivalently, at the end and beginning of an intervisit time of Q_i), one obtains after some rewriting (see [29] for the details):

$$\mathrm{E}(y^{\mathbf{N}_i}) = \frac{(1 - \rho_i)(1 - y)\beta_i(\lambda_i(1 - y))}{\beta_i(\lambda_i(1 - y)) - y} \frac{\mathrm{E}(y^{\mathbf{Y}_i}) - \mathrm{E}(y^{\mathbf{X}_i})}{(1 - y)(\mathrm{E}\mathbf{X}_i - \mathrm{E}\mathbf{Y}_i)}, \quad |y| \leq 1. \tag{12}$$

The first term in the right-hand side is the PGF $\mathrm{E}(y^{\mathbf{N}_{i|M/G/1}})$ of the queue length distribution in a 'corresponding' isolated $M/G/1$ queue of Q_i with arrival rate λ_i and service time LST $\beta_i(\cdot)$. The second term appears to be the PGF of the number of customers $\mathbf{N}_{i|I}$ at an arbitrary intervisit time of Q_i. Formula (12) implies that

$$\mathbf{N}_i \stackrel{d}{=} \mathbf{N}_{i|M/G/1} + \mathbf{N}_{i|I}, \tag{13}$$

the two terms in the righthand side being independent. This is the well-known Fuhrmann-Cooper queue length decomposition [96].

Remark 3.1
Fuhrmann and Cooper [96] state five conditions under which their decomposition holds; these conditions are contained in the 10 *assumptions of Section 2, except that it is explicitly assumed in [96] that service is non-preemptive, a condition that is violated when the service discipline is time-limited for example.*

Using the distributional form of Little's law, cf. Keilson and Servi [114], the above Fuhrmann-Cooper queue length decomposition (13) immediately translates into a waiting-time decomposition. In Section 4.1 we will return to this relation, for the case of polling models that satisfy Property 4.1. □

3.5 The joint queue length distribution at an arbitrary epoch

In Subsection 3.4 we focused on queue length vectors at visit beginnings and visit completions, and at service beginnings and service completions. Throughout the polling literature, the attention has always been on those epochs, as far as joint queue length distributions is concerned. However, in [41] it was shown that, for the general PS model, one can express the PGF $L(z)$ of

the steady-state joint queue length distribution *at an arbitrary epoch* in those at visit beginnings and visit completions, in the following way (with $z = (z_1, \ldots, z_n)$):

$$L(z) = \frac{1}{\mathrm{E}\mathbf{C}} \sum_{i=1}^{n} \left(\frac{V_{b_i}(z) - V_{c_i}(z)}{\Sigma(z)} \frac{z_i \left(1 - \beta_i(\Sigma(z))\right)}{z_i - \beta_i(\Sigma(z))} + \frac{V_{c_i}(z) - V_{b_{i+1}}(z)}{\Sigma(z)} \right), \tag{14}$$

with $\Sigma(z) := \sum_{j=1}^{n} \lambda_j (1 - z_j)$. Its proof in [41] is based on the following relations:
(i) Eisenberg's [73] relation (11) as generalized to PS polling models in [29].
(ii) Relation (10) between queue length PGF's at the beginning and end of a service.
(iii) an obvious relation between queue lengths at the beginning and end of a switchover:

$$V_{b_{i+1}}(z) = V_{c_i}(z) \sigma_i \left(\Sigma(z)\right), \quad i = 1, 2, \ldots, n. \tag{15}$$

(iv) a stochastic mean value theorem, expressing $L(z)$ as an average over the PGF's of the joint queue length distribution at an arbitrary moment during a visit to Q_i ($X_i(z)$) and during a switchover period between Q_i and Q_{i+1} ($Y_i(z)$):

$$L(z) = \frac{1}{\mathrm{E}\mathbf{C}} \sum_{i=1}^{n} \left(\frac{\mathrm{E}\mathbf{B}_i}{\gamma_i} X_i(z) + s_i Y_i(z) \right), \tag{16}$$

where, for $i = 1, 2, \ldots, n$,

$$X_i(z) = S_{b_i}(z) \beta_i^{\mathrm{past}}(\Sigma(z)), \tag{17}$$

$$Y_i(z) = V_{c_i}(z) \sigma_i^{\mathrm{past}}(\Sigma(z)), \tag{18}$$

where $\beta_i^{\mathrm{past}}(\cdot)$ and $\sigma_i^{\mathrm{past}}(\cdot)$ are the LST's of the past parts of $\mathbf{B}_i$ and $\mathbf{S}_i$, respectively, and therefore

$$\beta_i^{\mathrm{past}}(\Sigma(z)) = \frac{1 - \beta_i(\Sigma(z))}{\mathrm{E}\mathbf{B}_i \Sigma(z)}, \qquad \sigma_i^{\mathrm{past}}(\Sigma(z)) = \frac{1 - \sigma_i(\Sigma(z))}{\mathrm{E}\mathbf{S}_i \Sigma(z)}. \tag{19}$$

Starting from (16), substituting (17) and (18), and using (10) and (11) to eliminate all $S_{b_i}(z)$ and $S_{c_i}(z)$, yields (14).

Remark 3.2
In [41] also zero switchover times are allowed; the same result (14) is shown to hold.

In Theorem 1 of [41] it was subsequently observed that one may simplify (14) as follows, by using the fact that $\sum_{i=1}^{n} (V_{c_i}(z) - V_{b_{i+1}}(z)) = \sum_{i=1}^{n} (V_{c_i}(z) - V_{b_i}(z))$ and (11):

$$L(z) = \frac{\sum_{i=1}^{n} \lambda_i (1 - z_i) S_{c_i}(z)}{\sum_{i=1}^{n} \lambda_i (1 - z_i)}. \tag{20}$$

This formula is remarkably simple; notice that it does not involve the service time distributions, and that the service disciplines at the various queues do not play a role either, which confirms that (14) is based on very general principles. A short proof of this formula was subsequently presented in Boon *et al.* [22]. That proof is based on a very simple, yet very general, balance equation for n-dimensional queue length processes just before arrivals and just after departures, and on PASTA. For marginal queue lengths it reduces to a classical up- and down-crossing identity.

4 The joint queue length distribution at polling epochs

In Subsection 3.4, we have seen that Eisenberg's results [73] yield simple relations between the PGF $S_{c_i}(z)$ of the joint queue length vector at service completion epochs (or $S_{b_i}(z)$, at service beginning epochs) and the PGF's $V_{b_i}(z)$ and $V_{c_i}(z)$ of the joint queue length vector at visit beginning and visit completion epochs. Here, again, $z = (z_1, \ldots, z_n)$. We now restrict ourselves to polling models for which the service discipline at each queue satisfies the following property:

Property 4.1
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 PGF $h_i(z_1, \ldots, z_n)$, which may be any n-dimensional PGF.

Resing [144] (see also Fuhrmann [94]) 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. Resing [144] has pointed out that, for this class of polling systems, the joint queue length process at visit instants of a fixed queue is a so-called *multi-type branching process* with immigration. The theory of multi-type branching processes (cf. Athreya and Ney [11], Resing [143]) thus leads to an expression for the PGF of the joint steady-state queue length process at visit beginning (polling) instants (which exists if $\rho < 1$ and $s_i < \infty$ for all i). Property 4.1 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 $V_{b_i}(\cdot)$:

$$V_{c_i}(z) = V_{b_i}(z_1, \ldots, z_{i-1}, h_i(z), z_{i+1}, \ldots, z_n). \tag{21}$$

Next we relate $V_{b_i}(z)$ to $V_{c_{i-1}}(z)$. That will allow us – after n steps – to express, say, $V_{b_1}(\cdot)$ into itself, and finally to obtain an explicit expression for $V_{b_1}(z)$. The PGF's $V_{c_i}(\cdot)$, $S_{b_i}(\cdot)$ and $S_{c_i}(\cdot)$ then also follow.

In our analysis we follow Resing [144]. We distinguish 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

$$f(z) := (f_1(z), \ldots, f_n(z)), \tag{22}$$

with

$$f_i(z) := h_i(z_1, \ldots, z_i, f_{i+1}(z), \ldots, f_n(z)) \tag{23}$$

for $\mid z_j \mid \leq 1$, $j = 1, \ldots, n$. This is the *offspring* PGF, the PGF 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.
For $\mid z_j \mid \leq 1$, $j = 1, \ldots, n$, define

$$f^{(0)}(z) := z, \quad f^{(k)}(z) := f(f^{(k-1)}(z)), \qquad k \geq 1.$$

Case I: Non-zero switchover times
Observe that

$$V_{b_i}(z) = V_{c_{i-1}}(z)\sigma_{i-1}(\sum_{j=1}^{n} \lambda_j(1 - z_j)). \tag{24}$$

Substituting (21) into (24),

$$V_{b_i}(z) = V_{b_{i-1}}(z_1, \ldots, z_{i-2}, h_{i-1}(z), z_i, \ldots, z_n)\sigma_{i-1}(\sum_{j=1}^{n} \lambda_j(1 - z_j)). \tag{25}$$

Applying (25) n times (which corresponds to following the server during one full cycle w.r.t. Q_1),

$$V_{b_1}(z) = V_{b_1}(f(z))g(z), \tag{26}$$

with

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

The function $g(\cdot)$ represents the 'immigration process' of this multi-type branching process: it is the PGF 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 (26) yields

$$\begin{aligned} V_{b_1}(z) &= \prod_{k=0}^{\infty} g(f^{(k)}(z)) \qquad (27)\\ &= \prod_{k=0}^{\infty}\prod_{i=1}^{n} \sigma_i(\sum_{j=1}^{i} \lambda_j(1 - f_j^{(k)}(z)) + \sum_{j=i+1}^{n} \lambda_j(1 - f_j^{(k+1)}(z))), \end{aligned}$$

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

Case II: Zero switchover times
In the 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):

$$V_{b_i}^0(z) = V_{c_{i-1}}^0(z), \tag{28}$$

for $i = 2, \ldots, n$. The relation between $V_{b_1}^0(z)$ and $V_{c_n}^0(z)$ deserves special attention, because of our convention concerning the behavior of the server when the system is empty. When all queues in the model with zero switchover times become empty, S makes a full cycle, 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 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

$$V_{b_1}^0(z) = V_{c_n}^0(z) - V_{b_1}^0(0)[1 - g^0(z)], \tag{29}$$

with

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

The function $g^0(\cdot)$ represents the 'immigration process' of the multi-type branching process: it is the PGF of the arrival process of customers during periods in which the system is empty.

Remark 4.1
Although we 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 [144], who assumes that when the system is empty, 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 theory of multi-type branching processes and Eisenberg's approach.*

□

Substituting (21) into (28),

$$V_{b_i}^0(z) = V_{b_{i-1}}^0(z_1, \ldots, z_{i-2}, h_{i-1}(z), z_i, \ldots, z_n) \tag{30}$$

for $i = 2, \ldots, n$. Starting from (29) and (21) for $i = n$, and subsequently using (30) for $i = n, n-1, \ldots, 2$, one obtains

$$V_{b_1}^0(z) = V_{b_1}^0(f(z)) - V_{b_1}^0(0)[1 - g^0(z)]. \tag{31}$$

Iterating (31) yields

$$V_{b_1}^0(z) = 1 - V_{b_1}^0(0) \sum_{k=0}^{\infty} [1 - g^0(f^{(k)}(z))] = 1 - V_{b_1}^0(0) \sum_{k=0}^{\infty} \sum_{i=1}^{n} \frac{\lambda_i}{\lambda} (1 - f_i^{(k)}(z)), \tag{32}$$

with

$$V_{b_1}^0(0) = \left[1 + \sum_{k=0}^{\infty} [1 - g^0(f^{(k)}(0))] \right]^{-1} = \left[1 + \sum_{k=0}^{\infty} \sum_{i=1}^{n} \frac{\lambda_i}{\lambda} (1 - f_i^{(k)}(0)) \right]^{-1},$$

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

From (27) and (32) we see that $V_{b_1}(z)$ as well as $V_{b_1}^0(z)$ is determined by $\sum_{j=1}^{n} \lambda_j (1 - f_j^{(k)}(z))$.

Remark 4.2
It is worth observing that the Globally gated service discipline [45] as described in Section 2 does not *satisfy Property 4.1. At the same time, the PGF's* $V_{b_i}(\cdot)$ *and* $V_{c_i}(\cdot)$ *can all be expressed in terms of the joint queue length PGF* $V_{b_1}(\cdot)$ *at the start of a cycle. Indeed, Globally gated is, arguably, the most tractable service discipline, providing a useful testing ground for novel concepts. Altman* et al. *[6] consider the elevator variant of Globally gated, where the various queues are visited in alternating order. From an application perspective, it might be interesting to consider the concept of a reservation mechanism, which also underlies Globally gated, in more detail. For example, customers at some queue might have a certain window of opportunity to make a reservation for service in the next visit period of that queue. See Abidini* et al. *[1] for an application in optical switches.*

4.1 Marginal queue lengths and waiting times

Above, the joint queue length PGF'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 4.1 holds for the service disciplines at all queues. In Subsection 3.4, we already obtained a decomposition for the PGF of the marginal queue length distribution 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 PGF $\mathrm{E}(y^{\mathbf{N}_{i|I}})$). In particular, denoting

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

with y as i-th argument, it follows from (12) and (21) for the case of non-zero switchover times that

$$\mathrm{E}(y^{\mathbf{N}_{i|I}}) = \frac{\tilde{V}_{b_i}(\tilde{h}_i(y)) - \tilde{V}_{b_i}(y)}{(1-y)\tilde{V}'_{b_i}(1)(1-\tilde{h}'_i(1))}; \tag{33}$$

the same result holds for the case of zero switchover times, replacing $\tilde{V}_{b_i}(\cdot)$ by $\tilde{V}^0_{b_i}(\cdot)$ in (33). Similarly indicating queue lengths, and waiting times, by a superscript 0 in the case of zero switchover times, one finds [29]:

$$\mathrm{E}(y^{\mathbf{N}_i}) = \mathrm{E}(y^{\mathbf{N}^0_i})\frac{[\tilde{V}_{b_i}(\tilde{h}_i(y)) - \tilde{V}_{b_i}(y)]\tilde{V}^{0\prime}_{b_i}(1)}{[\tilde{V}^0_{b_i}(\tilde{h}_i(y)) - \tilde{V}^0_{b_i}(y)]\tilde{V}'_{b_i}(1)}, \tag{34}$$

$$\mathrm{E}(\mathrm{e}^{-\omega\mathbf{W}_i}) = \mathrm{E}(\mathrm{e}^{-\omega\mathbf{W}^0_i})\frac{[\tilde{V}_{b_i}(\tilde{h}_i(1-\omega/\lambda_i)) - \tilde{V}_{b_i}(1-\omega/\lambda_i)]\tilde{V}^{0\prime}_{b_i}(1)}{[\tilde{V}^0_{b_i}(\tilde{h}_i(1-\omega/\lambda_i)) - \tilde{V}^0_{b_i}(1-\omega/\lambda_i)]\tilde{V}'_{b_i}(1)}. \tag{35}$$

For exhaustive service, $\tilde{h}_i(\cdot) \equiv 1$; for gated service, $\tilde{h}_i(y) = \beta_i(\lambda_i(1-y))$.
Let us now (without loss of generality) concentrate on $\mathbf{W}_1$ and $\mathbf{W}^0_1$. After some calculations [29], one gets:

$$\mathrm{E}(\mathrm{e}^{-\omega\mathbf{W}_1}) = \mathrm{E}(\mathrm{e}^{-\omega\mathbf{W}^0_1})\frac{\tilde{V}_{b_1}(\tilde{h}_1(1-\omega/\lambda_1)) - \tilde{V}_{b_1}(1-\omega/\lambda_1)}{s[\tilde{H}(\tilde{h}_1(1-\omega/\lambda_1)) - \tilde{H}(1-\omega/\lambda_1)]}, \tag{36}$$

which for exhaustive service ($\tilde{h}_1(\cdot) \equiv 1$) and gated service ($\tilde{h}_1(1-\omega/\lambda_1) = \beta_1(\omega)$), corresponds to Theorems 2 and 5 in Srinivasan *et al.* [151], respectively.

Remark 4.3
The above results expose a close similarity between the cases with and without switchover times. Before [29], models with switchover times and 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 infinity, leading to degenerate distributions at such epochs, cf. Levy and Kleinrock [127] and Eisenberg [75]. The relationship between the two models has further been exposed in Cooper, Niu and Srinivasan [67], Fuhrmann [95], and Srinivasan, Niu and Cooper [151]; in [29] some of their results are unified and generalized.

4.2 Computational aspects

The above results provide a basis for a very efficient numerical calculation of the mean waiting times as well as higher-order moments [29]. The number of elementary operations (additions, multiplications) involved for calculating the mean waiting time at a single queue is $O(n\log_\rho(\epsilon))$, with ϵ the desired level of accuracy. This is comparable to the computational complexity of the so-called descendant-set approach developed by Konheim and Levy [117] and the so-called station time method of Ferguson and Aminetzah [85] which entail solving a system of n^2 equations for obtaining the mean waiting times at all n queues. These methods provided a significant reduction in computational complexity compared to the original buffer occupancy method described by Cooper [65], Cooper and Murray [66] and Eisenberg [73] which required solving a system of n^3 equations for determining the mean waiting times at all n queues. The Mean Value Analysis developed by Winands, Adan and Van Houtum [166], as further discussed in Section 7, also provides an efficient way to determine mean sojourn times, as demonstrated in Van der Gaast *et al.* [99] for a model with batch arrivals. It additionally offers a basis for approximations of mean queue lengths and mean delays.

We close this subsection by remarking that (i) Equation (11) of [144] provides *exact* (non-numerical) moment expressions for branching-type polling models, and (ii) Choudhury and

Whitt [58] present an elegant method to obtain moments and tail probabilities in polling models via numerical inversion of transform expressions.

5 Two-queue polling systems which are not of branching type

There appears to be a sharp division between 'easy' (branching-type) and 'complicated' polling models. Such a division is not uncommon in queueing theory; one also sees it, e.g., in queueing networks that do or do not satisfy the conditions to have a product form for their joint queue length distribution. If a polling system does not satisfy the branching property, then an exact analysis of queue length and waiting-time distributions generally seems out of reach. Just like in queueing networks, there are a few two-queue exceptions; in the present section we consider some of those. We restrict ourselves to the case of non-zero switchover times. Starting point is a relation between the two-dimensional queue length generating functions $V_{b_1}(z) = V_{b_1}(z_1, z_2)$ and $V_{b_2}(z_1, z_2)$ at server visits to Q_1 and Q_2, respectively. When the branching property holds, this relation is given by (25), which could be iterated to yield an infinite product. Now consider the case in which Q_1 receives exhaustive service and Q_2 receives 1-limited service. Then (26) is replaced by

$$\begin{aligned} V_{b_1}(z_1, z_2) &= \frac{\beta_2(z_1, z_2)\sigma_2(z_1, z_2)}{z_2}[\sigma_1(z_1, z_2)V_{b_1}(h_1(z_1, z_2), z_2) \\ &- \sigma_1(z_1, 0)V_{b_1}(h_1(z_1, 0), 0)] + \sigma_2(z_1, z_2)\sigma_1(z_1, 0)V_{b_1}(h_1(z_1, 0), 0). \end{aligned} \tag{37}$$

Ibe [109] has obtained the marginal queue length transform for Q_1 at polling instants of that queue; Groenendijk ([103], Section 6.3) used (37) to obtain an explicit expression for $V_{b_1}(z_1, z_2)$. The key to solving (37) is the observation that, because service at Q_1 is exhaustive, one has $h_1(z_1, z_2) = \pi_1(\lambda_2(1 - z_2))$ with $\pi_1(\cdot)$ the LST of a busy period of $M/G/1$ queue Q_1 in isolation. Because this function does not depend on z_1, $V_{b_1}(h_1(z_1, 0), 0)$ is a constant, not depending on z_1. Hence the only unknown *functions* in (37) are $V_{b_1}(z_1, z_2)$ and $V_{b_1}(h_1(z_1, z_2), z_2)$, and the substitution $z_1 = \pi_1(\lambda_2(1 - z_2))$ (plus the normalization condition) solves the problem. For a study of the two-queue case with exhaustive service at Q_1 and k-limited service at Q_2, we refer to Ozawa [140] and Winands *et al.* [167].

It is perhaps not that surprising that the two-queue exhaustive/1-limited model is easy to analyze; in the case of zero switchover times, it reduces to a classical queueing model with two customer classes and non-preemptive priority for class 1. It is surprising, though, that the two-queue *gated*/1-limited model has not succumbed to an exact analysis; in Boon *et al.* [23] it is suggested that determination of $V_{b_1}(z_1, z_2)$ for that model might be accomplished by solving a so-called boundary value problem of a complicated type.

Several two-queue polling models *have* been solved via a formulation as a boundary value problem; we now turn to this line of research.

Eisenberg [74] studies a two-queue polling model with 1-limited service at both queues, and without switchover times. He transforms the problem of determining $V_{b_1}(z_1, z_2)$ into the problem of solving a singular integral equation (a complex Fredholm integral equation of the second kind). As the author indicates, due to the difficult nature of the mathematics, some steps in the solution remain to be proven. In [64], a different approach for this same model is given. Below we sketch that approach, for the more general case of non-zero switchover times (cf. [38]). Starting-point in [38] again is the functional equation for $V_{b_i}(z_1, z_2)$:

$$\begin{aligned} K(z_1, z_2)V_{b_1}(z_1, z_2) &= V_{b_1}(0, z_2)\left[\beta_2(z_1, z_2)\sigma_1(z_1, z_2)\sigma_2(z_1, z_2)(z_1 - \beta_1(z_1, z_2))\right] \\ &+ V_{b_2}(z_1, 0)\left[z_1\sigma_2(z_1, z_2)(z_2 - \beta_2(z_1, z_2))\right], \end{aligned} \tag{38}$$

with $K(z_1, z_2)$ the *kernel* of the functional equation, defined as

$$K(z_1, z_2) := z_1 z_2 - \beta_1(z_1, z_2)\beta_2(z_1, z_2)\sigma_1(z_1, z_2)\sigma_2(z_1, z_2). \tag{39}$$

The appearance of the functions $V_{b_1}(0, z_2)$ and $V_{b_2}(z_1, 0)$ corresponds to a server arriving at an empty queue. Once they have been obtained, $V_{b_1}(z_1, z_2)$ is also known. The key in the analysis in [38] is that, according to its definition as a probability generating function, $V_{b_1}(z_1, z_2)$ should be analytic inside the product of unit circles $|z_1| < 1$, $|z_2| < 1$. Hence every zero of $K(z_1, z_2)$ in that region should also be a zero of the righthand side of (38). The ensuing relation between $V_{b_1}(0, z_2)$ and $V_{b_2}(z_1, 0)$ is thus translated into a Riemann boundary value problem – a problem in which two functions are related on a closed contour, while one function is analytic inside that contour (and continuous upto the boundary) and the other function is analytic outside that contour (and continuous upto the boundary). By solving such a Riemann problem, $V_{b_1}(0, z_2)$ and $V_{b_2}(z_1, 0)$ are obtained. In [64], for the case of zero switchover times, a similar approach was followed, resulting in a (somewhat simpler) Dirichlet boundary value problem.

Cohen [63] studies a two-queue polling model with semi-exhaustive (also called decrementing) service: the server stays in a non-empty queue until the number of customers present has become one smaller than the number found upon its arrival to the queue. The joint queue length distribution at visit completion epochs is obtained by formulating and solving a Riemann boundary value problem.

Several studies consider two-queue polling models with Bernoulli service. Under this service discipline, if both queues are non-empty and the server is at Q_j, a customer from Q_j is served with probability p_j and a customer from the other queue is served with probability $1 - p_j$. The case with $p_1 = 1$ and $0 \leq p_2 < 1$ was solved by Weststrate and Van der Mei [162] via an iterative process. The case that p_1, too, is less than one is harder. Both Lee ([124], zero switchover times) and Feng, Kowada and Adachi ([83], non-zero switchover times) treat this model by using boundary value techniques. Lee formulates a Riemann boundary value problem with a shift, and translates it to a Fredholm integral equation which he solves. Feng *et al.* [83] also formulate and solve a Riemann boundary value problem.

Finally, we would like to observe that it seems unlikely that an exact analysis will be provided for an n-queue polling model, with $n > 2$, in which none of the queues has a branching-type service discipline. This belief is based on the lack of a boundary value approach in dimensions higher than two. Analytic-numerical approaches like the power-series algorithm could be used in such cases, see Blanc [17].

6 The input process

The polling literature focuses almost exclusively on the case of customers arriving according to independent Poisson processes, the service requirements at the various queues moreover being independent sequences; the resulting input processes hence are independent compound Poisson processes. In this section we consider some generalizations of these assumptions.

(i) BMAP arrivals. Saffer and Telek [147] consider a polling model with either exhaustive or gated service, in which the arrival processes at the n queues are independent Batch Markovian Arrival Processes (BMAP). They developed a generalization of the so-called buffer occupancy method, a classical method for analyzing queue lengths in polling systems, first presented by Cooper and Murray [66].

(ii) Renewal arrivals. Bertsimas and Mourtzinou [16] consider a polling model with independent renewal arrival processes at the various queues. For the case of gated service at all queues, they derive expressions for the mean delays in heavy traffic, expressing these in cycle time variances which can be obtained by solving a system of $n \times n$ equations. Van der Mei and Winands [134] build upon their result, allowing general switchover times and providing closed-form expressions for scaled mean delays in heavy traffic. Boon *et al.* [27] combine light- and

heavy-traffic approximations, via interpolation, to come up with accurate mean waiting-time approximations for polling systems with both gated and exhaustive service.

Another type of approximation is provided in a few papers of Tran-Gia, see in particular [157]. He presents a discrete-time analysis of polling systems with finite buffers, 1-limited service and general renewal input. His method is based on the use of efficient discrete convolution operations, using fast convolution algorithms like the Fast Fourier transform.

(iii) Correlated arrivals. Levy and Sidi [128] study a polling system with correlated Poisson arrival streams. They consider gated and exhaustive service, and derive linear equations whose solution yields the mean delays. They also derive a pseudo-conservation law for the mean delays. They extend their analysis in [130] to the case of Poisson arrivals of customer batches with correlated numbers $(K_1, \ldots, K_n)$, destined for queues $Q_1, \ldots, Q_n$. A workload decomposition and general pseudo-conservation law for a polling model with such a batch Poisson arrival process is presented in [33]. Van der Gaast *et al.* [99] derive the sojourn time LST of a batch, for exhaustive, gated and Globally gated service; a batch here may contain customers of various queues.

(iv) Lévy input. Recently there has been a growing interest in queueing models with as input a Lévy process ('Lévy-driven queues', see Debicki and Mandjes [69]). Lévy processes are processes with stationary, independent increments. Compound Poisson processes, Brownian motion and linear increment processes are some special cases. The generalization from a compound Poisson input (as in an $M/G/1$ queue) to a Lévy input implies that one can no longer speak of customers and queue lengths; the focus naturally shifts to workloads. There is hardly any literature on Lévy-driven polling systems. A pioneering paper is due to Eliazar [76], who studies Lévy-driven polling systems under the gated discipline, using a dynamical systems approach. Czerniak and Yechiali [68] consider fluid input at all queues, which may be seen as a special case of Lévy input. In [39] a very general arrival process is allowed: the input process is an n-dimensional Lévy subordinator (i.e., non-decreasing sample paths, which is of course natural for an input process). Correlations between the inputs at the various queues are allowed. Moreover, the input process may change at polling and switchover instants, implying that one can have different input processes at different server positions. The transition from compound Poisson process to Lévy subordinator implies that one no longer has the branching Property 4.1, which is stated in terms of numbers of customers. The authors of [39] identify the analogous branching property in a continuous state space setting, that allows describing the multi-dimensional workload at successive polling instants at a fixed queue as a multi-type continuous state space, discrete-time, branching process. This is referred to as a multi-type Jirina branching process ([112]; MTJBP). The class of service disciplines that satisfy the new branching property is rich, and contains the exhaustive and gated disciplines. Altman and Fiems [4] had also observed the relation between Lévy-driven polling models and MTJBP's, in a special case in which all the queues are fed by identical Lévy subordinators. Employing the Kella-Whitt martingale, the LST of the joint steady-state workload distribution at an *arbitrary* epoch is also obtained in [39]. Martingales are also the main tool in proving a workload decomposition result for a polling system with multi-dimensional Lévy input [40].

7 Scheduling

Until ten years ago, very few papers diverged from the FCFS assumption for service within a queue of a polling system. In this section, we pay attention to two lines of research which deviate from the FCFS adagium: (i) polling systems with multiple classes of customers per queue, and fixed priorities, and (ii) polling systems in which there is only one class of customers per queue,

but with a service discipline within a queue that is not FCFS but, e.g., Last-Come First-Served (LCFS), processor sharing, Random Order of Service (ROS), or Shortest Job First (SJF).
(i) Multiple customer classes with fixed priorities. Shimogawa and Takahashi [149] derive a PCL for a polling system with fixed priorities within queues, and Fournier and Rosberg [92] consider polling systems with local priorities and with global priorities (where the server moves to the next queue if some queue has a customer of higher priority than the ones in the presently visited queue). They develop a PCL for both model variants.

While most polling+priority studies originated from a computer-communications background, polling systems with multiple customer classes and fixed priorities also arise naturally in the *Stochastic Economic Lot Scheduling Problem* (SELSP), where multiple types of products have to be produced on a single machine with significant setup times. In the SELSP, orders for the same product type are being placed by customers of different priority levels, giving rise to polling models with not only several queues (corresponding to orders for the various product types) but also several customer classes per queue, see Winands [165]. This formed one of the motivations for a series of papers of Boon *et al.* [19, 20, 21]. They analyze the joint queue length distribution for polling models which are of type PS, except for the additional assumption that within each queue there are several classes of customers with fixed priorities. That analysis relies on a relation to multi-type branching processes, cf. Section 4. Boon *et al.* also determine the waiting-time distributions of the various classes of customers. This is done for exhaustive, gated and Globally gated service. A key step of the approach is to determine the joint distribution of the past and residual cycle time at the arrival epoch of the tagged customer. For gated service, the waiting time of a customer of priority level k in Q_i consists of that residual cycle time, plus the services of higher priority customers arriving during the cycle, plus the services of customers of equal priority arriving during the past part of the cycle. For exhaustive service, the procedure is somewhat similar, with a slightly different definition of the cycle time: for gated service, a cycle for Q_i starts at the beginning of a visit to Q_i, whereas for exhaustive service it turns out to be convenient to let the cycle start at the completion of a visit to Q_i.

(ii) One customer class per queue; non-FCFS service. There are quite a few real-world examples of polling situations in which non-FCFS scheduling might be required. In the computer science community, polling models are used to study the Bluetooth and 802.11 protocols, and scheduling policies at routers and i/o subsystems in web servers. The high workload variability in many of these settings makes non-FCFS scheduling appealing, see Wierman *et al.* [163]. In [163] it is argued that the lack of research on scheduling in polling systems is not due to a lack of applications, but rather due to the beliefs that the impact of within-queue scheduling will be small, and that the ensuing mathematical analysis will be very hard. Using the *Mean Value Analysis* (MVA) framework that was developed for polling systems in Winands, Adan and Van Houtum [166], in [163] mean response (=sojourn) times in polling systems with exhaustive or gated service are determined for a wide array of service disciplines: LCFS, Processor Sharing, SJF and Shortest Remaining Processing Time First (SRPT). It turns out that, while varying the scheduling strategy at queues with gated service does not have a major effect, it *does* strongly affect mean delays in the case of exhaustive service. This holds in particular for SRPT, just as in an ordinary $M/G/1$ queue. The reason that the effect is particularly pronounced for exhaustive service is that small jobs which arrive during a visit of their queue take advantage of preemption and thus have very small delays.

The above analysis is extended to sojourn time *distributions* in [35]. The approach globally consists of the following steps: (i) determine the joint queue length distribution at server visit epochs to a queue (restricting attention to polling models which satisfy the branching property); (ii) determine the LST of the cycle time distribution for some queue Q_i; (iii) use this to determine the joint LST of the past and residual part of that cycle time, at the arrival epoch of a customer

at Q_i; (iv) for various service disciplines at Q_i, and now focusing on gated and Globally gated, careful bookkeeping yields the sojourn time LST at Q_i. The analysis for exhaustive service seems more complicated; in Ayesta, Boxma and Verloop [13] the sojourn time LST is obtained for the case of an $M/M/1$ processor sharing queue in a polling system, under the constraint that all other queues also satisfy the branching property. See also Kim and Kim [115] for the case of phase-type service at the processor sharing queue.

8 Asymptotics

In this section we consider two kinds of asymptotics: Many-queue asymptotics and heavy-traffic asymptotics.

8.1 Many-queue asymptotics

Asymptotic regimes where the number of queues in a polling system grows large have received little attention so far. A few authors have studied the case in which the switchover times between successive queues go to zero when the number of queues grows large. In the limit, the polling system then behaves as a "continuous" spatial system with a single server which moves at constant speed along a circle, stopping to perform services when it encounters customers. These customers arrive uniformly on the circle, according to a Poisson process. Initial studies of such a continuous polling system were provided in Coffman and Gilbert [60] and Fuhrmann and Cooper [97]. Their model is generalized by Kroese and Schmidt [122] via an approach that makes use of random measure theory and stochastic integration theory, and which thus also provides a rigorous mathematical basis for the study of continuous polling models.

An interesting model generalization is also proposed by Eliazar [77]. He considers a polling system with gated service and n queues, with a Levy input process and general interdependent switchover times. Letting $n \to \infty$, he proves convergence in law to a limiting polling system on the circle. His proof is based on an asymptotic analysis of stochastic Poincaré maps. The obtained limit is identified as a so-called snowplowing system on the circle (a snowplower cycling along a track, clearing off snow while moving; (cf. [116], pp. 254–255 and 259–264).

Motivated by applications in ferry-assisted wireless local-area networks, Kavitha and Altman have studied several continuous polling variants. See for instance [113], in which nonclassical service disciplines are considered, and in which the continuous polling system is analyzed by discretizing the system in such a way that known pseudo-conservation laws (cf. Subsection 3.3) can be applied. Their results rely heavily on fixed-point analysis of infinite-dimensional operators.

Kroese [120] considers a greedy service policy: after completion of a service, the server always moves in the direction of the nearest customer. The stability condition for this system, and several interesting open problems, are discussed in Rojas Nandayapa, Foss and Kroese [146]. Those open problems concern stability issues as well as characterization of the random measure describing the steady-state customer positions. This is done in a more general setting than polling on a circle; customers may arrive in some space, and are served by one or more servers roaming that space. We refer to [146] for an extensive set of references on continuous polling.

In Meyfroyt *et al.* [136] another type of scaling with a large number of queues is studied. Motivated by token passing algorithms for communication channels with medium access control and a large number of nodes, [136] considers the following scenario: the number of queues grows large while the total arrival rate is kept fixed and the individual switchover time and service time distributions remain the same. This asymptotic regime leads to cycles of infinite length and queue lengths with non-trivial distributions. Explicit pre-limit expressions are derived for the covariance of queue lengths, the covariance of visit times and the variance of the cycle time for symmetric polling systems in which the server uses a branching-type discipline. This leads

to explicit expressions for $\lim_{n\to\infty}\mathrm{E}[\mathbf{C}/n]$ and $\lim_{n\to\infty}n\mathrm{Var}(\mathbf{C}/n)$. Those results reveal that, since $\mathrm{Var}(\mathbf{C}/n)$ is of order $1/n$, the scaled cycle time $\mathbf{C}/n$ converges in probability to a deterministic value. This implies that the queue lengths at the various nodes become asymptotically independent. In the limit, the individual queues appear to behave as discrete-time bulk service queues. It is suggested in [136] that these properties of $\mathbf{C}/n$ and of the individual queues remain valid for symmetric polling systems with a large number of queues and more general non-idling service disciplines – which are not necessarily of the branching-type.

8.2 Heavy-traffic asymptotics

Pioneering papers regarding the heavy-traffic behavior of polling systems were written by Coffman, Puhalskii and Reiman [61, 62]. In [61], the focus is on a two-queue polling model with renewal arrival processes and exhaustive service at both queues, and with zero switchover times. The authors first apply standard heavy-traffic assumptions and scalings; they let $\sqrt{m}(1-\rho)$ approach a constant with m going to infinity, and show that the normalized total workload process $W(mt)/\sqrt{m})$ weakly converges to reflected Brownian motion (RBM). For this they can rely on a known $G/G/1$ result, because of work conservation. They subsequently show that the scaled workloads of individual queues change at a rate that becomes infinite in the limit. They then formulate an averaging principle for individual workloads, in which during one polling cycle these scaled workloads linearly decrease to zero (during visit periods of the corresponding queue) and linearly increase (during the subsequent intervisit period), while the total scaled workload in the system during such a cycle basically stays the same. Individual workloads change a factor $\sqrt{m}$ faster than the total workload. Put differently: when the total scaled workload equals x, the scaled workload at an individual queue is uniformly distributed on $[0, x]$. While in [61] a rigorous proof is only provided for the two-queue case with identical service time distributions, the authors convincingly argue that such an averaging principle should also hold in the n-queue case, with not necessarily identical service time distributions.

Coffman *et al.* prove in [62] that the averaging principle carries over to the case of non-zero switchover times. Because of those switchover times, they first have to replace the RBM heavy-traffic limit for the *total* workload by a Bessel-type diffusion limit. Two key elements of their subsequent approach are: (i) they first prove the averaging principle for a so-called threshold queue, a single queue in isolation with a server which only starts serving when the workload exceeds some value T; (ii) they strongly rely on a semi-martingale representation of the workload process, which allows them to use general convergence results for semi-martingales.

The Coffman-Puhalskii-Reiman papers have given rise to several lines of research. Olsen [138] provides a heuristic refinement of the averaging principle, which improves the accuracy of the resulting approximation for waiting-time distributions under moderate load. In several studies it is argued, without a rigorous proof, that the averaging principle of [61, 62] holds in far greater generality. We refer to Section 2 of Markowitz, Reiman and Wein [131] for an excellent discussion of the heavy-traffic averaging principle and further references, here only mentioning the interesting extensions to polling systems in tandem in Reiman and Wein [142] and to the stochastic economic lot scheduling problem [131]. Olsen and Van der Mei [139], too, conjecture that the heavy-traffic averaging principle holds in considerable generality, and apply it to polling models with renewal arrivals, exhaustive or gated service at the queues, and service according to a polling table. They also use their heavy-traffic limiting result to provide accurate approximations for waiting-time distributions under moderate to heavy load. A similar approach is followed in Boon *et al.* [25] for a network with a single roving server, leading to a heavy-traffic limiting result for the distribution of the total sojourn time of a customer in the network, when following a specific path. In combination with a novel light-traffic approximation, this yields an approximation for the mean total sojourn time along a specific path, which is highly accurate

for a wide range of traffic loads. Jennings [111] uses a new technique to prove the validity of a heavy-traffic averaging principle for a vector of *weighted queue lengths* in a polling system with zero switchover times, and with a certain parameterized set of gated and exhaustive service disciplines. Each queue length is weighted by its mean processing time.

Finally we mention three results of a different type. Firstly, Van der Mei [132] develops a heavy-traffic approach which is quite different from the one in [61, 62]. He restricts himself to branching-type polling systems, and then exploits Theorem 4 of Quine [141] for multi-type Galton-Watson branching processes, in which the maximal eigenvalue of the so-called mean matrix (of numbers of descendants) approaches the critical value 1. Using Resing's relation between the numbers of particles in multi-type branching processes [144] and the numbers of customers in the various queues at server polling epochs, he is able to obtain the heavy-traffic limiting behavior of the queue lengths. See also Abidini, Dorsman and Resing [2] for a related heavy-traffic result for a polling model with retrials and so-called glue periods that models the dynamics of optical switches; in [2] heavy-traffic asymptotics for the *joint* queue length process are derived. Interestingly, Kroese [121] provides a heavy-traffic analysis of a continuous polling system on the circle (cf. Subsection 8.1), by exploiting the relationship between such systems and *age-dependent* branching processes.

Secondly, Boon and Winands [26] consider a two-queue polling system with zero switchover times and k_i-limited service at Q_i, $i = 1, 2$, under Markovian assumptions. Applying a singular perturbation technique, they derive the heavy-traffic behavior of the joint queue length vector. The queue length of the critically loaded queue (Q_2) appears to be exponentially distributed after an appropriate scaling, whereas the queue length of Q_1 is distributed as that of a queue in isolation with Erlang-k_2 distributed vacations. This reveals a heavy-traffic behavior that is quite different from the heavy-traffic behavior of the branching-type polling models studied in the papers mentioned above.

Thirdly, Bekker *et al.* [15] consider polling models with the gated or Globally gated service policy and several *non-FCFS* service disciplines. They derive asymptotic closed-form expressions for the LST of scaled (by a factor $1 - \rho$) waiting times and sojourn times in heavy traffic. For FCFS, it was already known that the scaled sojourn times are of the form $\mathbf{U\Gamma}$, with $\mathbf{U}$ and $\mathbf{\Gamma}$ independent, $\mathbf{U}$ uniformly distributed and $\mathbf{\Gamma}$ Gamma distributed. In [15] this result is also shown to hold for LCFS, while one has $\tilde{\mathbf{U}}\mathbf{\Gamma}$ for ROS with $\tilde{\mathbf{U}}$ having a trapezoidal distribution; for processor sharing and SJF one gets $\tilde{\mathbf{U}}^*\mathbf{\Gamma}$, with $\tilde{\mathbf{U}}^*$ having a generalized trapezoidal distribution. These results lead to accurate waiting- and sojourn time approximations. Vis, Bekker and Van der Mei [158] consider the same heavy-traffic problem for the case of exhaustive service at all queues. In that case, the scaled sojourn times are of the form $\mathbf{\Theta\Gamma}$, where $\mathbf{\Theta}$ is related to a uniformly distributed random variable.

9 Some miscellaneous topics

In this section we discuss some miscellaneous topics, which did not fit in the framework of the previous sections: (i) multiple-server polling systems, (ii) disciplines with service limits, (iii) optimization of polling systems and (iv) queue-length dependent server behavior. Unfortunately, we could not cover some interesting topics like the concept of the dormant server which stays at a queue when the system has become empty; the concept of the smart customer whose arrival rate is determined by the server location; and the concept of fairness. The latter concept may deserve more attention than it has so far received in the literature [148, 164], because it is closely related to the important question which queue to serve next, and which service discipline to use at a queue.

9.1 Multiple-server polling systems

As reflected in Assumption 1 in Section 2, we have focused on polling systems with a single server. Although multiple-server polling systems find a wide range of applications, they have received relatively limited attention, and hardly any exact distributional results are available, since the combination of several queues and multiple servers yields highly complex behavior.

In multiple-server models, it is useful to distinguish between two scenarios with either *coupled* servers which always visit the various queues as a group or *independent* servers which each visit the queues individually. Browne and Weiss [51] establish index-type rules for determining the visit order that myopically minimizes the expected length of individual cycles in systems with coupled servers and exhaustive or gated service. Browne *et al.* [49] and Browne and Kella [50] analyze two-queue models with an infinite number of coupled servers and deterministic service times at one of the two queues. Vlasiou and Yechiali [160] analyze polling systems with an infinite number of coupled servers and random visit durations. Borst [28] explores the class of models with multiple coupled servers that satisfy a slight generalization of branching Property 4.1, and allow an exact analysis of the joint queue length distribution at polling epochs, the marginal queue length distributions, and the waiting-time distributions.

Models with multiple independent servers arise in scenarios where several queues may be served concurrently, as is commonly the case in a wide range of applications, e.g., token ring and optical communication networks, elevator systems, and signalized traffic intersections. In a pioneering paper Morris and Wang [137] derive the mean cycle time of each server and the mean intervisit time to a queue, and present approximate expressions for the mean sojourn time for both a gated-type and a limited-type service discipline. An interesting phenomenon observed in [137] is that multiple independent servers have a tendency to cluster if they follow identical routes, especially in high load conditions. This phenomenon is somewhat reminiscent of the tendency for city buses and trams to bunch together on heavily traveled routes, and may be visualized and understood as follows. A server that is behind will tend to move fast, as it only encounters recently served queues, whereas a server at the front will tend to be slowed down by queues that have not been served for a while. Thus the servers tend to form bunches while constantly leapfrogging over one another. Obviously, the bunching of servers is alleviated if they follow different routes, and Morris and Wang [137] therefore advocate the use of 'dispersive' policies to improve the system performance. Gamse and Newell [100, 101] use multiple-server polling models to study elevator operations where similar bunching behavior can occur.

Borst and Van der Mei [32] provide mean waiting-time approximations which exploit pseudo-conservation-like concepts (which had proven to be valuable in the single-server case, cf. Boxma and Meister [46, 47]) and explicitly account for the tendency of the servers to cluster as function of their visit orders. Van der Mei and Borst [133] demonstrate how performance metrics in polling systems with multiple independent servers may be calculated using the so-called power-series algorithm.

In recent papers Antunes *et al.* [8] and Robert and Roberts [145] propose mean-field approximations for the capacity of multiple-server polling systems with a large number of queues and a limit on the number of servers that can visit a queue simultaneously, motivated by applications in passive optical networks. Finally, it is worth observing that the analysis of optimal dynamic routing policies and service disciplines for polling systems with multiple independent servers is closely related to that of selecting an optimal service vector in 'switched' networks with several potential schedules and reconfiguration delays as considered in [9, 54, 56, 108, 161].

9.2 Disciplines with service limits

Disciplines with service limits as described in Section 2 are commonly adopted in practice to regulate the amount of service provided to each of the queues during a visit. Such service limits

can either be specified in terms of the maximum number of customers served during a visit or the maximum time duration of a visit, and can be leveraged to bound the cycle time. Moreover, these limits provide a mechanism to achieve prioritization, by assigning different service limits to different queues, according to their relative importance.

Although these disciplines are widely implemented, they are difficult to analyze and hence it is not well understood how to set the service limits so as to achieve target performance levels. Note that k-limited service disciplines do not satisfy Property 4.1 and exact results are only available for special cases, such as completely symmetric systems with 1-limited service and a few two-queue scenarios as discussed in Section 5. Polling systems with time-limited service have not yielded to an exact analysis in any degree of generality either. Coffman, Fayolle and Mitrani [59] and De Haan, Boucherie and Van Ommeren in a series of papers (see, e.g., [104]) present interesting results for exponentially distributed time limits. Leung [126] develops a numerical technique for analyzing systems with a probabilistically limited service discipline.

The fact that disciplines with service limits are widely deployed, yet extremely hard to analyze, has considerably added to the importance of the pseudo-conservation laws discussed in Section 3 in constructing and validating waiting-time approximations. Boxma and Meister [47] use the pseudo-conservation law to derive waiting-time approximations for 1-limited service. Groenendijk [103] presents a more refined procedure to compute such approximations. For the general case of k-limited service, the pseudo-conservation law still contains an unknown term. Fuhrmann and Wang [98] obtain waiting-time approximations for k-limited service by bounding that term. Everitt [78, 80] derives such approximations by approximating that term. Chang and Sandhu [57] present a more refined procedure to calculate waiting-time approximations for k-limited service.

9.3 Optimization of polling systems

Optimization in polling systems is a multi-faceted problem which has been actively pursued, though it remains somewhat under-explored compared to the analysis of polling systems. We refer to Boxma [34] (static optimization) and Yechiali [168] (semi-dynamic optimization) for surveys, and here only highlight a few of the main issues.

In the optimization of polling systems, there are two key factors that play a role: first, what is the performance metric to be optimized, and second, what is the class of feasible scheduling policies. As for the first factor, a commonly adopted optimization criterion is a weighted sum of the mean waiting times at the various queues. Concerning the second factor, usually the class of feasible scheduling policies consists of a family of strategies of similar structure that differ by some (vectorial) parameter. Typical examples include a routing vector (routing probabilities, or polling table), a service vector (service probabilities, or service limits), or a routing vector and a service vector simultaneously, which we now briefly discuss in succession.

Optimization of the routing policy for a given service policy

A considerable research effort has been devoted to static optimization, i.e., optimization of static routing policies, in which the routing decisions are made independently of the state of the system. Boxma, Levy and Weststrate [43] consider a system with random polling, and either exhaustive or gated service at each of the queues. They address the problem of finding the routing probabilities that minimize $\sum_{i=1}^{n} \rho_i \mathrm{E}\mathbf{W}_i$, the latter quantity being explicitly known from the pseudo-conservation law for random polling, cf. [48]. They subsequently use this to determine a polling table that minimizes $\sum_{i=1}^{n} \rho_i \mathrm{E}\mathbf{W}_i$, or, more generally [44], to determine a polling table that minimizes a weighted sum of the mean waiting times, the latter quantity now being approximated in terms of the occurrence ratios of the queues in the polling table. Kruskal [123] studies a similar problem with deterministic arrival, service, and switchover processes. In all

cases, the optimal visit ratios are given by surprisingly simple square-root formulas.

Also, a considerable amount of research effort has been put to semi-dynamic optimization, i.e., optimization of semi-dynamic routing policies, in which periodically the visit order for some future period is determined, based on some partial knowledge of the state of the system; see for instance Browne and Yechiali [52] and Fabian and Levy [81].

Optimization of the service policy for a given routing policy

Borst, Boxma and Levy [31] consider a system with a k-limited service strategy at each of the queues, and address the problem of determining the vector of service limits $(k_1, \ldots, k_n)$ that minimizes a weighted sum of the mean waiting times. Blanc and Van der Mei [18] study a similar optimization problem in a system with a Bernoulli service strategy at each of the queues.

Simultaneous optimization of routing policy and service policy

Borst *et al.* [30] consider a system operated with a fixed time polling (ftp) scheme. An ftp scheme specifies which queue should be visited at what time, i.e., it specifies not only the *order* of the visits, but also the *starting times* of the visits, and addresses the problem of constructing an ftp scheme that minimizes a weighted sum of the mean waiting times.

For a model with zero switchover times, the optimal (non-preemptive) polling policy is known to be given by the $c\mu$-rule, cf. Meilijson and Yechiali [135], and Buyukkoc, Varaiya and Walrand [55]. For a symmetric two-queue model with non-zero switchover times, Hofri and Ross [106] show that the policy that minimizes the sum of discounted switchover times and the holding cost, is exhaustive service in a nonempty system, and is of threshold type for switching from an empty queue to another. For an asymmetric two-queue model with switchover *costs* rather than switchover *times*, Koole [118] shows that the policy that minimizes the sum of discounted switching cost and holding cost, is *not* a threshold policy, but that the best threshold policy approaches the optimal policy very well. See the next subsection for some further threshold policies.

9.4 Queue-length dependent server behavior

We briefly mention some studies which are devoted to the exact analysis of two-queue polling models with threshold switching. Lee and Sengupta [125] consider a two-queue system without switchover times. If there are more than L customers at Q_1 after a customer departure, then the server next serves a type-1 customer. If there are at most L customers at Q_1 after a customer departure, then the server checks the type of the last served customer, and serves a customer of the other type (if there is one). Boxma, Koole and Mitrani [42] study a two-queue model with exponential service times and exhaustive service at Q_1. Service at Q_2 is also exhaustive, unless the size of Q_1 reaches a threshold T during a service at Q_2; in the latter case the server switches to Q_1, either preemptively or at the end of the service. The same model, but with general service time distributions and without preemption, is considered in Boxma & Down [36]; that paper also contains a simple, yet accurate, mean queue length approximation which is suitable for optimization purposes. The two-queue model with general service time distributions in Feng *et al.* [84] has two thresholds M and $N > M$ in Q_2. After a service completion in Q_1 that leaves Q_1 non-empty, the server still moves to Q_2 if its queue length exceeds N. After a service completion in Q_2 that leaves Q_2 with at most M customers while Q_1 is not empty, the server switches to Q_1; otherwise, it stays at Q_2. The analysis in each of these four papers [36, 42, 84, 125] focuses on queue length PGF's, and relies on arguments from complex function theory. The two-queue model of Avrachenkov, Perel and Yechiali [12] also has a threshold based policy, but the capacities of both queues are finite. They use a matrix-analytic approach, and expose an interesting oscillation phenomenon. We also refer to this paper for further references on threshold switching.

Remark 9.1

Next to queue-length dependent server behavior, one could also allow queue-length dependent customer behavior. In Adan et al. *[3], a two-queue polling model is analyzed in which customers join the shortest queue. The joint queue length distribution is determined both via the compensation approach and via reduction to a Riemann-Hilbert boundary value problem. Alternatively, one could allow customers to use some form of information about the server position. For example, the arrival rate of customers could depend on whether the server is visiting Q_i, or switching to Q_j. In the case of branching-type service disciplines, one can then still obtain joint queue-length distributions by exploiting properties of multi-type branching processes; a similar statement even holds in the case of Lévy-driven polling models [39].*

10 Suggestions for further research

Polling is a quite broad topic, and there are several ways of listing suggestions for further polling studies. One option is to link an open problem to each of the 10 assumptions of Section 2. Indeed, it would be interesting to obtain more results for multiple-server polling systems (Ass. 1); to devote more attention to spatial polling models (Ass. 2; see Altman and Foss [5]); to relax the assumption of Poisson arrivals (Ass. 3); to allow the loss of customers (Ass. 4); to consider non-cyclic routing, in particular Markovian routing (Ass. 5. This is, a.o., relevant in the setting of random access in wireless communications, see Dorsman *et al.* [70]. It is interesting to notice [144] that the joint queue length process in Markovian polling models is not a multi-type branching process, even if the service policies at all queues are of branching type); to get a better grip on service policies which are not of branching type (Ass. 6); to obtain more results for polling systems with non FCFS service order (Ass. 7); to study the effect of large switchover times, and also to allow the possibility that a switchover time is skipped when the corresponding queue is empty (Ass. 8; see Boon *et al.* [23]); to consider a network of queues with one or more polling servers (Ass. 9; [7, 10, 14, 26, 107, 150]); and to study stability conditions [90] but also the transient behavior of polling systems (Ass. 10).

Rather than "polling" these 10 topics in an exhaustive manner, we prefer to focus on what in our opinion are a few particularly relevant directions for further research:
(i) Exact results for non-branching models are quite scarce, and exact results for branching-type polling models are typically given in the form of sums of infinite products of generating functions. Hence, there is a strong need for readily applicable expressions, which give useful qualitative insights and reasonably accurate quantitative results, like those provided in Federgruen and Katalan [82]. In particular, there seems to be a need for more asymptotic analysis. Firstly, we need to improve our insight into the heavy-traffic behavior of the class of branching-type polling systems, possibly based on the theory of multi-type branching processes, see Van der Mei [132], but it is even more important to develop a methodology to study the heavy-traffic behavior of those polling systems in which the branching property is violated. Secondly, large-deviation asymptotics for polling systems deserve attention. Finally, we have only begun to understand the asymptotics for n, the number of queues, growing large. The scaling limits which are developed via asymptotic analysis may subsequently provide the basis for useful approximations, see Bertsimas and Mourtzinou [16] and Boon *et al.* [27].
(ii) Relatively few studies have been devoted to the dynamic optimization of polling systems: which queue to serve next, and for how long? From an application perspective, it seems important to develop a methodology, possibly based on Markov decision processes, to tackle such problems systematically, also covering scenarios with simultaneous service of several queues subject to certain constraints.

Acknowledgment. The authors gratefully acknowledge several discussions with Marko Boon and Jacques Resing. The research was financially supported by the NWO Gravitation NET-

WORKS Grant 024.002.003 (both authors), NWO TOP-GO grant 623.001.012 (Sem Borst), and NWO TOP-C1 grant 613.001.352 (Onno Boxma).

References

[1] M.A. Abidini, O.J. Boxma, B. Kim, J. Kim and J.A.C. Resing (2017). Performance analysis of polling systems with retrials and glue periods. *Queueing Systems*, **87**, 293–324.

[2] M.A. Abidini, J.L. Dorsman and J.A.C. Resing (2017). Heavy traffic analysis of a polling model with retrials and glue periods. Submitted for publication.

[3] I.J.B.F. Adan, O.J. Boxma, S. Kapodistria and V. Kulkarni (2016). The shorter queue polling model. *Ann. Oper. Res.*, **241**, 167–200.

[4] E. Altman and D. Fiems (2007). Expected waiting time in symmetric polling systems with correlated walking times. *Queueing Systems*, **56**, 241–253.

[5] E. Altman and S. Foss (1997). Polling on a space with general arrival and service time distribution. *Oper. Res. Lett.*, **20**, 187–194.

[6] E. Altman, A. Khamisy and U. Yechiali (1992). On elevator polling with globally gated regime. *Queueing Systems* **11**, *Special Issue on Polling Models*, 85–90.

[7] E. Altman and U. Yechiali (1994). Closed polling systems. *Prob. Eng. Inf. Sci.*, **18**, 327–343.

[8] N. Antunes, C. Fricker, Ph. Robert and J.W. Roberts (2010). Traffic capacity of large WDM passive optical networks. In: *Proc. ITC 22.*

[9] M. Armony and N. Bambos (2003). Queueing dynamics and maximal throughput scheduling in switched processing systems. *Queueing Systems*, **44**, 209–252.

[10] R. Armony and U. Yechiali (1999). Polling systems with permanent and transient customers. *Stochastic Models*, **15**, 395–427.

[11] K.B. Athreya and P.E. Ney (1972). *Branching Processes* (Springer, Berlin).

[12] K. Avrachenkov, E. Perel and U. Yechiali (2016). Finite-buffer polling systems with threshold-based switching policy. *TOP*, **24 (3)**, 541–571.

[13] U. Ayesta, O.J. Boxma and I.M. Verloop (2012). Sojourn times in a processor sharing queue with multiple vacations. *Queueing Systems*, **71**, 53–78.

[14] P. Beekhuizen and J.A.C. Resing (2008). Reduction of a polling network to a single node. *Queueing Systems*, **58**, 303–319.

[15] R. Bekker, P. Vis, J.L. Dorsman, R.D. van der Mei and E.M.M. Winands (2015). The impact of scheduling policies on the waiting-time distributions in polling systems. *Queueing Systems*, **79**, 145–172.

[16] D. Bertsimas and G. Mourtzinou (2009). Multiclass queueing systems in heavy traffic: an asymptotic approach based on distributional and conservation laws. *Oper. Res.*, **45**, 470–487.

[17] J.P.C. Blanc (1991). The power-series algorithm applied to cyclic polling systems. *Stochastic Models*, **7**, 527–545.

[18] J.P.C. Blanc and R.D. van der Mei (1995). Optimization of polling systems with Bernoulli schedules. *Perf. Eval.*, **21**, 139–158.

[19] M.A.A. Boon and I.J.B.F. Adan (2009). Mixed gated/exhaustive service in a polling model with priorities. *Queueing Systems*, **63**, 383–399.

[20] M.A.A. Boon, I.J.B.F. Adan and O.J. Boxma (2010). A two-queue polling model with two priority levels in the first queue. *Discrete Event Dyn. Syst.*, **20**, 511–536.

[21] M.A.A. Boon, I.J.B.F. Adan and O.J. Boxma (2010). A polling model with multiple priority levels. *Perf. Eval.*, **67**, 468–484.

[22] M.A.A. Boon, O.J. Boxma, O. Kella and M. Miyazawa (2017). Queue-length balance equations in multiclass multiserver queues and their generalizations. *Queueing Systems*, **86**, 277–299.

[23] M.A.A. Boon, O.J. Boxma and E.M.M. Winands (2011). On open problems in polling systems. *Queueing Systems*, **68 (3)**, 365–374.

[24] M.A.A. Boon, R.D. van der Mei and E.M.M. Winands (2011). Applications of polling systems. *Surveys Oper. Res. Mgmt. Sci.*, **16**, 67–82.

[25] M.A.A. Boon, R.D. van der Mei and E.M.M. Winands (2016). Heavy traffic analysis of roving server networks. *Stochastic Models*, **33**, 171–209.

[26] M.A.A. Boon and E.M.M. Winands (2014). Heavy-traffic analysis of k-limited polling systems. *Prob. Eng. Inf. Sci.*, **28**, 451–471.

[27] M.A.A. Boon, E.M.M. Winands, I.J.B.F. Adan and A.C. van Wijk (2014). Closed-form waiting time approximations for polling systems. *Perf. Eval.*, **68**, 290–306.

[28] S.C. Borst (1995). Polling systems with multiple coupled servers. *Queueing Systems*, **20**, 369–393.

[29] S.C. Borst and O.J. Boxma (1997). Polling models with and without switchover times. *Oper. Res.*, **45**, 536–543.

[30] S.C. Borst, O.J. Boxma, J.H.A. Harink and G.B. Huitema (1994). Optimization of fixed time polling schemes. *Telecommunication Systems*, **3**, 31–59.

[31] S.C. Borst, O.J. Boxma and H. Levy (1995). The use of service limits for efficient operation of multi-station single-medium communication systems. *IEEE/ACM Trans. Netw.*, **3**, 602–612.

[32] S.C. Borst and R.D. van der Mei (1998). Waiting-time approximations for multiple-server polling systems. *Perf. Eval.*, **31**, 163–182.

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

[34] O.J. Boxma (1991). Analysis and optimization of polling systems. In: *Queueing, Performance and Control in ATM*, eds. J.W. Cohen and C.D. Pack (North-Holland, Amsterdam), 173–183.

[35] O.J. Boxma, J. Bruin and B. Fralix (2009). Sojourn times in polling systems with various service disciplines. *Perf. Eval.*, **66**, 621–639.

[36] O.J. Boxma and D. Down (1997). Dynamic server assignment in a two-queue model. *Eur. J. Oper. Res.*, **103**, 595–609.

[37] O.J. Boxma and W.P. Groenendijk (1987). Pseudo-conservation laws in cyclic service systems. *J. Appl. Prob.* **24**, 949–964.

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

[39] O.J. Boxma, J. Ivanovs, K. Kosinski and M.R.H. Mandjes (2011). Lévy-driven polling systems and continuous-state branching processes. *Stochastic Systems*, **1**, 411–436.

[40] O.J. Boxma and O. Kella (2014). Decomposition results for stochastic storage processes and queues with alternating Lévy inputs. *Queueing Systems*, **77**, 97–112.

[41] O.J. Boxma, O. Kella and K.M. Kosinski (2011). Queue lengths and workloads in polling systems. *Oper. Res. Lett.*, **39**, 401–405.

[42] O.J. Boxma, G.M. Koole and I. Mitrani (1995). Polling models with threshold switching. In: *Quantitaive Methos in Parallel Systems*, F. Baccelli, A. Jean-Marie and I. Mitrani (eds.) (Springer, Berlin), 129–140.

[43] O.J. Boxma, H. Levy and J.A. Weststrate (1990). Optimization of polling systems. In: *Proc. Performance '90*, eds. P.J.B. King, I. Mitrani, R.J. Pooley (North-Holland, Amsterdam), 349–361.

[44] O.J. Boxma, H. Levy and J.A. Weststrate (1993). Efficient visit orders for polling systems. *Perf. Eval.*, **18**, 103–123.

[45] O.J. Boxma, H. Levy and U. Yechiali (1992). Cyclic reservation schemes for efficient operation of multiple-queue single-server systems. *Ann. Oper. Res.*, **35**, 187–208.

[46] O.J. Boxma and B.W. Meister (1987). Waiting-time approximations for cyclic-service systems with switchover times. *Perf. Eval.*, **7**, 299–308.

[47] O.J. Boxma and B.W. Meister (1987). Waiting-time approximations in multi-queue systems with cyclic-service. *Perf. Eval.*, **7**, 59–70.

[48] O.J. Boxma and J.A. Weststrate (1989). Waiting times in polling systems with Markovian server routing. In: *Messung, Modellierung und Bewertung von Rechensystemen und Netzen*, eds. G. Stiege, J.S. Lie (Springer, Berlin), 89–104.

[49] S. Browne, E.G. Coffman, Jr., E.N. Gilbert and P.E.W. Wright (1992). Gated, exhaustive, parallel service. *Prob. Eng. Inf. Sci.*, **6**, 217–239.

[50] S. Browne and O. Kella (1995). Parallel service with vacations. *Oper. Res.*, **43**, 870–878.

[51] S. Browne and G. Weiss (1992). Dynamic priority rules when polling with multiple parallel servers. *Oper. Res. Lett.*, **12**, 129–137.

[52] S. Browne and U. Yechiali (1989). Dynamic priority rules for cyclic-type queues. *Adv. Appl. Prob.*, **21**, 432–450.

[53] S.L. Brumelle (1971). On the relation between customer and time averages in queues. *J. Appl. Prob.*, **8**, 508–520.

[54] A. Brzezinski and E. Modiano (2005). Dynamic reconfiguration and routing algorithms for IP-over-WDM networks with stochastic traffic. *IEEE J. Lightwave Techn.*, **23**, 3188–3205.

[55] C. Buyukkoc, P. Varaiya and J. Walrand (1985). The $c\mu$ rule revisited. *Adv. Appl. Prob.*, **17**, 237–238.

[56] G. Celik, S.C. Borst, P.A. Whiting and E. Modiano (2016). Dynamic scheduling with reconfiguration delays. *Queueing Systems*, **83 (1–2)**, 87–129.

[57] K.C. Chang and D. Sandhu (1992). Mean waiting time approximations in cyclic-service systems with exhaustive limited service policy. *Perf. Eval.*, **15**, 21–40.

[58] G.L. Choudhury and W. Whitt (1996). Computing distributions and moments in polling models by numerical transform inversion. *Perf. Eval.*, **25**, 267–292.

[59] E.G. Coffman, Jr., G. Fayolle and I. Mitrani (1988). Two queues with alternating service periods. In: *Performance '87*, eds. P.-J. Courtois and G. Latouche (North-Holland, Amsterdam), 227–239.

[60] E.G. Coffman, Jr. and E.N. Gilbert (1986). A continuous polling system with constant service times. *IEEE Trans. Inf. Theory*, **33**, 584–591.

[61] E.G. Coffman, Jr., A.A. Puhalskii and M.I. Reiman (1995). Polling systems with zero switchover times: A heavy-traffic averaging principle. *Ann. Appl. Prob.*, **5**, 681–719.

[62] E.G. Coffman, Jr., A.A. Puhalskii and M.I. Reiman (1998). Polling systems in heavy traffic: A Bessel process limit. *Math. Oper. Res.*, **23**, 257–304.

[63] J.W. Cohen (1987). A two-queue model with semi-exhaustive service. In: *Performance '87*, eds. P.-J. Courtois and G. Latouche (North-Holland, Amsterdam), 19–37.

[64] J.W. Cohen and O.J. Boxma (1981). The $M/G/1$ queue with alternating service formulated as a Riemann-Hilbert boundary value problem. In: *Proc. Performance '81*, ed. F.J. Kylstra (North-Holland, Amsterdam), 181–199.

[65] R.B. Cooper (1970). Queues served in cyclic order: waiting times. *Bell Syst. Techn. J.*, **49**, 399–413.

[66] R.B. Cooper and G. Murray (1969). Queues served in cyclic order. *Bell Syst. Techn. J.*, **48**, 675–689.

[67] R.B. Cooper, S.-C. Niu and M.M. Srinivasan (1996). A decomposition theorem for polling models: the switchover times are effectively additive. *Oper. Res.*, **44**, 629–633.

[68] O. Czerniack and U. Yechiali (2009). Fluid polling systems. *Queueing Systems*, **63**, 401–435.

[69] K. Debicki and M.R.H. Mandjes (2015). *Queues and Lévy Fluctuation Theory*. Springer, New York.

[70] J.L. Dorsman, S.C. Borst, O.J. Boxma and M. Vlasiou (2015). Markovian polling systems with an application to wireless random-access networks. *Perf. Eval.*, **85**, 33–51.

[71] B.T. Doshi (1986). Queueing systems with vacations – a survey. *Queueing Systems*, **1**, 29–66.

[72] B.T. Doshi (1990). Single server queues with vacations. In: *Stochastic Analysis of Computer and Communication Systems*, ed. H. Takagi (North-Holland, Amsterdam), 417–465.

[73] M. Eisenberg (1972). Queues with periodic service and changeover times. *Oper. Res.*, **20**, 440–451.

[74] M. Eisenberg (1979). Two queues with alternating service. *SIAM J. Appl. Math.*, **36**, 287–303.

[75] M. Eisenberg (1994). The polling system with a stopping server. *Queueing Systems*, **18**, 387–431.

[76] I. Eliazar (2005). Gated polling systems with Lévy inflow and inter-dependent switchover times: a dynamical systems approach. *Queueing Systems*, **49**, 49–72.

[77] I. Eliazar (2005). From polling to snowplowing. *Queueing Systems*, **51**, 115–133.

[78] D.E. Everitt (1986). A conservation-type law for the token ring with limited service. *Br. Telecom Techn. J.*, **4**, 51–61.

[79] D. Everitt (1986). Simple approximations for token rings. *IEEE Trans. Commun.*, **34 (7)**, 719–721.

[80] D.E. Everitt (1989). An approximation procedure for cyclic service queues with limited service. In: *Performance of Parallel and Distributed Systems*, eds. T. Hasegawa, H. Takagi and Y. Takahashi (North-Holland, Amsterdam), 141–156.

[81] O. Fabian and H. Levy (1994). Pseudo-cyclic policies for multi-queue single server systems. *Ann. Oper. Res.* **48**, *Special Issue on Queueing Networks*, ed. N.M. van Dijk, 127–152.

[82] A. Federgruen and Z. Katalan (1994). Approximating queue size and waiting time distributions in general polling systems. *Queueing Systems*, **18**, 353–386.

[83] W. Feng, M. Kowada and K. Adachi (1998). A two-queue model with Bernoulli service schedule and switching times. *Queueing Systems*, **30**, 405–434.

[84] W. Feng, M. Kowada and K. Adachi (2001). Performance analysis of a two-queue model with an (M, N)-threshold service policy. *J. Oper. Res. Soc. Japan*, **44**, 101–124.

[85] M.J. Ferguson and Y.J. Aminetzah (1985). Exact results for nonsymmetric token ring systems. *IEEE Trans. Commun.*, **33**, 223–231.

[86] S.G. Foss and N.I. Chernova (1996). Dominance theorems and ergodic properties of polling systems. *Prob. Inf. Trans.*, **32 (4)**, 342–364.

[87] S.G. Foss and N.I. Chernova (1996). On polling systems with infinitely many stations. *Sib. Math. J.*, **37 (4)**, 832–846.

[88] S.G. Foss, N.I. Chernova and A. Kovalevskii (1996). Stability of polling systems with state-independent routing. In: *Proc. 34th Allerton Conf.*, Monticello Il, 220–227.

[89] S.G. Foss and A.P. Kovalevskii (1999). A stability criterion via fluid limits and its application to a polling system. *Queueing Systems*, **32 (1)**, 131–168.

[90] S. Foss and G. Last (1996). Stability of polling systems with exhaustive service policies and state-dependent routing. *Ann. Appl. Prob.*, **6**, 116–137.

[91] S.G. Foss and G. Last (1998). On the stability of greedy polling systems with general service policies. *Prob. Eng. Inf. Sci.*, **12 (1)**, 49–68.

[92] L. Fournier and Z. Rosberg (1991). Expected waiting times in cyclic service systems under priority disciplines. *Queueing Systems* **9**, 419–439.

[93] C. Fricker and M.R. Jaïbi (1994). Monotonicity and stability of periodic polling models. *Queueing Systems*, **15**, 211–238.

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

[95] S.W. Fuhrmann (1992). A decomposition result for a class of polling models. *Queueing Systems*, **11**, *Special Issue on Polling Models*, 109–120.

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

[97] S.W. Fuhrmann and R.B. Cooper (1985). Application of decomposition principle in $M/G/1$ vacation model to two continuum cyclic queueing models - especially token-ring LANs. *AT&T Techn. J.*, **64**, 1091–1099.

[98] S.W. Fuhrmann and Y.T. Wang (1988). Analysis of cyclic service systems with limited service: bounds and approximations. *Perf. Eval.*, **9**, 35–54.

[99] J.P. van der Gaast, M.B.M. de Koster and I.J.B.F. Adan (2017). The analysis of batch sojourn-times in polling systems. *Queueing Systems*, **85**, 313–335.

[100] B. Gamse and G.F. Newell (1982). An analysis of elevator operation in moderate height buildings - I. A single elevator. *Transp. Res. B*, **16**, 303–319.

[101] B. Gamse and G.F. Newell (1982). An analysis of elevator operation in moderate height buildings - II. Multiple elevators. *Transp. Res. B*, **16**, 321–335.

[102] 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), 659–698.

[103] W.P. Groenendijk (1990). *Conservation Laws in Polling Systems.* Ph.D. Thesis, University of Utrecht, Utrecht.

[104] R. de Haan, R.J. Boucherie and J.-K. van Ommeren (2009). A polling model with an autonomous server. *Queueing Systems*, **62**, 279–308.

[105] D.P. Heyman and M.J. Sobel (1982). *Stochastic Models in Operations Research, Vol. I.* McGraw-Hill Book Company, New York.

[106] M. Hofri and K.W. Ross (1987). On the optimal control of two queues with server set-up times and its analysis. *SIAM J. Comput.*, **16**, 399–420.

[107] B. Van Houdt (2010). Numerical solution of polling systems for analyzing networks on chips. In: *Proc. of NSMC.*

[108] Y.-C. Hung and C.-C. Chang (2008). Dynamic scheduling for switched processing systems with substantial service-mode switching times. *Queueing Systems*, **60**, 87–109.

[109] O.C. Ibe (1990). Analysis of polling systems with mixed service disciplines. *Stochastic Models*, **6**, 667–689.

[110] J. Ivanovs and O. Kella (2013). Another look into decomposition results. *Queueing Systems*, **75**, 19–28.

[111] O.B. Jennings (2010). Averaging principles for a diffusion-scaled, heavy-traffic polling station with K job classes. *Math. Oper. Res.*, **35**, 669–703.

[112] M. Jirina (1958). Stochastic branching processes with continuous state space. *Czechosl. Math. J.*, **8**, 292–313.

[113] V. Kavitha and E. Altman (2012). Continuous polling models and application to ferry assisted WLAN. *Annals of Operations Research*, **198**, 185–218.

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

[115] B. Kim and J. Kim (2017). Sojourn time distribution in polling systems with processor-sharing policy. *Perf. Eval.*, **114**, 97–112.

[116] D.E. Knuth (1973). *The Art of Computer Programming, Vol. III.* Addison-Wesley Publ. Cy., Massachusetts.

[117] A.G. Konheim and H. Levy (1992). Efficient analysis of polling systems. In: *Proc. INFOCOM '92*, 2325–2331.

[118] G.M. Koole (1998). Assigning a single server to inhomogeneous queues with switching costs. *Theor. Comp. Science*, **182**, 203–216.

[119] A.P. Kovalevskii, V.A. Topchii and S.G. Foss (2005). On the stability of a queueing system with uncountably branching fluid limits. *Prob. Inf. Trans.*, **41 (3)**, 254–279 (Rus: 76–104).

[120] D.P. Kroese and V. Schmidt (1994). Single-server queues with spatially distributed arrivals. *Queueing Systems*, **17**, 317–345.

[121] D.P. Kroese (1997). Heavy traffic analysis for continuous polling models. *J. Appl. Prob.*, **34**, 720–732.

[122] D.P. Kroese and V. Schmidt (1992). A continuous polling system with general service times. *Ann. Appl. Prob.*, **2**, 906–927.

[123] J.B. Kruskal (1969). Work-scheduling algorithms: a non-probabilistic queueing study (with possible applications to No. 1 ESS). *Bell Syst. Techn. J.*, **48**, 2963–2974.

[124] D.-S. Lee (1997). Analysis of a two-queue model with Bernoulli schedules. *J. Appl. Prob.*, **34**, 176–191.

[125] D.-S. Lee and B. Sengupta (1993). Queueing analysis of a threshold based priority scheme for ATM networks. *IEEE/ACM Trans. Netw.*, **1**, 709–717.

[126] K.K. Leung (1991). Cyclic-service systems with probabilistically-limited service. *IEEE J. Sel. Areas Commun.*, **9 (2)**, 185–193.

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

[128] H. Levy and M. Sidi (1989). Polling systems with correlated arrivals. In: *Proc. IEEE INFOCOM '89*, 907–913.

[129] H. Levy and M. Sidi (1990). Polling models: applications, modeling and optimization. *IEEE Trans. Commun.*, **38**, 1750–1760.

[130] H. Levy and M. Sidi (1991). Polling systems with simultaneous arrivals. *IEEE Trans. Commun.*, **39**, 823–827.

[131] D.M. Markowitz, M.I. Reiman and L.M. Wein (2000). The stochastic economic lot scheduling problem: Heavy traffic analysis of dynamic cyclic policies. *Oper. Res.*, **48**, 136–154.

[132] R.D. van der Mei (2007). Towards a unifying theory on branching-type polling systems in heavy traffic. *Queueing Systems*, **57**, 29–46.

[133] R.D. van der Mei and S.C. Borst (1997). Analysis of multiple-server polling systems by means of the power-series algorithm. *Stochastic Models*, **13**, 339–369.

[134] R.D. van der Mei and E.M.M. Winands (2008). A note on polling models with renewal arrivals and nonzero switch-over times. *Oper. Res. Lett.*, **36**, 500–505.

[135] I. Meilijson and U. Yechiali (1977). On optimal right-of-way policies at a single-server station when insertion of idle times is permitted. *Stoch. Proc. Appl.*, **6**, 25–32.

[136] T.M.M. Meyfroyt, M.A.A. Boon, S.C. Borst and O.J. Boxma (2018). Performance of large-scale polling systems with branching-type and limited service. Eurandom Report 2018-003; submitted for publication.

[137] R.J.T. Morris and Y.T. Wang (1984). Some results for multi-queue systems with multiple cyclic servers. In: *Performance of Computer-Communication Systems*, eds. W. Bux, H. Rudin (North-Holland, Amsterdam), 245–258.

[138] T.L. Olsen (2001). Approximations for the waiting time distribution in polling models with and without state-dependent setups. *Oper. Res. Lett.*, **28**, 113–123.

[139] T.L. Olsen and R.D. van der Mei (2005). Polling systems with periodic server routeing in heavy traffic: Renewal arrivals. *Oper. Res. Lett.*, **33**, 17–25.

[140] T. Ozawa (1990). Alternating service queues with mixed exhaustive and k-limited services. *Perf. Eval.*, **11**, 165–175.

[141] M.P. Quine (1972). The multitype Galton-Watson process with ρ near 1. *Adv. Appl. Prob.*, **4**, 429–452.

[142] M.I. Reiman and L.M. Wein (1999). Heavy traffic analysis of polling systems in tandem. *Oper. Res.*, **47**, 524–534.

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

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

[145] Ph. Robert and J.W. Roberts (2010). A mean field approximation for the capacity of server-limited, gate-limited multi-server polling systems. *ACM SIGMETRICS Perf. Eval. Rev.*, **38**, 24–26.

[146] L. Rojas Nandayapa, S.G. Foss and D.P. Kroese (2011). Stability and performance of greedy server systems: A review and open problems. *Queueing Systems*, **68**, 221–227.

[147] Z. Saffer and M. Telek (2010). Unified analysis of $BMAP/G/1$ cyclic polling models. *Queueing Systems*, **64**, 69–102.

[148] G. Shapira and H. Levy (2015). On fairness in polling systems. Technical Report, Tel-Aviv University; to appear in *Ann. Oper. Res.*

[149] S. Shimogawa and Y. Takahashi (1992). A note on the pseudo-conservation law for a multi-queue with local priority. *Queueing Systems* **11**, *Special Issue on Polling Models*, 145–151.

[150] M. Sidi, H. Levy and S.W. Fuhrmann (1992). A queueing network with a single cyclically roving server. *Queueing Systems* **11**, *Special Issue on Polling Models*, 121–144.

[151] M.M. Srinivasan, S.-C. Niu and R.B. Cooper (1995). Relating polling models with nonzero and zero switchover times. *Queueing Systems*, **19**, 149–168.

[152] H. Takagi (1991). Application of polling models to computer networks. *Comput. Netw. ISDN Syst.*, **22**, 193–211.

[153] H. Takagi (1991). *Queueing Analysis. Vol. 1: Vacation and Priority Systems, Part 1.* North-Holland, Amsterdam.

[154] H. Takagi (1997). Queueing analysis of polling models: progress in 1990-1994. In: J.H. Dshalalow, editor, *Frontiers in Queueing: Models, Methods and Problems*, ed. J.H. Dshalalow (CRC Press, Boca Raton), 119–146.

[155] H. Takagi (2000). Analysis and application of polling models. In: *Performance Evaluation: Origins and Directions*, LNCS 1769, eds. G. Haring, C. Lindemann and M. Reiser (Springer, Berlin), 424–442.

[156] N. Tian and Z.G. Zhang (2006). *Vacation Queueing Models.* Springer, New York.

[157] P. Tran-Gia (1992). Analysis of polling systems with general input process and finite capacity. *IEEE Trans. Commun.*, **40**, 337–344.

[158] P. Vis, R. Bekker and R.D. van der Mei (2015). Heavy-traffic limits for polling models with exhaustive service and non-FCFS service order policies. *Adv. Appl. Prob.*, **47**, 989–1014.

[159] V.M. Vishnevskii and O.V. Semenova (2006). Mathematical methods to study the polling systems. *Autom. Remote Control*, **67**, 173–220.

[160] M. Vlasiou and U. Yechiali (2008). M/G/∞ polling systems with random visit times. *Prob. Eng. Inf. Sci.*, **22 (1)**, 81–106.

[161] C.-H. Wang and T. Javidi (2017). Adaptive policies for scheduling with reconfiguration delay: an end-to-end solution for all-optical data centers. *IEEE/ACM Trans. Netw.* **25 (3)**, 1555–1568.

[162] J.A. Weststrate and R.D. van der Mei (1994). Waiting times in a two-queue model with exhaustive and Bernoulli service. *Zeitschr. für Oper. Res.*, **40**, 289–303.

[163] A. Wierman, E.M.M. Winands and O.J. Boxma (2007). Scheduling in polling systems. *Perf. Eval.*, **64**, 1009–1028.

[164] A.C.C. van Wijk, I.J.B.F. Adan, O.J. Boxma and A. Wierman (2012). Fairness and efficiency for polling systems with the κ-gated service discipline. *Perf. Eval.*, **69**, 274–288.

[165] E.M.M. Winands (2007). *Polling, Production & Priorities*. PhD Thesis, Eindhoven University of Technology.

[166] E.M.M. Winands, I.J.B.F. Adan and G.J. van Houtum (2006). Mean value analysis for polling systems. *Queueing Systems*, **54**, 45–54.

[167] E.M.M. Winands, I.J.B.F. Adan, G.J. van Houtum and D.G. Down (2009). A state-dependent polling model with k-limited service. *Prob. Eng. Inf. Sci.*, **23**, 385–408.

[168] U. Yechiali (1991). Optimal dynamic control of polling systems. In: *Queueing, Performance and Control in ATM*, eds. J.W. Cohen, C.D. Pack (North-Holland, Amsterdam), 205–217.