

Applications of polling systems

M.A.A. Boon*
marko@win.tue.nl

R.D. van der Mei ^{† ‡}
mei@cw.nl

E.M.M. Winands [†]
emm.winands@few.vu.nl

February 24, 2011

Abstract

Since the first paper on polling systems, written by Mack in 1957, a huge number of papers on this topic has been written. A typical polling system consists of a number of queues, attended by a single server. In several surveys, the most notable ones written by Takagi, detailed and comprehensive descriptions of the mathematical analysis of polling systems are provided. The goal of the present survey paper is to complement these papers by putting the emphasis on *applications* of polling models. We discuss not only the capabilities, but also the limitations of polling models in representing various applications. The present survey is directed at both academicians and practitioners.

Keywords: Polling systems, applications, survey

1 Introduction

A typical *polling system* consists of a number of queues, attended by a single server in a fixed order. There is a huge body of literature on polling systems that has developed since the late 1950s, when the papers of Mack et al. [117, 118] concerning a patrolling repairman model for the British cotton industry were published. The term “polling” originates from the so-called polling data link control scheme, in which a central computer (server) interrogates each terminal (queue) on a communication line to find whether it has any information (customers) to transmit. The addressed terminal transmits information and the computer then switches to the next terminal to check whether that terminal has any information to transmit. In a broader perspective, polling models are applicable in situations in which several types of users compete for access to a common resource which is available to only one type of user at a time. The ubiquity of polling systems can be observed in many applications, e.g., in computer-communication, production, transportation and maintenance systems. The present paper discusses the main application areas of polling systems along with an extensive list of references, examines how these various applications can be represented and analyzed via

*EURANDOM and Department of Mathematics and Computer Science, Eindhoven University of Technology, P.O. Box 513, 5600MB Eindhoven, The Netherlands

[†]Department of Mathematics, Section Stochastics, VU University, De Boelelaan 1081a, 1081HV Amsterdam, The Netherlands

[‡]Centre for Mathematics and Computer Science (CWI), Department of Probability and Stochastic Networks, 1098 SJ Amsterdam, The Netherlands

polling models and outlines several topics for future research inside and outside the described areas.

To illustrate the fact that polling systems have become a powerful tool for the performance analysis of a wide variety of important applications, we would like to cite Hideaki Takagi, one of the fathers of the success of polling models [152]: *The analysis of polling models gained momentum as queueing systems that are easy to understand, analyze, and extend. The study has been accelerated largely by applications to the modeling of communication, manufacturing and transportation systems. I believe that it is one of the few successful theoretical performance evaluation models developed in the last decades.* The fact that nowadays polling systems are still fully alive can be illustrated by observing that the recent application-oriented conferences Performance 2007, Informatics Applied Probability 2007 and 2009, and ValueTools 2008 all scheduled a dedicated session for polling systems. Very recently, Annals of Operations Research dedicated a special issue to polling systems.

Until the mid nineties, Takagi maintained a fairly complete bibliography on polling models, which contained over 700 publications, including journal and conference papers, books, theses, and technical reports. Since research on polling models has continued in the past years (although not at as high a pace as in the years around 1990), the number of papers might now well be above 1000. The main reason for this tidal wave of papers on polling models in the past 50 years can be found in the diversity of applications, which have led to innumerable enhancements of the basic cyclical polling system. Such enhancements gave rise to interesting new queueing theoretic problems which in turn stimulated research in various directions.

In this respect one of the most remarkable mathematical results in the polling literature is the striking dichotomy in the complexity of the analysis of polling systems, which has been independently illuminated by Fuhrmann [74] and Resing [135] via the use of a multitype branching approach. That is, if the service discipline (i.e., the criterion which determines how many customers are served during a visit of the server to a queue) satisfies a certain branching property, in many cases the polling system allows for an exact analysis by rather standard methods. If this branching property is, however, violated, the corresponding polling system defies an exact analysis except for some special (two-queue or symmetric) cases.

The most important members of the class of policies satisfying the branching property are the *exhaustive* service discipline, under which the server continues to work until the queue has become empty, the *gated* service discipline, under which exactly those customers are served who were present at the queue at the beginning of the visit, and *globally gated* service, which states that only those customers are served that were present at the server's arrival epoch at some predetermined "parent queue". (Actually the globally gated service discipline does not strictly satisfy the branching property as defined in [135], but a weaker variant that still allows an exact analysis.) Unfortunately, the *k-limited* service policy, under which the server continues working at a queue until either a predefined number of k customers is served or until the queue becomes empty, is on the wrong side of the borderline implying that even mean queue lengths are in general not known.

In the past various survey papers dedicated to polling systems have appeared in the open literature (see, e.g., [115, 149, 151, 152, 168, 182]) which give detailed and comprehensive descriptions of the mathematical analysis of polling systems. Although we present a brief overview of the basic mathematical analysis techniques, the main goal of the present survey paper is to complement these papers by putting the emphasis on *applications* of polling

models. We unearth not only the capabilities, but also the limitations of polling models in representing various applications. Moreover, since the publication of the previous survey papers around 10 till 15 years ago, a large number of papers on polling models have been published. The present survey is directed at both academicians and practitioners. The list of references is relatively recent, but deliberately far from exhaustive. The reader who is interested in going deeper into specific mathematical details of polling systems will find references to the aforementioned survey papers aimed at that objective. In selecting the literature to cite in this survey, a preference has been given to current papers and research on applications of polling systems.

The rest of the present paper is organized as follows. Section 2 gives a global overview of various aspects of polling systems, and gives an illustration of how one can analyze a typical polling system. In Section 3 we describe the main applications of polling systems in communication networks, production systems, traffic and transportation problems and various other fields of engineering. Finally, the last section indicates some possible directions for further research.

2 Polling systems

A polling model consists of a number of queues, attended by a single server who visits the queues in some order to render service to the customers waiting at the queues, typically incurring some switch-over time while moving from one queue to the next. In Section 2.1 we give a structural overview of the variety of polling models considered in the literature, successively discussing model variants with respect to the arrival process, the buffer size, the service process, the switch-over process, the server routing, the service discipline and the queueing discipline. In Section 2.2 we briefly illustrate how polling systems may be analyzed.

2.1 Typical features

Arrival process. In most polling models, customers arrive at the queues according to mutually independent homogeneous single Poisson arrival processes. In many cases arrival processes can indeed be described adequately by Poisson processes, e.g., in the case of telephone calls and traffic accidents. Moreover, due to the memorylessness property of the Poisson process, polling models with Poisson arrivals often allow for a tractable analysis.

Yet, in many situations the assumption of Poisson arrivals is unrealistic, e.g., in case of bursty traffic (voice, video) and several production applications. Only very recently, variations of Poisson arrival processes have been considered. Van Vuuren and Winands [167] have presented an approximate algorithm for polling systems with k -limited service and Markovian Arrival Processes (MAPs). Boxma et al. [32] consider polling systems with Lévy-driven input. Bertsimas and Mourtzinou [15] study polling systems with Mixed General Erlang arrival processes. Slightly more general arrival processes are discussed by Saffer and Telek [136] who consider BMAPs (Batch Markov Arrival Processes). Some limiting cases lead to closed-form expressions for, e.g., queue lengths or waiting time distributions, even under the assumption of general renewal arrivals. For example, Van der Mei and Winands [162, 163] derive the distribution of the scaled waiting time under heavy traffic conditions, i.e., when the load tends to one. For switch-over times tending to infinity, the scaled waiting time distribution

is conjectured in [175], but the proof is still an open problem. Closed-form approximations for waiting times in polling systems with general renewal arrivals are constructed in [24, 57] using an interpolation between light-traffic and heavy-traffic limits.

In polling models it is almost exclusively assumed that customers arrive from some external infinite population. In some applications this assumption is unrealistic, e.g., in maintenance environments where customers represent machine breakdowns. Altman and Yechiali [7] study a closed polling model in which customers, after having received service at a queue, proceed to another (possibly the same) queue. This model is extended in [8] where a model with transient as well as permanent customers is considered. Sidi and Levy [143] consider a system in which (external) customers arrive at a queue according to a Poisson process, and in which a customer, after having received service at a queue, either leaves the system or moves to another queue (with some given probability). Applications of this type of customer routing can be found, for example, in manufacturing when products undergo service in a number of stages or in the context of rework. Finally, we mention work of Levy and Sidi [116] in which they study a polling model with simultaneous arrivals at the queues.

A final remark regarding arrival processes, is that some papers study discrete-time arrival processes. For some applications, it is natural to divide time into slots. In communication systems discrete-time models are used by, e.g., Kleinrock and Scholl [99] for the analysis of the Minislotted Alternating Priorities (MSAP) multiple-access scheme, by Kleinrock and Levy [98] for the analysis of polling systems with random server routing to model the Slotted ALOHA system, and by Beekhuizen [11] to model networks on chips. An example in a completely different application area, is provided by Van Leeuwen [166] who analyses the Fixed Cycle Traffic Light queue in discrete time.

Buffer size. In most polling models the buffer size is assumed to be infinite. In a number of applications the buffer capacities are obviously finite, e.g., transportation and manufacturing. In some applications it is natural to assume that the buffers are unit sized, i.e., each queue can only accommodate one customer at a time. Applications of polling models with unit-sized buffers are found, e.g., in the area of maintenance. If one would like to study a multi-item production system which produces products in a make-to-order fashion, a polling system with finite buffers appears naturally as well. For an analysis of polling models with finite buffers, see, e.g., Takagi [150], or Browne and Yechiali [39].

Service process. The service times at a queue are typically assumed to be samples from a probability distribution which is characteristic for that queue. The service times are usually assumed to be mutually independent and independent of the actual state of the system. Polling models with batch service are studied in [164] and [35]. Van der Wal and Yechiali [164] study a model where, at the beginning of each cycle, the server determines a tour along each of the non-empty queues. This model, with customers being served in one batch, has applications to videotex, telex, and time-division multiple access (TDMA) systems, as well as for central database systems and Automated Guided Vehicles (AGVs). Boxma et al. [35] study a polling model with batch service. Server routing is either strictly cyclic, or the server visits the queue with the oldest customer first.

The usual assumption in the polling literature is that a *single* server is visiting all the queues. Only few results have been obtained on *multiple*-server polling systems. For example, some

numerical results are obtained by Ajmone Marsan et al. [1], who perform several numerical studies using generalized stochastic Petri nets (GSPNs). Analytical approximative results have been obtained by Morris and Wang [125], and also in [29, 157]. Morris and Wang observe the interesting phenomenon that the servers tend to cluster if they follow identical routes, deteriorating the system performance. Because the methods used in all of these papers are based on approximations, many challenges remain for future research. However, some interesting exact results are found for a polling system with multiple *coupled* servers, visiting the queues always together [25].

Switch-over process. The switch-over times needed by the server to move from one queue to another queue are typically assumed to be samples from some prespecified probability distribution which is characteristic for that couple of queues. In a few cases the switch-over times are assumed to be decomposed into switch-out times (which are characteristic for the queue being departed from) and switch-in times (which are characteristic for the queue being switched to), putting a special structure on the switch-over times.

Nearly all of the existing literature on polling systems makes the assumption of state-independent switch-overs, i.e., switch-overs are assumed to be independent of the current state of the system. Notable exceptions are the recent studies of Altman et al. [3], Günalay and Gupta [86], Gupta and Srinivasan [87], Singh and Srinivasan [144] and Winands et al. [177]. The choice of modeling state-independent setups is generally not motivated by an application but by the tractability of the resulting analysis.

Altman and Fiems [4] allow correlation between switch-over times in the polling model. That is, they assume that the switch-over times constitute a stationary ergodic series of random variables. A wireless LAN, where an access point polls mobiles, is one of the applications of this type of switch-over process.

Server routing. The order in which the server visits the queues is determined by some routing mechanism. Such a mechanism may depend on the actual state of the system (dynamic) or may be independent of the state of the system (static).

We first discuss *static* routing mechanisms. The traditional routing mechanism is the cyclic server routing. To model systems in which particular queues are visited more frequently than others, cyclic polling has been extended to periodic polling, in which the server visits the queue periodically according to some service order table of finite length, cf. [9]. Alternatively, the server may be routed along the queues according to some probabilistic routing mechanism. Kleinrock and Levy [98] introduce the so-called random polling mechanism in which, after a departure from a queue, the server proceeds to queue j with some given probability p_j .

Boxma and Weststrate [36] and Srinivasan [146] generalize the random polling scheme to the so-called Markovian routing scheme in which the server, after a departure of the server from queue i , proceeds to queue j with given probability $p_{i,j}$.

Under *dynamic* server routing, the decision of the server as to the order in which the queues are visited may depend on a certain amount of information available to the server, such as the queue lengths. For instance, it might not make sense to move to an empty queue while customers are waiting at other queues (cf., e.g., [91]). An example of dynamic server routing in systems with full information about the buffer contents is the 'serve-longest-queue' policy.

Yechiali [181] has introduced the so-called *semi-dynamic* server routing, in which the server, at the end of each tour along the queues, receives information about the queue lengths. Based on this information, the server makes a decision as to the order in which it will visit the queues in the next tour. This order cannot be changed during the course of that visit tour.

Service discipline. The service discipline specifies the number of customers that is served during one visit of the server to a queue. In the literature, a whole abundance of service disciplines has been proposed by putting some cut-off mechanism (which eventually depends on the evolution of the queue length during the server visit) on the classical exhaustive and gated service disciplines. The service disciplines can be classified into the class of *customer-limited* service disciplines, in which restrictions are put on the number of customers served during a visit of the server to a queue, and the class of *time-limited* service strategies, putting restrictions on the amount of time spent by the server during one visit of the server to a queue.

Alternatively, service disciplines can be classified into the so-called *exhaustive-type* policies and the *gated-type* policies, depending on whether customers who arrive at a queue while the server is working at that queue are candidates for service during the same visit of the server to that queue, e.g., [115]. Under an exhaustive-type policy the customers arriving at a queue in service are *candidates* for service in the same visit period, whereas under a gated-type policy they are not. The best known variant of the gated service discipline is *globally gated*, which also allows for an exact analysis [34]. This service discipline states that only those customers are served who were present at the beginning of a visit to one fixed “parent-queue”. Numerous hybrid variants of service disciplines can be conceived by combining the various types of cut-off mechanisms. A number of these service disciplines has been studied in the literature, such as customer-limited type service policies like the k -limited service, Bernoulli service [94], probabilistically-limited service [112], binomial-gated service [114], binomial-exhaustive, Bernoulli-gated and Bernoulli-exhaustive service [38], fractional-exhaustive service [113], mixed gated/exhaustive service based on customer priority levels [20], Bernoulli-type service [135], and time-limited service disciplines with exponential time limits [44, 111], and with constant time limits [55]. A gated-type service discipline with multiple gates is studied in [159–161].

Queueing discipline. The queueing discipline specifies the order in which the customers present at the same queue are served. The most common queueing discipline is the classical *First-Come-First-Served* (FCFS) discipline. Very recently, Wierman et al. [173] showed that optimization of the queueing discipline in a polling system can improve the performance of the system significantly without having to purchase additional resources. Other recent papers in this direction are, for example, [21, 22, 30]. It is important to note that the queue length distribution is independent of the service order, and so are, by Little’s law, the mean waiting times (provided the queueing discipline does not depend on the service times). However, the distribution of the waiting time does depend on the queueing discipline.

2.2 Mathematical analysis

In this subsection we give a brief impression of how polling systems can be analyzed. As stated in the introduction, the present paper surveys the application areas of polling systems, where it is not our intention to present an encyclopedic overview of all available mathematical results. Consequently, we have chosen to discuss only one (frequently used) analysis method, studying the marginal queue length and waiting time distributions. The reader should bear in mind that several alternatives exist. For detailed and comprehensive surveys of the analysis methodology of polling systems, we refer to Takagi's surveys [149, 151, 152], and furthermore to [115, 168, 182]. Moreover, Takagi's seminal book "Analysis of polling systems" [148] should definitely be mentioned here.

The model studied in the present subsection can be considered as the basic, standard polling model. We assume that the polling model consists of N queues ($N \geq 1$), denoted by $Q_1, \dots, Q_N$, attended by one server in cyclic order. Whenever the server switches from Q_i to Q_{i+1} , a switch-over time S_i is incurred ($i = 1, \dots, N$; since the system is cyclic, Q_{N+1} actually refers to Q_1 etc.). The switch-over times are independent random variables. In [28] an adaptation of the analysis is given for polling models without switch-over times. The arrival processes to the different queues are independent Poisson processes with intensities λ_i , $i = 1, \dots, N$. The service times of customers arriving in Q_i are independent random variables denoted by B_i . At the end of this section we will briefly mention some alternative techniques that may be required if the polling model of interest does not satisfy all of these assumptions.

The individual loads of the queues are denoted by $\rho_i = \lambda_i \mathbb{E}[B_i]$, for $i = 1, \dots, N$, and the total load of the system is denoted by $\rho = \sum_{i=1}^N \rho_i$. The Laplace-Stieltjes transform (LST) of a continuous, positive random variable X is denoted by $\tilde{X}(\cdot)$. We use the same notation for the probability generating function (PGF) in case X is a discrete random variable.

Cycle times, visit times and intervisit times. The distribution of the length of one cycle depends on its starting point. For this reason we denote by C_i the time between two successive visit *beginnings* to Q_i , and by C_i^* the time between two consecutive visit *completions* to Q_i . The *intervisit time* of Q_i , denoted by I_i , is the time between the visit completion of Q_i and the next visit beginning to this queue. The *mean* cycle time and the *mean* visit times can easily be computed in a cyclic polling system. The first obvious, but important, observation is that the mean cycle time does *not* depend on the starting point of the cycle. For this reason, we denote the mean cycle time by $\mathbb{E}[C] = \mathbb{E}[C_i] = \mathbb{E}[C_i^*]$ for all $i = 1, \dots, N$. Since each customer arriving in the system gets served eventually, the server must be working a fraction ρ of the time, and is switching between queues a fraction $1 - \rho$ of the time. This leads to the simple relation $\mathbb{E}[S] = (1 - \rho)\mathbb{E}[C]$, where $S = \sum_{i=1}^N S_i$. A similar way of reasoning can be applied to the workload of the individual queues, yielding expressions for the mean visit and intervisit times.

$$\begin{aligned}\mathbb{E}[C] &= \frac{\mathbb{E}[S]}{1 - \rho}, \\ \mathbb{E}[V_i] &= \rho_i \mathbb{E}[C], \\ \mathbb{E}[I_i] &= (1 - \rho_i) \mathbb{E}[C].\end{aligned}$$

The stability condition of the system depends on the service disciplines in the various queues. For exhaustive, gated, and globally gated service (and many other disciplines as well) the

stability condition is simply $\rho < 1$, yielding $\mathbb{E}[C] < \infty$. Note that for these service disciplines the switch-over times do not play a role in the stability condition, as they become negligible compared to the visit times, as $\rho \rightarrow 1$.

Queue lengths and waiting times. The key performance measures of a polling system are the queue lengths and the waiting times. One of the most remarkable results regarding the marginal queue length distribution, denoted by L_i , has been obtained by Fuhrmann and Cooper [75]. They observed that the marginal queue length distribution can be decomposed as the sum of two independent random variables:

$$L_i \stackrel{d}{=} L_i^{(M/G/1)} + L_i^{(I_i)},$$

where $L_i^{(M/G/1)}$ is the queue length of an $M/G/1$ queue with only type i customers in isolation, often referred to as “the corresponding $M/G/1$ queue”, and $L_i^{(I_i)}$ is the queue length at an arbitrary moment during the intervisit period I_i . This results in the following PGF of the marginal queue length distribution:

$$\tilde{L}_i(z) = \frac{(1 - \rho_i)(1 - z)\tilde{B}_i(\lambda_i(1 - z))}{\tilde{B}_i(\lambda_i(1 - z)) - z} \times \frac{\tilde{L}_i^{(V_{c_i})}(z) - \tilde{L}_i^{(V_{b_i})}(z)}{\lambda_i \mathbb{E}[I_i](1 - z)}, \quad (1)$$

for $i = 1, \dots, N$. In (1) we have used $L_i^{(V_{b_i})}$ and $L_i^{(V_{c_i})}$ to denote the number of customers in Q_i at, respectively, the beginning and completion of a visit to Q_i . An important observation is that one only needs the queue length distribution at the beginnings and completions of a visit to each queue, in order to find the marginal queue length distributions. Moreover, Keilson and Servi show that the LSTs of the waiting time distributions follow directly from $\tilde{L}_i(\cdot)$ by applying the distributional form of Little’s Law [95]:

$$\widetilde{W}_i(\omega) = \frac{\tilde{L}_i(1 - \omega/\lambda_i)}{\tilde{B}_i(\omega)},$$

where $\widetilde{W}_i(\cdot)$ is the LST of the waiting time distribution of type i customers, *excluding* the service time B_i .

Joint queue length at polling epochs. Determining the marginal queue length distributions at visit beginnings and completions, requires studying the *joint* queue length distributions at these embedded epochs. It turns out that, a few exceptions aside, an exact analysis is only possible if all service disciplines in the polling system satisfy the following *branching property*, described by Fuhrmann [74] and Resing [135].

Property 2.1 (Branching Property). *If the server arrives at Q_i to find k_i customers there, then during the course of the server’s visit, each of these k_i customers will effectively be replaced in an i.i.d. manner by a random population having PGF $h_i(z_1, \dots, z_N)$, which can be any N -dimensional PGF.*

The PGF mentioned in the branching property depends on the service discipline. As an example, we give the expressions for exhaustive and gated service.

$$\begin{aligned} \text{Gated:} \quad h_i(z_1, \dots, z_N) &= \tilde{B}_i\left(\sum_{j=1}^N \lambda_j(1 - z_j)\right), \\ \text{Exhaustive:} \quad h_i(z_1, \dots, z_N) &= \pi_i\left(\sum_{j \neq i} \lambda_j(1 - z_j)\right), \end{aligned}$$

where $\pi_i(\cdot)$ is the LST of a busy period distribution in an $M/G/1$ system with only type i customers, so it is the root in $(0, 1]$ of the equation $\pi_i(\omega) = \tilde{B}_i(\omega + \lambda_i(1 - \pi_i(\omega)))$, $\omega \geq 0$ (cf. [45], p. 250).

Recall that the globally gated service discipline does not strictly satisfy Property 2.1, but still allows for an exact analysis. Borst [26] defines a generalization of the branching property that still ensures mathematical tractability. The globally gated discipline satisfies this weaker property.

Assuming that each queue satisfies the branching property, the PGF of the joint queue length distribution at visit completions, denoted by $\tilde{L}^{(V_{c_i})}(z_1, \dots, z_N)$, can be expressed in terms of the joint queue length distribution at visit beginnings, denoted by $\tilde{L}^{(V_{b_i})}(z_1, \dots, z_N)$. The following relations, which are commonly referred to as the *laws of motion*, hold:

$$\tilde{L}^{(V_{c_1})}(z_1, \dots, z_N) = \tilde{L}^{(V_{b_1})}(h_1(z_1, \dots, z_N), z_2, \dots, z_N), \quad (2)$$

$$\tilde{L}^{(V_{b_2})}(z_1, \dots, z_N) = \tilde{L}^{(V_{c_1})}(z_1, \dots, z_N) \tilde{S}_1\left(\sum_{j=1}^N \lambda_j(1 - z_j)\right), \quad (3)$$

$\vdots$

$$\tilde{L}^{(V_{c_N})}(z_1, \dots, z_N) = \tilde{L}^{(V_{b_N})}(z_1, z_2, \dots, z_{N-1}, h_N(z_1, \dots, z_N)), \quad (4)$$

$$\tilde{L}^{(V_{b_1})}(z_1, \dots, z_N) = \tilde{L}^{(V_{c_N})}(z_1, \dots, z_N) \tilde{S}_N\left(\sum_{j=1}^N \lambda_j(1 - z_j)\right). \quad (5)$$

The recursive equation for $\tilde{L}^{(V_{b_1})}(z_1, \dots, z_N)$ itself can be used to compute the moments of this joint queue length distribution. An explicit expression for $\tilde{L}^{(V_{b_1})}(z_1, \dots, z_N)$ can be found by clever definition of *offspring generating functions* and *immigration generating functions*. We refer to [135] for more details.

Mean waiting times. By differentiation of the LSTs of the waiting time distributions, the following expressions for $\mathbb{E}[W_i]$ are found for exhaustive, gated, and globally gated service respectively:

$$\text{Exhaustive:} \quad \mathbb{E}[W_i] = (1 - \rho_i) \frac{\mathbb{E}[C_i^{*2}]}{2\mathbb{E}[C]}, \quad (6)$$

$$\text{Gated:} \quad \mathbb{E}[W_i] = (1 + \rho_i) \frac{\mathbb{E}[C_i^2]}{2\mathbb{E}[C]}, \quad (7)$$

$$\text{Globally gated:} \quad \mathbb{E}[W_i] = \sum_{j=1}^{i-1} \mathbb{E}[S_j] + \left(1 + 2 \sum_{j=1}^{i-1} \rho_j + \rho_i\right) \frac{\mathbb{E}[C_1^2]}{2\mathbb{E}[C]}. \quad (8)$$

For a polling system with globally gated service, we have assumed that the server's arrival epoch at Q_1 determines which customers (in *all* queues) are eligible for service. Although the expressions (6)–(8) seem quite simple at first sight, it should be noted that the computation of the second moments of C_i^* and C_i is not straightforward. Although they can be obtained by solving the set of equations resulting from differentiation of the recursive expression for $\widehat{L}^{(V_{b_1})}(z_1, \dots, z_N)$, obtained from the laws of motion, with respect to $z_1, \dots, z_N$, more efficient methods have been developed. The two most frequently used alternatives to find mean waiting times are the *descendant set approach*, developed by Konheim et al. [101], and mean value analysis for polling systems, developed by Winands et al. [176].

Pseudo-conservation law. Boxma and Groenendijk [31] derive a so-called *pseudo-conservation law* (PCL) for the mean waiting times under various service disciplines. This PCL states that

$$\sum_{i=1}^N \rho_i \mathbb{E}[W_i] = \rho \frac{\sum_{i=1}^N \lambda_i \mathbb{E}[B_i^2]}{2(1-\rho)} + \rho \frac{\mathbb{E}[S^2]}{2\mathbb{E}[S]} + \frac{\mathbb{E}[S]}{2(1-\rho)} \left(\rho^2 - \sum_{i=1}^N \rho_i^2 \right) + \sum_{i=1}^N \mathbb{E}[Z_{ii}], \quad (9)$$

where Z_{ii} denotes the amount of work left behind by the server at Q_i at the completion of a visit, which solely depends on the service discipline of Q_i . For exhaustive service at Q_i , we have $\mathbb{E}[Z_{ii}] = 0$, and for gated service $\mathbb{E}[Z_{ii}] = \rho_i^2 \mathbb{E}[S]/(1-\rho)$.

Other models and/or techniques In the present section we have described only one technique, often referred to as the *buffer occupancy method*, first used by Cooper and Murray [47] and Cooper [46]. An advantage of this technique is its applicability to several variations of the basic polling model, described in the current subsection, such as polling systems with other service disciplines than exhaustive and/or (globally) gated service, systems with simultaneous arrivals [116], or fluid polling systems [49]. However, if the assumptions stated in the beginning of this subsection are not all valid, one might have to resort to alternative ways to analyze the system. For example, if the arrival processes are not Poisson, but BMAPs, a generalization of the buffer occupancy method using Kronecker product notation can be used [136]. Altman and Fiems [4] show how stochastic recursive equations can be used to analyse a polling model with correlated switch-over times. If the service disciplines do not satisfy the branching property, one may have to resort to different techniques, or even approximations. For example, a two-queue polling system with Bernoulli service (a generalization of 1-limited as well as exhaustive service) has been analyzed by Lee [109] by formulating the model as a Riemann boundary value problem with a shift. Unfortunately, to analyze polling systems with non-branching service disciplines consisting of more than two queues, one may have to use approximations or numerical methods. Van Houdt [165] uses a numerical algorithm based on the power method, Kronecker matrix representations, and the shuffle algorithm, to analyze discrete-time polling systems for networks on chips (which will be discussed later in this paper). This algorithm can also be used for general polling systems with branching-type or non-branching-type service disciplines.

3 Applications

In this section, we will first describe the three most successful application areas of polling models within the classical fields of engineering of communication systems, production sys-

tems, traffic and transportation systems. Subsequently, the last subsection summarizes a variety of miscellaneous applications of polling systems.

3.1 Computer-communication systems

Polling systems find a wealth of applications in the area of computer-communication systems, where resources (e.g., bandwidth, CPU power) are shared among different users. The reader is referred to surveys by Grillo [85], Levy and Sidi [115], Takagi [152] and Weststrate [172] for overviews of applications up until the early 1990s. For completeness, the main applications cited in these survey papers are outlined below, and supplemented with more recent applications of polling models in communication systems.

Time-sharing computer systems. Classical applications of polling models are time-sharing computer systems [100], consisting of a number of terminals connected by multi-drop lines to a central computer. The data transfer from the terminals to the computer - and back - is controlled via a polling scheme in which the computer polls the terminals, requesting their data, one terminal at a time. In such applications of polling models, the server represents the central computer, the queues represent the terminals and customers represent the data.

In communication networks different terminals compete for access to a shared medium. If multiple terminals transmit or receive data over the medium at the same time, packet collisions and interference problems may occur. Motivated by this many medium access control (MAC) protocols have been proposed for different network technologies, in many cases leading to polling models.

Token-ring networks. Bux [42] uses polling models to study the performance of token-passing schemes in Local Area Networks (LANs) where a token - representing the right for transmission - is circulated among the different users. In such cases, the token-passing scheme is usually configured in a ring or a bus topology. A token-ring network can be characterized as a set of stations connected to a common transmission medium in a ring topology. All messages travel over a fixed route from station to station around the loop. A token can be in two states: occupied or free. A station with data to transmit reads the free token and changes it to the occupied state before retransmitting it. The occupied token is then incorporated as part of the header of the data transmitted on the ring by the station. Thus, other stations on the ring can read the header, note the occupied token and refrain from transmission. When the token is back at the station that changed it to the occupied state and the station decides to transfer the right for transmission to another station, it changes the token to the free state. The token-ring network allows the transmission of packets in a conflict-free manner. However, transmission of packets may still fail due to errors and distortions on the ring itself. Typically, these errors are rare and a so-called Selective-Repeat ARQ (SR-ARQ) could be used to recover these errors. In such a scheme, a station that receives an erroneous message transmits a negative acknowledgment to the transmitting station to indicate that the message has to be retransmitted. To analyze the performance of SR-ARQ schemes, Levy and Sidi [115] use a polling model in which each station is represented by two queues, one for messages that need to be sent out and one for negative acknowledgments to be sent back when erroneous messages are received. Altman and Kofman [6] propose a solution to deal with the irregular,

bursty, correlated arrival processes in token-ring networks. To this end, they use a polling model with so-called Cruz-type traffic, filtering the arrival streams by leaky buckets.

Token-bus networks. The token-bus network consists of a set of stations connected to each other in a bus topology. The intention behind this technique is to combine the attractive features of the bus topology with those of a conflict-free medium-access protocol. In a token bus, as the token is passed, a logical ring is formed. Since the bus topology does not impose any sequential ordering of the stations, the logical ring is defined by a sequence of station addresses. Conceptually, token passing on buses and rings is very similar and the same type of performance model can be used to describe these techniques. The difference between a token ring and a token bus from a modeling point of view is that the server in the token ring network visits the queues in a cyclic manner, whereas in the token-bus model the server moves along the queues in a non-cyclic periodic manner, which can be modeled by a polling table. An example of a token-bus network is presented by Manfield [119], where a communication network constitutes the transmission medium between a master processor and a set of peripheral processors. The operation of these systems works as follows: the master processor polls each of the peripheral processors in turn. Subsequently, it receives a response which indicates whether the peripheral processor has any packet to send. If it has, the packet at the head of that queue is sent to the master processor. After polling a peripheral incoming queue, the master processor checks its own queue for outgoing packets, and if the queue is not empty, seizes control of the bus. The advantage of this system, is that it ensures that outgoing messages are rapidly sent and it avoids message deadlocks in the central control facility (i.e., the master processor). This polling scheme is called *star polling* in the literature.

Slotted-ring networks. Another class of communication protocols for networks with a ring topology is a so-called slotted-ring network. In a slotted-ring network one or more slots circulate along the stations. If there is a packet at a station ready for transmission and an empty slot comes along, the packet is put into the slot, together with the address of the destination station. That slot is then examined by each of the other stations in turn, until the destination station recognizes it and copies its content. There are two possibilities to empty the slot: either it is emptied by the source station (this is called the Cambridge Ring Protocol), or by the destination system (this is called the Orwell Ring Protocol). We refer to Bux [42] for a polling model with source release, and to Van Arem [153] for a polling model of a slotted-ring protocol with destination release.

Fibre Distributed Data Interface networks. The FDDI is a token-passing protocol for LANs with a ring topology in which the access to the ring for transmission is controlled via a so-called timed-token protocol, i.e. where the transmission time for each station is bounded. This type of models leads to the formulation of polling models with time-limited service policies [141].

Distributed Queue Dual Bus networks. The DQDB protocol is a multiple access protocol for communication networks consisting of two unidirectional buses carrying information in the opposite directions. The stations are distributed along the two buses and have the

capability to transmit/receive information to/from both buses. The DQDB protocol is intended to integrate data, voice and video traffic in a single communication network. Bisdikian [17] uses a polling model to study medium access mechanisms of a single station in a DQDB network.

Random access schemes. Opposite to the scheduled multiple access protocols are the random multiple access protocols, where an entity with a message will transmit it regardless of potential collisions. The ALOHA access scheme is an example of a random access scheme. As soon as a packet arrives at a station it is transmitted. When the transmission fails, for example because other packets were transmitted at the same time causing a collision, the packet is scheduled for transmission after a random period of time. An alternative for this scheme is the so-called reservation ALOHA access scheme. In this scheme, a station is granted the exclusive right to transmit, without being interfered by any other station, for a certain amount of time. When a transmitting station no longer reserves the channel, some - or all - stations of the system start contending in order to seize the channel. The length of the contention period is random and the next station that will seize the channel is also random. This type of protocols naturally leads to the formulation of polling models with random routing [36, 98]. A detailed comparison of various multi-access schemes, including random access schemes like ALOHA and CSMA, is given by Kleinrock [97].

Optical networks. Polling models also find applications in the area of Ethernet Passive Optical networks (EPONs), where packets from different Optical Network Units (ONUs) share channel capacity in the upstream direction. An EPON is a point-to-multipoint network in the downstream direction and a multipoint-to-point network in the upstream direction. The Optical Line Terminal (OLT) resides in the local office, connecting the access network to the Internet. The OLT allocates the bandwidth to the Optical Network Units (ONUs) located at the customer premises, providing interfaces between the OLT and end-user network to send voice, video and data traffic. In an EPON the process of transmitting data downstream from the OLT to the ONUs is broadcast in variable-length packets according to the 802.3 protocol [104]. However, in the upstream direction the ONUs share capacity, and various polling-based bandwidth allocation schemes can be implemented. Simple time-division multiple access (TDMA) schemes based on fixed time-slot assignment suffer from the lack of statistical multiplexing, making inefficient use of the available bandwidth, which raises the need for dynamic bandwidth allocation (DBA) schemes. A dynamic scheme that reduces the time-slot size when there are no data to transmit would allow excess bandwidth to be used by other ONUs. Kramer et al. [103, 105] propose an OLT-based interleaved polling scheme similar to hub-polling to support dynamic bandwidth allocation. To avoid monopolization of bandwidth usage of ONUs with high data volumes they propose an interleaved DBA scheme with a maximum transmission window size limit.

Bluetooth. Bluetooth is a wireless technology standard, used for exchanging data between mobile devices such as mobile phones, laptops, and headsets. These devices form small networks, referred to as Wireless Personal Area Networks (WPANs). The basic Bluetooth network topology is called a *piconet*, consisting of one master device and up to seven slave devices. Miorandi et al. [124] observe that the structure of a piconet consisting of N slaves,

can be modeled adequately using a polling system consisting of $2N$ queues. One queue is used for each master-to-slave communication link, and one additional queue is required for each slave-to-master link. Approximations of the mean delays are found for the Pure Round-Robin (corresponding to 1-limited service), gated, and exhaustive disciplines. Zussman et al. [185] study the same model, deriving exact results regarding the packet delays.

I/O subsystems. Polling models occur in the context of I/O-subsystems of file servers or database management systems as well. File servers are expected to handle large volumes of information-retrieval requests within a reasonable amount of time. After the requested information (either static or dynamic) has been gathered, the information is ready for transmission to the client over the communication network that connects the client to the server. To this end, the information is placed in an I/O-subsystem, where it is ready to be drained to the client over the network. This can be done for example via the Transport Control Protocol (TCP), implementing a window-based control mechanism. The I/O-subsystem will typically consist of a number of parallel TCP-buffers. The draining of the different buffers is managed by an I/O-controller that controls the access of the buffer content to the network. To this end, the I/O controller “visits” the buffers in some order to check whether a buffer has information to be drained, and if so, whether the corresponding TCP-window is open, allowing for transmission of data segments residing in the buffer via the network to the client. We refer to [89, 158] for performance models of Web servers including I/O-controlled buffers. This type of I/O-control leads to the formulation of polling models where the service discipline represents the complicated dynamics of TCP-window control. In [48], Czerniak et al. analyze TCP systems using polling models to describe the different transmission methods.

Mobile networks. Polling models can also be found in the area of mobile networks, where different users compete for access to the shared scarce radio resources. In such environments, the base station is typically in charge of assigning time slots to the different users in some way. In this context, the server represents the right for transmission and the customers represent data packets to be transmitted. Typical examples of polling mechanisms occur in the context of the Code Division Multiple Access (CDMA) based High Speed Packet Access (HSPA), where the base station controller grants access to the medium in a per-timeslot basis. There are different scheduling mechanisms for deciding which of the terminals gets access to the medium for the duration of a single time slot. A common implementation is simple Round-Robin (RR) scheduling, where medium-access is circulated among the terminals, independent of the quality of the signal [154]. This immediately leads to polling models with limited service policies and cyclic server routing. A straightforward extension of RR scheduling is Weighted Round-Robin (WRR), which leads to the formulation of polling models with periodic server routing. To enhance the efficiency of medium access for HSPA-networks, highly sophisticated channel-aware scheduling mechanisms have been proposed, granting time slots to terminals based on the measured instantaneous (possibly relative) Signal-to-Noise Ratios (SINRs) of each of the terminals [18, 27]. The inherent randomness in the channel conditions, and hence in the order in which the stations are granted access to the medium naturally leads to the formulation of polling models with random or Markovian server routing [36, 98].

Ferry-based Wireless LANs. Polling models are applicable in the context of designing message ferry routes in Ferry-based Wireless LANs. In such FWLANs, a number of isolated nodes are scattered over some geographical area where communication between a node and the outer world, or communication between nodes, is made possible via a message ferry. The ferry follows a predetermined cyclic path, collecting messages from and delivering messages to nodes whenever it is *in the vicinity* of the node (implying that the routes do not have to pass *through* all the points in space). The transmission and reception range is flexible: assuming a fixed transmission power, the range can be increased at the cost of decreasing throughput. Taking into account the relation between the range and the throughput, one can design an optimal cyclic route of the ferry. We refer to Kavitha and Altman [93] (and references therein) for results in FWLANs.

Mobile adhoc networks. Polling models occur naturally in the modeling of mobile ad-hoc networks (MANETs), consisting of both mobile and fixed wireless terminals. A typical feature of these kind of networks, is that wireless devices create their own wireless network in a distributional fashion. Mobile users can change location and thereby change communication links in the network. Examples of MANETs are animal-monitoring systems, collaborative conference computing, vehicular networks, peer-to-peer file-sharing systems and disaster-relief networks. Unlike most classical wireless networks, MANETs allow for multi-hop communication. Communication links appear and break down dynamically. Their lifetimes are random and independent of the amount of traffic offered to or transmitted over such links. To capture this phenomenon, De Haan [54] proposes the so-called pure time-limited service policy, where the server visits a queue for a random amount of time, independent of anything else in the system; note that this policy is not work conserving. The reader is referred to [53] (Chapter 1) for an overview of applications of polling models in MANETs.

Networks on chips. Another interesting application area of polling models is networks on chips (NOCs), which have been proposed as a solution for the inefficiency caused by traditional bus connections [12, 51]. In NOCs, intellectual property blocks (a general term for on-chip modules) are not connected to a single shared link, but to network interfaces that implement communication protocols. Data is transmitted using switches that consist of input and output ports. If multiple ports have data for the same output port, only one input port can transmit its data, and the switch selects which one. Data that is not transmitted immediately is stored in buffers and will be transmitted later. We refer to [11] for overview of the applicability and the state-of-the-art on analysis of polling models for networks on chips, and to [165] for an efficient numerical algorithm.

3.2 Production systems

A completely different application area of polling systems can be found in the so-called *stochastic economic lot scheduling problem* (SELSP). The SELSP deals with the make-to-stock production of multiple standardized products on a single machine with limited capacity under *random demands, possibly random setup times and possibly random production times*. The SELSP is a common problem in practice, e.g., in glass and paper production, injection molding, metal stamping and semi-continuous chemical processes, but also in bulk production of

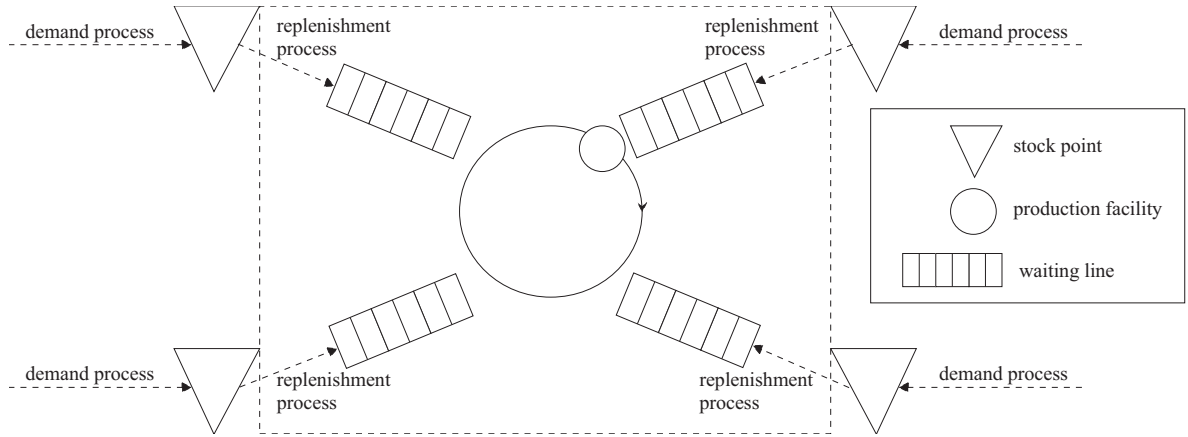


Figure 1: The fixed-sequence base-stock system.

consumer products such as detergents and beers.

In many firms encountering the SELSP, the following class of *fixed-sequence base-stock policies* is used for the control of the inventory of each product. We distinguish N products, which are numbered $1, 2, \dots, N$. Subsequently, to each individual product a stock point is assigned which is controlled by a base-stock inventory policy. Under such a policy, for each product there exists a pre-defined desired number of items in stock, the base-stock level b_i , $i = 1, 2, \dots, N$. When demand arrives at a stock point and the requested product is on stock, the demand is immediately fulfilled. Otherwise, demand is backlogged and fulfilled as soon as the product becomes available after production. A production order, also called replenishment order, is placed immediately after demand for the corresponding product has arrived. These production orders queue up at the production facility, where each product has its own designated queue.

On the strategy deployed by the production facility, the following two restrictions are imposed,

1. The products are produced according to a fixed production sequence;
2. When the machine starts production of a product, it will continue production until either the base-stock level has been reached or a second local criterion, i.e., only dependent on the stock level of the product currently setup, has been fulfilled.

The combination of stock points and production facility is visualized in Figure 1.

For given values of the base-stock levels b_i the steady-state net stock level N_i for product i is given by, for $i = 1, 2, \dots, N$,

$$N_i = b_i - L_i, \quad (10)$$

where L_i denotes the steady-state *shortfall* (the number of outstanding production orders at the production facility) of product i . Notice that the net stock level of a product becomes negative, when the shortfall of this product is larger than its base-stock level. One can verify that the shortfall of a product is independent of the base-stock levels implying that the performance of the production facility can be analyzed independently of these base-stock levels. Moreover, the shortfall distribution of a product at the production facility is identical

to the queue length distribution in a polling system. The interarrival, service and setup time processes in such a polling system are identical to the demand, processing and setup time processes in the SELSP, respectively. *For given base-stock levels, the evaluation of a fixed-sequence base-stock policy is, therefore, tantamount to evaluation of the corresponding polling system.*

Literature on the SELSP. Seminal papers analyzing the above fixed-sequence base-stock policy via polling systems are by Federgruen and Katalan [67, 69, 70]. Besides the basic assumptions for the SELSP, products are produced by an exhaustive or a gated base-stock policy. The production manager is allowed to insert a fixed idle time prior to the setup for a product in order to reduce the setup frequencies and, hence, the average setup costs.

Grasman et al. [83] extend the exhaustive base-stock model of Federgruen and Katalan by adding random yields for the cases of backlogging, lost sales and expediting. In case of backlogging they derive similar newsboy equations in order to obtain optimal base-stock levels. In case of lost sales or expediting they have to resort to a heuristic for computing the (approximate) optimal base-stock levels. Krieg and Kuhn [106, 107] introduce continuous-time models for single-stage multi-product *Kanban systems*, which are completely identical to the SELSP with lost sales. In Krieg and Kuhn [106], a system is analyzed with *state-independent* setups, whereas in Krieg and Kuhn [107] *state-dependent* setups are modeled, i.e., no setup for a product is incurred when there is no shortfall. Production quantities are in both models determined by an exhaustive base-stock policy.

Winands et al. [177] analyze a two-product system, in which a high-priority product is produced exhaustively and a low-priority product according to the k -limited service strategy. Chapter 4 of [174] is devoted to a numerical simulation study which shows that the k -limited lot-sizing policy outperforms the standard exhaustive policy for a wide variety of environments. In particular, the k -limited policy proves its worth in asymmetric production systems.

In most applications, the service discipline determines how many items are produced, and, hence, what the (random) duration of a cycle in the polling system is. In some cases, however, the cycle length is fixed implying that the corresponding analysis often reduces to a one-dimensional problem. See [81] for a case study of a chemical plant, for which a fixed production sequence strategy in combination with a fixed cycle length has been developed under the assumption of deterministic production and setup times. Erkip et al. [66] introduce a discrete-time model under the assumption of backlogging, in which the production and setup times are deterministic. Other work in this direction is by Bruin [40], who presents a generating function approach for the fixed cycle strategy under general traffic settings, and by Dellaert [56], who develops a heavy-traffic approximation for the optimal base-stock levels.

Make-to-order. Summarizing, we can say that the SELSP is an extension of a standard polling system by an additional inventory dimension. The make-to-order scheduling counterpart (in which no stock is kept for a product and thus the base-stock levels equal 0) is, however, precisely equivalent to a standard polling model. The polling system again consists of a server, mostly referred to as machine, and multiple customer classes, which will be referred to as products.

In the literature, one can notice that (almost) all modeling variants as discussed in Section 2

have been studied in the context of such a make-to-order production situation. This implies that an enormous amount of papers on polling systems have appeared motivated by make-to-order manufacturing applications. Since the primary goal of the present paper is to give an overview of the various application areas of polling systems - and not to give an exhaustive list of papers within each area - we only give a small subset of recent papers on make-to-order polling systems (i.e., [13, 59, 60, 108, 127, 132, 142, 145]).

Finally, in Markowitz et al. [120, 121] heavy traffic polling results are applied to all kinds of related stochastic multi-product single-machine scheduling problems.

The impact of setup times. Besides papers on the performance analysis of polling systems in the context of production environments, much research has been published on the impact of setup times (see [58, 80, 122, 139, 183, 184]). In these papers, it is shown that reduction of setup times can, counterintuitively, increase the mean queue lengths in polling systems and cyclic production systems for a variety of settings. Furthermore, the conditions are studied under which mean queue lengths increase when setup times are reduced [68, 69, 137, 138, 174].

3.3 Traffic and transportation systems

The third important area in which polling systems are frequently applied in practice are traffic and transportation systems. In road traffic a polling system is the natural way to model a situation where queues arise due to the fact that multiple flows of traffic have to share one single lane. The plainest form is a two-way road which is partly blocked because of an accident or road maintenance, but any common traffic intersection also qualifies. Besides road traffic, polling models are employed in transportation systems consisting of driverless electric vehicles that follow a predetermined track. The actual transportation is performed by so-called Automated Guided Vehicles (AGVs). AGV systems are used in warehouses and container terminals, but also in other application areas like, e.g., health care.

Traffic signals

We discuss applications of polling systems in road traffic first. Queues are formed by the lines of cars waiting before a red traffic signal. The time that is required for one vehicle to pass the traffic signal can be viewed as service time, while the times that all queues face a red light simultaneously, i.e. the clearance times of the intersection, can be considered as switch-over times. There are several aspects that make polling models for signalized intersections different from the models in other application areas. Firstly, the assumption of independent, identically distributed service times is in practice not valid. When a green period commences, a certain time elapses while vehicles are accelerating to normal speed. After this time, which typically lasts a few seconds, the queue discharges at a more or less constant rate, which is called the saturation flow. In practice the saturation flow may vary within a cycle and between cycles, but Webster [171] finds that the assumption of a constant saturation flow agrees well with values observed over a fairly large number of cycles. A typical feature of the traffic light queue is that cars approaching an empty intersection while facing a green light do not slow down and require almost no service time at all, especially if they do not have to take a turn. The second difference between traffic intersections and many other polling applications

is the arrival process. The assumption of Poisson arrivals might be realistic for an isolated traffic intersection, but many intersections are part of an arterial system, which means that an output process of the first intersection forms an input process for the next intersection. Only few papers deal with these so-called *platooned* arrivals. The third difference that we discuss is that most traffic intersections are divided into groups of flows that face a green light simultaneously. In the polling system this can be modeled as a (possibly varying) number of multiple coupled servers serving queues simultaneously. The moment at which a switch-over is incurred depends on the service discipline that is used for the intersection. The best known discipline is the *Fixed Cycle Traffic Light* discipline, which uses deterministic green, red and amber times. The obvious disadvantage of fixed settings is that the system does not respond to the current situation, which might be less efficient because traffic lights remain green even when no cars are waiting in the corresponding queue. An alternative to fixed settings that is very commonly used is an *adaptive control mechanism* for traffic signals. This is a mechanism that detects the presence of vehicles and makes decisions on whether or not to switch a signal based upon this information. A typical policy is to wait until all traffic flows that face a green light are empty before switching to red. For some intersections, a time limited policy might be more suitable. See [130] for a discussion on this topic. We discuss literature on polling models for traffic intersections in more detail by splitting it into two groups: papers concerning the fixed cycle traffic light queue, and papers on vehicle-actuated signals.

Literature on the Fixed Cycle Traffic Light queue. A traffic intersection with fixed green, amber and red times can be viewed as a polling model with deterministic visit times and switch-over times. One may argue whether this system should be considered a real polling system, since the analysis of delays is a one-dimensional problem and one can focus on one queue in particular. One of the most influential papers on the fixed cycle traffic light queue is [171], in which Webster describes a procedure to find optimal settings (i.e., green and red times for all traffic flows) for traffic lights. He derives an approximate expression for the optimal cycle length and an expression for the average delay per vehicle. His expressions are partly based on theoretical grounds, and partly on simulation results. Although more sophisticated methods have been developed ever since, his results are still quite popular in practice. Few years later, Miller [123], under the assumption of Poisson arrivals, and Newell [128], for general arrivals, develop approximations for the distributions of the delay and the queue length. Exact expressions for these distributions have been obtained by Heidemann [90], but again under the assumption of Poisson arrivals. In a recent paper Van Leeuwen [166] derives the queue length and the waiting time distribution for general arrival processes. Van den Broek et al. [155] derive approximations for the mean overflow, i.e., the mean queue length at the end of a green period, that are easier to compute and provide more intuitive insight.

In practice, intersections are often part of an arterial system and interarrival times are correlated. Alfa and Neuts [2] regard a model with platooned arrivals, which allows for distinguishing between interarrival times between cars within a platoon, and interplatoon interarrival times. The number of cars within one platoon can have any discrete phase probability distribution. The platooned arrival process is applied to road traffic, and to the fixed cycle traffic light queue. Through a numerical example they conclude that ignoring correlation in the arrival process leads to an underestimation of the mean queue length at high traffic intensities.

Literature on vehicle-actuated traffic signal control. Nowadays fixed cycles are less and less frequently implemented at traffic intersections, but it might still be a realistic assumption when the intersection is congested during rush-hour. Most traffic signal systems are vehicle-actuated, which means that the system contains detectors that gather information about the number of cars present at each flow and regulate the green and red times based on this information. In particular, when a queue becomes empty, the corresponding traffic light turns red and the traffic light of the next group of flows becomes green. This situation corresponds to a polling model with exhaustive service. However, in general, vehicle-actuated systems are designed to have minimum and maximum green times for each flow. This makes them very suitable to model as a polling system with time-limited service, but it also makes them difficult to analyse. Frigui and Alfa [72, 73] propose an iterative algorithm to approximate waiting times in a discrete-time polling system with time-limited service. Their papers focus on applications in communication systems, but in [71] the model is also applied to traffic intersections.

The earliest literature on vehicle-actuated systems dates from the early 1960s when Darroch et al. [52] analyze a system that consists of two intersecting traffic streams that are served exhaustively. A model with two lanes that does not assume Poisson input has been studied by Lehoczký [110], who uses an alternating priority queueing model. Newell [129] analyzes an intersection with two one-way streets using fluid and diffusion queueing approximations. Daganzo [50] studies a polling system with more general arrival and service processes, with applications in traffic and transportation systems (as long as only one flow of traffic is served at a time). In [131], Newell and Osuna study a four-lane intersection where two opposite flows face a green light simultaneously. A variation of the two-lane intersection is introduced by Greenberg et al. [84], who analyse mean delays on a single rail line that has to be shared by trains arriving from opposite directions. This model is extended by Yamashita et al. [180], who study alternating traffic crossing a narrow one-lane bridge on a two-lane road. In many of the discussed papers traffic is modeled as fluid passing through the road. This approximation is fairly accurate when the traffic intensity is relatively high. Vlasiou and Yechiali [170] use a different approach, modeling a traffic intersection as a polling system with an infinite number of servers visiting each queue simultaneously.

All in all it is rather surprising how few mathematical models have been developed that deal with intersections where multiple flows face a green light simultaneously. In [22] a two-queue polling model is analyzed with two priority levels in one queue. This model can be applied to an intersection where two *conflicting* traffic flows share the same green slot. In [88] a Markov Decision Problem (MDP) decomposition approach is used to find optimal traffic signal settings at intersections with larger numbers of combined flows.

Automated guided vehicles

Next to traffic intersections, a large transportation related application area of polling systems are Automated Guided Vehicles. The first AGV system was simply a tow truck that followed a wire in the floor instead of a rail, nowadays AGVs are mostly laser navigated. AGVs are mostly used for the transport of materials in manufacturing systems, but also in other areas like public transportation systems at airports. Not all AGV systems can be modeled as a polling system, but we discuss some exceptions. In a conventional AGV system, each vehicle can pick up a load from any station and deliver it to any other station. In order to

avoid collisions between vehicles, most systems use the *zone blocking* concept where the entire system is divided into zones. The control system allows only one vehicle in each zone at a time. Whenever an AGV system consists of a single loop, it can be modeled as a polling system. The vehicle corresponds to the server in the polling system, and the stations form the queues of the system. In some cases each station is modeled as two queues, one for the picking-up, and one for the drop-off. The inter-station travel times are modelled as switch-over times in the polling model. AGV systems with multiple vehicles can be modeled as a polling system with multiple, independently moving servers, but it is more common to approximate this situation by modeling it as a system with one vehicle that moves at a higher speed (see, e.g., [61]). It is typical for AGV systems that the vehicle in general does not visit the stations in a deterministic order. Although most papers concerning polling systems focus on a fixed visiting order of the queues, several papers discuss an alternative server routing. Srinivasan [146] uses a polling system with Markovian server routing to model AGV systems. In [33] and [16] optimization of server routing in polling systems is discussed. In AGV systems the visit order is generally determined by a decentralized cargo dispatching rule. The most commonly used cargo dispatching rule in AGV systems is First-Encountered-First-Served (FEFS) rule. The FCFS rule is less efficient, because it leads to unnecessary empty travel. Note that FCFS in the context of AGV systems generally does not refer to the order in which items are picked up within a queue, but to the route that the vehicle takes. In terms of queues, it travels towards the queue with the oldest waiting customer. The FEFS rule is presented for single-loop AGV systems in [10] and states that an empty vehicle continues to travel along the loop until it finds some load to pick up at some station. If it has available space, it picks up the load and drops it off at its destination. The number of packages that are picked up depends on the service discipline and may vary per station.

Bozer and Srinivasan [37] discuss an AGV system with tandem configurations. The original AGV system is divided into non-overlapping, single vehicle closed loops with load transfer stations in between (see Figure 2). Each loop can be modeled as a single server polling system. An advantage of a tandem AGV system is that each vehicle is dispatched over a smaller number of stations. Another advantage is that traffic management problems within each loop are completely eliminated. In [37] the mean throughput capacity for the loop is estimated under the FEFS rule. Srinivasan et al. [147] study this rule as well, and also introduce a modified FCFS dispatching rule, under which a vehicle deposits a load at a station and then checks that station for a new load-transfer request. If there is one, the vehicle serves this request; otherwise, it travels empty to other stations based on the FCFS dispatching rule. Xu et al. [179] focus on the loop configuration rather than on the cargo dispatching rule. Ganesharajah et al. [79] consider both simultaneously. Van der Heijden et al. [156] study an underground transportation system that contains a single traffic lane which has to be shared by AGVs from opposite directions. The system has similarities with a road traffic situation, but the extremely long clearance times require a more dedicated approach. The research is continued in [62] where more intelligent, adaptive control rules and dynamic programming algorithms are used to minimize vehicle waiting times.

3.4 Miscellaneous

The present subsection provides an overview of miscellaneous applications of polling systems observed in the open literature.

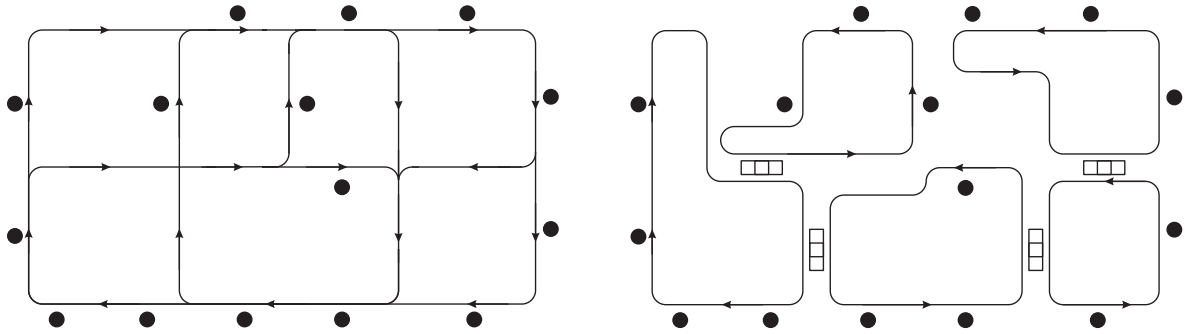


Figure 2: A conventional AGV system (left) and a tandem AGV system (right) as proposed by Bozer and Srinivasan [37]. The dots represent stations with pick-and-drop points. In the tandem AGV system, each loop is modeled as a single server polling system, and additional pick-and-drop points are placed as interface between adjacent loops. This figure appeared originally in [37].

Health care. In [43] a medical emergency room is modeled via a multi-server polling model. At the emergency room different types of patients arrive and are placed in queues depending on the type of surgical procedure required. Each emergency surgical procedure has a separate queue of infinite buffer capacity. The number of emergency room theatres is limited, and each emergency room theatre is modeled by a server in the polling model. Since setting up for surgery procedures takes longer than the actual procedures themselves, the service times within the polling model are much smaller than the switch-over times. Finally, an urgency parameter in terms of average waiting time is assigned to each patient which dictates the existence of (local) priority levels within each queue. From the application point of view, Cicin-Sain offers, albeit in a highly idealized mathematical setting, a preliminary exploration of how to design emergency rooms such that the most urgent patients have the shortest waiting times [43]. Another application in health care is depicted in [169], where a polling model, consisting of two queues with infinite supply, is used to model scheduling strategies by surgeons performing eye surgeries.

Mail delivery. Another polling application is an internal mail delivery system, which was modeled via a continuous polling system (a system with an infinite number of queues) by Nahmias and Rothkopf [126]. In their model, a clerk traverses at a constant rate a cyclic route along which mail is generated according to a Poisson process. The clerk picks up the mail that has been generated since the last traversal and delivers the mail (that was previously picked up and sorted) to locations distributed uniformly along the route. At the end of the route, the clerk sorts the mail just picked up. Then, the clerk again traverses the route, delivers the mail that has just been sorted and picks up the new mail and so on. In [126] also the extension is studied in which there are several independent routes for a couple of clerks and a single room where sorting takes place.

Sarkar and Zangwill [140] discuss a finite-queue variant of the above problem. That is, they study a polling system where the workload at a queue either comes from outside the system or from another queue within the system. In terms of the above mail delivery application the sorting of the clerk would be done at one designated queue N , whereas at the other queues

the mail is generated by exogenous Poisson processes. The sorting work done by the clerk at queue N obviously depends upon the amount of mail generated at the other queues. Besides the mail delivery application, Sarkar and Zangwill also describe applications of this model related to rework in manufacturing systems, computer file transfers and buses circulating at airports.

Snow blowing. Eliazar [64, 65] studies a so-called snowblower problem. That is, on a closed-loop racetrack snow is falling randomly and a snowblower machine is continuously circling this racetrack and clearing off the snow. This snowy racetrack can be modeled by a continuous polling system by drawing the following analogies: server $\leftrightarrow$ snowblower; job arrivals $\leftrightarrow$ snowfall; workload $\leftrightarrow$ snowload.

Shipyard loading. Another application of a polling system is a shipyard loading problem in which containers arrive by truck to the port. Trucks drive the containers up to their destination and the receiving yard crane transfers them to their assigned position. The crane represents the server of a polling system, whereas the trucks arriving for a specific destination are customers of a specific type. As stated by Daganzo [50], storage room (or: queue lengths) is the most critical performance measure for transportation problems dealing with freight. For transportation problems dealing with passengers, the waiting time is the most commonly used performance measure. In the aforementioned shipyard loading problem, one is especially interested in the characteristics of the *total* queue length in the polling system. That is, early arriving trucks wait in a common area until they are due for service. To dimension up this common area, information on the total queue length is needed.

Dynamic Picking Systems. Gong and De Koster [82] use a polling model to describe a dynamic order picking system (DPS). In a DPS, a worker picks orders that arrive in real time during the picking operations and the picking information can dynamically change in a picking cycle. It is very important to achieve short delivery times, especially for online retailers. One of the challenging questions that online retailers now face is how to organize the logistic fulfillment processes during and after order receipt. In traditional stores, purchased products can be taken home immediately. However, in the case of online retailers, the customer must wait for the shipment to arrive. In [82] polling models are used to describe and identify a DPS for online retailers. Polling-based picking systems can lead to shorter throughput times than traditional batch picking systems, particularly for high order arrival rates.

Elevators. A (multi-server) polling model has been studied by Gamse and Newell [76, 77] for an application of elevators in a building. In the model assumptions are made pertaining to the relative movements of servers within the system, with the assumptions being formed based on the nature of the specific application of elevators. In this context it is also interesting to spend some words on the so-called *elevator* server routing scheme. In such a system, the server first serves queues in the “up” direction, i.e., in the order $1, 2, \dots, N-1, N$, and subsequently serves these queues in the opposite (“down”) direction, i.e., visiting them in the order $N, N-1, \dots, 2, 1$ (see, e.g., Altman et al. [5]). It turns out that the globally gated service discipline in combination with elevator-type server movement results in equal mean waiting times for customers in *all* queues, establishing a perfectly fair polling system.

Another way of modeling an elevator system using a polling model can be found in [78]. In this paper an MDP approach is used to dynamically schedule elevators in a building. However, the model under consideration is only suitable for a limited amount of practical purposes, because all customers are assumed to have one common destination, which is the ground floor.

Besides the different routing scheme, another difficulty in modeling elevator systems is the fact that each idle elevator should return to the floor that is its home base. In polling literature some work has been done on this topic, which is called a stopping or dormant server, cf., [26, 63].

Maintenance. We would like to end this survey with actually the first application of a polling system which appeared in the open literature, i.e., in the area of maintenance. That is, Mack et al. [117, 118] use a polling model to describe a patrolling repairman who inspects a number of machines to check whether a breakdown has occurred and if so, eliminates such breakdowns. Evidently, in a polling model the repairman is represented by the server, the breakdowns are represented by the customers and the times needed by the repairman to travel from one machine to the next are represented by the switch-over times. In [102] a similar model is studied in which an operator at a fixed position serves a number of storage locations on a rotating carousel conveyor. Models with several independent rotating carousels have also been considered, cf. [41, 96]. Weststrate [172] studies a polling model which captures the behavior of a single repairman who is not only concerned with corrective maintenance, i.e., maintenance after a breakdown has occurred, but who can also perform preventive maintenance.

4 Outlook

Let us conclude the present survey with some personal views on possible future research directions related to applications and practical considerations. The present section is structured identically to the preceding section.

4.1 Communication systems

The technological advances in information and communication technology lead to the formulation of polling models with features that are not included in the classical polling models framework for which results are known today, and as such will lead to a variety of scientific challenges in the years to come. In this subsection we address a number of such open research problems.

The applicability of the results on polling models that are known today is strongly limited by the usual *Poisson assumption*, often needed to obtain exact results. Despite the fact that the packet-level dynamics on communication networks are well-known to be highly complicated and often far from Poisson, the vast majority of research results explicitly rely on the Poisson assumptions, or some straightforward relaxation of that. Extension of Poisson-based results to non-Poisson and/or dependent arrival processes raises tremendous scientific challenges, requiring new queueing-theoretical analysis methods. The development of such methods would strongly enhance the applicability of queueing-theoretical results on polling models. In

this context, encouraging results (but still only for renewal arrivals) have been obtained in [14, 133, 162, 163].

Another aspect that limits the applicability of results on polling models is *inflexibility with respect to the service policies*. In fact, the modeling of many computer-communication networks often leads to polling models where the service policies, typically representing the right for transmission/receipt of data to a shared medium, are of limited-type, and as such do not possess the convenient branching structure that allows for exact analysis. This raised the great need for the development of techniques to handle limited-type service policies, although there is not much hope for detailed exact analysis of such models. This stresses the importance of the development of asymptotic results (e.g., asymptotics on heavy traffic, large numbers of queues, heavy tails), a direction of research that has been highly successful in analyzing many queueing models in the past couple of years. Development of asymptotic results on polling models with limited-type service policies will continue to form a challenging area of research in the years to come.

The assumption of having a *single* server serving multiple queues is also an important limitation for many real applications. As mentioned before, only few results exist on polling models consisting of multiple servers.

In the area of mobile and wireless networks, a promising means to increase bandwidth, and to enhance robustness of the user-perceived service quality, is to make use of multiple antennas, enabling users to simultaneously utilize multiple networks (e.g., MIMO). In such environments, packet streams can be split over multiple parallel networks, each of which can be modeled as a polling model, where the packet routing may depend on the actual state of these networks. The packet-level analysis of such multi-network scenarios using this type of concurrent-access techniques will lead to the development of coupled polling models, where the arrival processes at the different polling models are correlated and state dependent. For this type of models, only few results are known today. Possibly, results from [116] (correlated arrivals) and [32] (correlated and state dependent arrivals) could be used to model these kind of networks.

4.2 Production systems

Although polling models appear naturally as a modeling tool in a wide variety of production applications, one can certainly observe common characteristics among these applications. The first similarity is the fact that a *high utilization of capacity* is typically prevalent due to either the magnitude of the setup times or the customer demands.

Secondly, in [92, 178] it is shown that assuming exponential interarrival times is not appropriate, whereas the independence assumptions do appear to be valid. Although these studies represent only a minuscule sample of production systems, they do show that the analysis of queueing systems with *general (renewal) arrival processes* is of practical interest for production applications.

Thirdly, in many other production systems the single most important performance measure is not an aggregate measure like the mean queue length, rather the probability that the queue length exceeds a pre-defined threshold.

In view of these three observations, the importance of explicit expressions of the complete

queue length distribution in the asymptotic regimes of high utilization of capacity (due to either the presence of setup times or the customer demands) for general arrival processes is evident.

4.3 Traffic and transportation systems

Although traffic intersections with vehicle actuated signals are typical examples of polling systems, there is still a lot of research to be done in this area before a realistic model is obtained. In traffic and transportation systems the assumption of independent interarrival times is reasonable for isolated intersections, but not for intersections that are part of an arterial system. Just like in the area of communication and production systems, it is necessary to develop techniques to analyse polling systems with other arrival assumptions, in particular platooned arrivals. For practical purposes there is a great need of good models for intersections that are part of an arterial system, i.e., the output of a certain queue forms the input process of a queue in a neighboring polling system. The assumption of independent interarrival times is strongly violated in such systems. One could consider the study of Reiman and Wein [134] as an important first step in this direction. They study waiting times of customers in a sequence of two-queue, exhaustive polling systems under heavy traffic conditions. Obviously, the analysis of networks of polling systems would not only be of interest for traffic and transportation systems, but also for other application areas.

The assumption of independent service times is not realistic either in most cases. It is also of great importance that models are developed for situations where multiple (conflicting and/or non-conflicting) traffic flows face a green light simultaneously. Some initial results on this topic are obtained in [19, 23], but (also here) many challenges remain for future research. Finally, and also this is not just important in the area of transportation, it is important to obtain good waiting-time approximations for time-limited systems.

Another area in transportation systems where few results have been obtained, is a system with one or multiple elevators. It would be of much practical relevance to have a realistic (polling) model that can be used to find optimal settings for the visiting policies of elevators. Difficulties that have to be addressed in elevator polling are the different routing schemes, but also the fact that each idle elevator returns to the floor that is its home base.

Acknowledgments

The authors are very grateful to Ivo Adan and Onno Boxma for providing valuable comments on earlier drafts of the present paper.

References

- [1] M. Ajmone Marsan, S. Donatelli, F. Neri, and G. Fantini. Some numerical results on finite-buffer markovian multiserver polling systems. In: *IEEE International Telecommunications Symposium, Rio de Janeiro, Brazil*, 1994.
- [2] A. S. Alfa and M. F. Neuts. Modelling vehicular traffic using the discrete time Markovian arrival process. *Transportation Science*, 29(2):109–117, 1995.

- [3] E. Altman, J. P. C. Blanc, A. Khamisy, and U. Yechiali. Gated-type polling systems with walking and switch-in times. *Stochastic Models*, 10:741–763, 1994.
- [4] E. Altman and D. Fiems. Expected waiting time in symmetric polling systems with correlated walking times. *Queueing Systems*, 56(3–4):241–253, 2007.
- [5] E. Altman, A. Khamisy, and U. Yechiali. On Elevator polling with globally gated regime. *Queueing Systems*, 11:85–90, 1992.
- [6] E. Altman and D. Kofman. Bounds for performance measures of token rings. *IEEE/ACM Transactions on Networking*, 4(2):292–299, 1996.
- [7] E. Altman and U. Yechiali. Polling in a closed network. *Probability in the Engineering and Informational Sciences*, 8(3):327–343, 1994.
- [8] R. Armony and U. Yechiali. Polling systems with permanent and transient jobs. *Stochastic Models*, 15(3):395–427, 1999.
- [9] J. E. Baker and I. Rubin. Polling with a general-service order table. *IEEE Transactions on Communications*, 35(3):283–288, 1987.
- [10] J. J. Bartholdi, III and L. K. Platzman. Decentralized control of Automated Guided Vehicles on a simple loop. *IEEE Transactions*, 21:76–81, 1989.
- [11] P. Beekhuizen. *Performance Analysis of Networks on Chips*. PhD thesis, Eindhoven University of Technology, 2010.
- [12] L. Benini and G. De Micheli. Networks on chips: a new SoC paradigm. *Computer*, 35(1):70–78, 2002.
- [13] S. Benjaafar and D. Gupta. Workload allocation in multi-product, multi-facility production systems. *IEEE Transactions*, 31(4):339–352, 1999.
- [14] D. Bertsimas and G. Mourtzinou. Multiclass queueing systems in heavy traffic: an asymptotic approach based on distributional and conservation laws. *Operations Research*, 45:470–487, 1997.
- [15] D. Bertsimas and G. Mourtzinou. Decomposition results for general polling systems and their applications. *Queueing Systems*, 31:295–316, 1999.
- [16] D. Bertsimas and H. Xu. Optimization of polling systems and dynamic vehicle routing problems on networks. Working paper, Operations Research Center, MIT, 1993.
- [17] C. Bisdikian. A queueing model for a data station within the IEEE 802.6 MAN. In: *Proceedings 17th Conference on Local Computer Networks*, 1992, pages 52–61, 1992.
- [18] T. Bonald and A. Proutière. Wireless downlink data channels: user performance and cell dimensioning. In: *Proceedings Mobicom '03*, San Diego, USA, 2003.
- [19] M. A. A. Boon. *Polling Models - From Theory to Traffic Intersections*. PhD thesis, Eindhoven University of Technology, 2011.
- [20] M. A. A. Boon and I. J. B. F. Adan. Mixed gated/exhaustive service in a polling model with priorities. *Queueing Systems*, 63:383–399, 2009.
- [21] M. A. A. Boon, I. J. B. F. Adan, and O. J. Boxma. A polling model with multiple priority levels. *Performance Evaluation*, 67:468–484, 2010.

- [22] M. A. A. Boon, I. J. B. F. Adan, and O. J. Boxma. A two-queue polling model with two priority levels in the first queue. *Discrete Event Dynamic Systems*, 20(4):511–536, 2010.
- [23] M. A. A. Boon, I. J. B. F. Adan, E. M. M. Winands, and D. G. Down. Delays at signalised intersections with exhaustive traffic control. EURANDOM report 2010-043, EURANDOM, 2010.
- [24] M. A. A. Boon, E. M. M. Winands, I. J. B. F. Adan, and A. C. C. van Wijk. Closed-form waiting time approximations for polling systems. *Performance Evaluation*, 68:290–306, 2011.
- [25] S. C. Borst. Polling systems with multiple coupled servers. *Queueing Systems*, 20:369–393, 1995.
- [26] S. C. Borst. *Polling Systems*, volume 115 of *CWI Tracts*. 1996.
- [27] S. C. Borst. User-level performance of channel-aware scheduling algorithms in wireless data networks. In: *Proceedings Infocom '03*, San Francisco, USA, 2003.
- [28] S. C. Borst and O. J. Boxma. Polling systems with and without switch-over times. *Operations Research*, 45:536–543, 1997.
- [29] S. C. Borst and R. D. van der Mei. Waiting-time approximations for multiple-server polling systems. *Performance Evaluation*, 31:163–182, 1998.
- [30] O. J. Boxma, J. Bruin, and B. H. Fralix. Waiting times in polling systems with various service disciplines. *Performance Evaluation*, 66:621–639, 2009.
- [31] O. J. Boxma and W. P. Groenendijk. Pseudo-conservation laws in cyclic-service systems. *Journal of Applied Probability*, 24(4):949–964, 1987.
- [32] O. J. Boxma, J. Ivanovs, K. Kosiński, and M. Mandjes. Lévy-driven polling systems and continuous-state branching processes. EURANDOM report 2009-026, EURANDOM, 2009.
- [33] O. J. Boxma, H. Levy, and J. A. Weststrate. Optimization of polling systems. In: P. J. B. King, I. Mitrani, and R. J. Pooley (Eds.), *Proceedings Performance '90*. North-Holland, Amsterdam, 1990.
- [34] O. J. Boxma, H. Levy, and U. Yechiali. Cyclic reservation schemes for efficient operation of multiple-queue single-server systems. *Annals of Operations Research*, 35(3):187–208, 1992.
- [35] O. J. Boxma, J. van der Wal, and U. Yechiali. Polling with batch service. *Stochastic Models*, 24(4):604–625, 2008.
- [36] O. J. Boxma and J. A. Weststrate. Waiting times in polling systems with Markovian server routing. In: G. Stiege and J. S. Lie (Eds.), *Messung, Modellierung und Bewertung von Rechen-systemen und Netzen*, pages 89–105. Springer Verlag, Berlin, 1989.
- [37] Y. A. Bozer and M. M. Srinivasan. Tandem configurations for Automated Guided Vehicle systems and the analysis of single vehicle loops. *IIE Transactions*, 23:72–82, 1991.
- [38] S. Browne and U. Yechiali. Dynamic server routing in binomial-gated, binomial-exhaustive, bernoulli-gated, bernoulli exhaustive and mixed polling systems. Technical report, Dept. of Statistics & Operations Research, Tel-Aviv University, 1990.
- [39] S. Browne and U. Yechiali. Dynamic scheduling in single server multi-class service systems with unit buffers. *Naval Research Logistics*, 38:383–396, 1991.
- [40] J. Bruin. Cyclic multi-item production systems. In: *Proceedings of Analysis of Manufacturing Systems*, pages 99–104, 2007.

- [41] B. D. Bunday and W. K. El-Badri. The efficiency of M groups of N machines served by a travelling robot: comparison of two models. *International Journal of Production Research*, 26:299–308, 1988.
- [42] W. Bux. Local-area subnetworks: a performance comparison. *IEEE Transactions on Communications*, 29(10):1465–1473, 1981.
- [43] M. Cicin-Sain, C. E. M. Pearce, and J. Sunde. On the application of a polling model with non-zero walk times and priority processing to a medical emergency-room environment. *Proceedings 23rd International Conference on Information Technology Interfaces, 2001*, 1:49–56, 2001.
- [44] E. G. Coffman Jr., G. Fayolle, and I. Mitrani. Two queues with alternating service periods. In: *Proceedings 12th IFIP WG 7.3 International Symposium on Computer Performance Modelling, Measurement and Evaluation*, pages 227–239. North-Holland, Amsterdam, 1987.
- [45] J. W. Cohen. *The Single Server Queue*. North-Holland, Amsterdam, revised edition, 1982.
- [46] R. B. Cooper. Queues served in cyclic order: waiting times. *The Bell System Technical Journal*, 49(3):399–413, 1970.
- [47] R. B. Cooper and G. Murray. Queues served in cyclic order. *The Bell System Technical Journal*, 48(3):675–689, 1969.
- [48] O. Czerniak, E. Altman, and U. Yechiali. Analysis of a TCP system under polling-type reduction-signal procedures. In: *Proceedings of ValueTools 2009, Pisa, Italy, 2009*.
- [49] O. Czerniak and U. Yechiali. Fluid polling systems. *Queueing Systems*, 63:401–435, 2009.
- [50] C. Daganzo. Some properties of polling systems. *Queueing Systems*, 6(2):137–154, 1990.
- [51] W. J. Dally and B. Towles. Route packets, not wires: on-chip interconnection networks. In: *Proceedings of the Design Automation Conference*, pages 684–689, 2001.
- [52] J. N. Darroch, G. F. Newell, and R. W. J. Morris. Queues for a vehicle-actuated traffic light. *Operations Research*, 12(6):882–895, 1964.
- [53] R. de Haan. *Queueing Models for Mobile Adhoc Networks*. PhD thesis, University of Twente, 2009.
- [54] R. de Haan, R. J. Boucherie, and J.-K. van Ommeren. A polling model with an autonomous server. *Queueing Systems*, 62(3):279–308, 2009.
- [55] E. De Souza e Silva, H. R. Gail, and R. R. Muntz. Polling systems with server timeouts and their application to token passing networks. *IEEE/ACM Transactions on Networking*, 3(5):560–575, 1995.
- [56] N. P. Dellaert. *Production To Order: Models and Rules for Production Planning*, volume 333 of *Lecture Notes in Economics and Mathematical Systems*. Springer Verlag, Berlin, 1989.
- [57] J. L. Dorsman, R. D. van der Mei, and E. M. M. Winands. A new method for deriving waiting-time approximations in polling systems with renewal arrivals. *To appear in Stochastic Models*, 2011.
- [58] I. M. Duenyas. The limitations of suboptimal policies. *Interfaces*, 24(5):77–84, 1994.
- [59] I. M. Duenyas and P. Van Oyen. Multi-product systems with both setup times and costs: Fluid bounds and schedules. *Queueing Systems*, 19:421–444, 1995.

- [60] I. M. Duenyas and P. Van Oyen. Heuristic scheduling of parallel heterogeneous queues with setups. *Management Science*, 42:814–829, 1996.
- [61] I. M. Dukhovnyy. An approximate model of motion of urban passenger transportation over annular routes. *Engineering Cybernetics*, 17:161–162, 1979.
- [62] M. Ebben, D.-J. van der Zee, and M. van der Heijden. Dynamic one-way traffic control in automated transportation systems. *Transportation Research Part B*, 38:441–458, 2004.
- [63] M. Eisenberg. The polling system with a stopping server. *Queueing Systems*, 18(3):387–431, 1994.
- [64] I. Eliazar. The snowblower problem. *Queueing Systems*, 45(4):357–380, 2003.
- [65] I. Eliazar. From polling to snowplowing. *Queueing Systems*, 51(1–2):115–133, 2005.
- [66] N. Erkip, R. Güllü, and A. Kocabiyikoglu. A quasi-birth-and-death model to evaluate fixed cycle time policies for stochastic multi-item production/inventory problem. In: *Proceedings of MSOM conference*, 2000.
- [67] A. Federgruen and Z. Katalan. Customer waiting-time distributions under base-stock policies in single-facility multi-item production systems. *Naval Research Logistics*, 43:533–548, 1996.
- [68] A. Federgruen and Z. Katalan. The impact of setup times on the performance of multiclass service and production systems. *Operations Research*, 44:989–1001, 1996.
- [69] A. Federgruen and Z. Katalan. The stochastic Economic Lot Scheduling Problem: cyclical base-stock policies with idle times. *Management Science*, 42(6):783–796, 1996.
- [70] A. Federgruen and Z. Katalan. Determining production schedules under base-stock policies in single facility multi-item production systems. *Operations Research*, 46(6):883–898, 1998.
- [71] I. Frigui. *Analysis of a time-limited polling system with Markovian Arrival Process and phase type service*. PhD thesis, University of Manitoba, Winnipeg, 1997.
- [72] I. Frigui and A. S. Alfa. Analysis of a time-limited polling system. *Computer Communications*, 21:558–571, 1998.
- [73] I. Frigui and A. S. Alfa. Analysis of a discrete time table polling system with MAP input and time-limited service discipline. *Telecommunication Systems*, 12:51–77, 1999.
- [74] S. W. Fuhrmann. Performance analysis of a class of cyclic schedules. Technical Memorandum 81-59531-1, Bell Laboratories, 1981.
- [75] S. W. Fuhrmann and R. B. Cooper. Stochastic decompositions in the $M/G/1$ queue with generalized vacations. *Operations Research*, 33(5):1117–1129, 1985.
- [76] B. Gamse and G. F. Newell. An analysis of elevator operations in moderate high buildings - I. single elevator. *Transportation Research Part B*, 16(4):303–319, 1982.
- [77] B. Gamse and G. F. Newell. An analysis of elevator operations in moderate high buildings - II. multiple elevators. *Transportation Research Part B*, 16(4):321–335, 1982.
- [78] A. D. Gandhi and C. G. Cassandras. Optimal control of polling models for transportation applications. *Mathematical and Computer Modelling*, 23(11–12):1–23, 1996.
- [79] T. Ganesharajah, N. G. Hall, and C. Sriskandarajah. Design and operational issues in AGV-served manufacturing systems. *Annals of Operations Research*, 76:109–154, 1998.

- [80] Y. Gerchak and Z. Zhang. The cheaper/faster-yet-more-expensive phenomenon: are Zangwill's "paradoxes" indeed paradoxical? *Interfaces*, 24(5):84–87, 1994.
- [81] R. B. L. M. Giezenaar. Ontwerp voor een productie- en voorraadstrategie voor de productieplant Laurox bij AKZO Nobel Chemicals te Deventer. Master's thesis, University of Twente (in Dutch), 1997.
- [82] Y. Gong and R. de Koster. A polling-based dynamic order picking system for online retailers. *IIE Transactions*, 40:1070–1082, 2008.
- [83] S. E. Grasman, T. L. Olsen, and J. R. Birge. Setting basestock levels in multi-product systems with setups and random yield. *IIE Transactions*, 40(12):1158–1170, 2008.
- [84] B. S. Greenberg, R. C. Leachman, and R. W. Wolff. Predicting dispatching delays on a low speed, single track railroad. *Transportation Science*, 22:31–38, 1988.
- [85] D. Grillo. Polling mechanism models in communication systems – some application examples. In: H. Takagi (Ed.), *Stochastic Analysis of Computer and Communication Systems*, pages 659–699. North-Holland, Amsterdam, 1990.
- [86] Y. Günlal and D. Gupta. A polling system with a patient server and state-dependent setup times. *IIE Transactions*, 29:479–480, 1997.
- [87] D. Gupta and M. M. Srinivasan. Polling systems with state-dependent setup times. *Queueing Systems*, 22:403–423, 1996.
- [88] R. Haijema and J. van der Wal. An MDP decomposition approach for traffic control at isolated signalized intersections. *Probability in the Engineering and Informational Sciences*, 22:587–602, 2007.
- [89] R. Hariharan, W. K. Ehrlich, P. K. Reeser, and R. D. van der Mei. Performance of web servers in a distributed computing environment. In: J. M. de Souza, N. L. S. da Fonseca, and E. D. de Souza e Silva (Eds.), *Teletraffic Engineering in the Internet Era*, pages 137–148, 2001.
- [90] D. Heidemann. Queue length and delay distributions at traffic signals. *Transportation Research Part B*, 28(5):377–389, 1994.
- [91] M. Hofri and K. W. Ross. On the optimal control of two queues with server setup times and its analysis. *SIAM Journal on Computing*, 16(2):399–42, 1987.
- [92] R. R. Inman. Empirical evaluation of exponential and independence assumptions in queueing models of manufacturing systems. *Production and Operations Management*, 8(4):409–432, 1999.
- [93] V. Kavitha and E. Altman. Queueing in space: design of message ferry routes in static adhoc networks. In: *Proceedings ITC21*, 2009.
- [94] J. Keilson and L. D. Servi. Oscillating random walk models for $GI/G/I$ vacation systems with Bernoulli schedules. *Journal of Applied Probability*, 23:790–802, 1986.
- [95] J. Keilson and L. D. Servi. The distributional form of Little's Law and the Fuhrmann-Cooper decomposition. *Operations Research Letters*, 9(4):239–247, 1990.
- [96] W. B. Kim and E. Königsberg. The efficiency of two groups of N machines derved by a single robot. *The Journal of the Operational Research Society*, 38(6):523–538, 1987.
- [97] L. Kleinrock. Performance evaluation of distributed computer-communication systems. In: O. J. Boxma and R. Syski (Eds.), *Queueing Theory and its Applications – Liber Amicorum for J.W. Cohen*, pages 1–57. North-Holland, Amsterdam, 1988.

- [98] L. Kleinrock and H. Levy. The analysis of random polling systems. *Journal of Operations Research*, 36(5):716–732, 1988.
- [99] L. Kleinrock and M. O. Scholl. Packet switching in radio channels: New conflict-free access schemes. *IEEE Transactions on Communication*, 28(7):1015–1029, 1980.
- [100] G. P. Klimov. Time-sharing service systems I. *Theory of Probability and its Applications*, 19:532–551, 1974.
- [101] A. G. Konheim, H. Levy, and M. M. Srinivasan. Descendant set: an efficient approach for the analysis of polling systems. *IEEE Transactions on Communications*, 42(2–4):1245–1253, 1994.
- [102] E. Königsberg and J. Mamer. The analysis of production systems. *International Journal of Production Research*, 20:1–16, 1982.
- [103] G. Kramer, B. Mukherjee, S. Dixit, Y. Yinghua, and R. Hirth. Supporting differentiated classes of services in ethernet passive optical networks. *Journal of Optical Networking*, 1(9):280–290, 2002.
- [104] G. Kramer, B. Mukherjee, and G. Pesavento. Ethernet PON (ePON): design and analysis of an optical access network. *Photonic Network Communications*, 3:307–319, 2001.
- [105] G. Kramer, B. Mukherjee, and G. Pesavento. Interleaved polling with adaptive cycle time (IPACT): a dynamic bandwidth allocation scheme in an optical access network. *Photonic Network Communications*, 4:89–107, 2002.
- [106] G. N. Krieg and H. Kuhn. A decomposition method for multi-product kanban systems with setup times and lost sales. *IIE Transactions*, 34(7):613–625, 2002.
- [107] G. N. Krieg and H. Kuhn. Analysis of multi-product kanban systems with state-dependent setups and lost sales. *Annals of Operations Research*, 125:141–166, 2004.
- [108] W.-M. Lan and T. L. Olsen. Multiproduct systems with both setup times and costs: Fluid bounds and schedules. *Operations Research*, 54(3):505–522, 2006.
- [109] D.-S. Lee. Analysis of a two-queue model with Bernoulli schedules. *Journal of Applied Probability*, 34:176–191, 1997.
- [110] J. P. Lehoczky. Traffic intersection control and zero-switch queues under conditions of Markov chain dependence input. *Journal of Applied Probability*, 9(2):382–395, 1972.
- [111] K. Leung. Cyclic service systems with non-preemptive, time-limited service. *IEEE Transactions on Communication*, 42(8):2521–2524, 1994.
- [112] K. K. Leung. Cyclic-service systems with probabilistically-limited service. *IEEE Journal on Selected Areas in Communications*, 9(2):185–193, 1991.
- [113] H. Levy. Optimization of polling systems: the fractional exhaustive service method. Report, Tel-Aviv University, 1988.
- [114] H. Levy. Analysis of cyclic polling systems with binomial gated service. In: T. Hasegawa, H. Takagi, and Y. Takahashi (Eds.), *Performance of Distributed and Parallel Systems*, pages 127–139. North-Holland, Amsterdam, 1989.
- [115] H. Levy and M. Sidi. Polling systems: applications, modeling, and optimization. *IEEE Transactions on Communications*, 38:1750–1760, 1990.

- [116] H. Levy and M. Sidi. Polling systems with simultaneous arrivals. *IEEE Transactions on Communications*, 39(6):823–827, 1991.
- [117] C. Mack. The efficiency of N machines uni-directionally patrolled by one operative when walking time is constant and repair times are variable. *Journal of the Royal Statistical Society Series B*, 19(1):173–178, 1957.
- [118] C. Mack, T. Murphy, and N. L. Webb. The efficiency of N machines uni-directionally patrolled by one operative when walking time and repair times are constants. *Journal of the Royal Statistical Society Series B*, 19(1):166–172, 1957.
- [119] D. R. Manfield. Analysis of a priority polling system for two-way traffic. *IEEE Transactions on Communications*, 33:1001–1006, 1985.
- [120] D. M. Markowitz, M. I. Reiman, and L. M. Wein. The stochastic economic lot scheduling problem: heavy traffic analysis of dynamic cyclic policies. *Operations Research*, 48(1):136–154, 2000.
- [121] D. M. Markowitz and L. M. Wein. Heavy traffic analysis of dynamic cyclic policies: a unified treatment of the single machine scheduling problem. *Operations Research*, 49(2):246–270, 2001.
- [122] B. McIntyre. A comment on Zangwill’s “the limits of Japanese production theory”. *Interfaces*, 24(5):87–89, 1994.
- [123] A. J. Miller. Settings for fixed-cycle traffic signals. *Operational Research Quarterly*, 14:373–386, 1963.
- [124] D. Miorandi, A. Zanella, and G. Pierobon. Performance evaluation of bluetooth polling schemes: An analytical approach. *Mobile Networks and Applications*, 9:63–72, 2004.
- [125] R. J. T. Morris and Y. T. Wang. Some results for multi-queue systems with multiple cyclic servers. In: W. Bux and H. Rudin (Eds.), *Performance of Computer-Communication Systems*, pages 245–258. North-Holland, Amsterdam, 1984.
- [126] S. Nahmias and M. H. Rothkopf. Stochastic models of internal mail delivery systems. *Management Science*, 30(9):1113–1120, 1984.
- [127] K. B. Narasimha and S. Bhattacharya. Lead time minimization of a multi-product, single-processor system: A comparison of cyclic policies. *International Journal of Production Economics*, 106(1):28–40, 2007.
- [128] G. F. Newell. Approximation methods for queues with application to the fixed-cycle traffic light. *SIAM Review*, 7(2):223–240, 1965.
- [129] G. F. Newell. Properties of vehicle-actuated signals: I. one-way streets. *Transportation Science*, 3(2):31–52, 1969.
- [130] G. F. Newell. The rolling horizon scheme of traffic signal control. *Transportation Research Part A*, 32(1):39–44, 1998.
- [131] G. F. Newell and E. E. Osuna. Properties of vehicle-actuated signals: II. two-way streets. *Transportation Science*, 3(2):99–125, 1969.
- [132] T. L. Olsen. A practical scheduling method for multiclass production systems with setups. *Management Science*, 45(1):116–130, 1999.
- [133] T. L. Olsen and R. D. van der Mei. Periodic polling systems in heavy-traffic: renewal arrivals. *Operations Research Letters*, 33:17–25, 2005.

- [134] M. I. Reiman and L. M. Wein. Heavy traffic analysis of polling systems in tandem. *Operations Research*, 47(4):524–534, 1999.
- [135] J. A. C. Resing. Polling systems and multitype branching processes. *Queueing Systems*, 13:409–426, 1993.
- [136] Z. Saffer and M. Telek. Unified analysis of $BMAP/G/1$ cyclic polling models. *Queueing Systems*, 64(1):69–102, 2010.
- [137] S. Samaddar and C. A. Hill. Controlling adverse effect on work in process inventory while reducing machine setup time. *European Journal of Operational Research*, 180(1):249–261, 2007.
- [138] S. Samaddar and T. Whalen. Improving performance in cyclic production systems by using forced variable idle setup time. *Manufacturing & Service Operations Management*, 10(2):173–180, 2008.
- [139] D. Sarkar and W. I. Zangwill. Variance effects in cyclic production systems. *Management Science*, 37:444–453, 1991.
- [140] D. Sarkar and W. I. Zangwill. File and work transfers in cyclic queue systems. *Management Science*, 38(10):1510–1523, 1992.
- [141] K. C. Sevcik and M. J. Johnson. Cycle time properties of the FDDI token ring protocol. *IEEE Transactions on Software Engineering*, 13:376–385, 1987.
- [142] M. Sharafali, H. C. Co, and M. Goh. Production scheduling in a flexible manufacturing system under random demand. *European Journal of Operational Research*, 158(1):89–102, 2004.
- [143] M. Sidi and H. Levy. Customer routing in polling systems. In: P. King, I. Mitrani, and R. Pooley (Eds.), *Proceedings Performance '90*, pages 319–331. North-Holland, Amsterdam, 1990.
- [144] M. P. Singh and M. M. Srinivasan. Exact analysis of the state dependent polling model. *Queueing Systems*, 41:371–399, 2002.
- [145] M. P. Singh and M. M. Srinivasan. Performance bounds for flexible systems requiring setups. *Management Science*, 53(6):991–1004, 2007.
- [146] M. M. Srinivasan. Non-deterministic polling systems. *Management Science*, 37:667–681, 1991.
- [147] M. M. Srinivasan, Y. A. Bozer, and M. Cho. Trip-based material handling systems: throughput capacity analysis. *IIE Transactions*, 26:70–89, 1994.
- [148] H. Takagi. *Analysis of polling systems*. MIT Press (Cambridge, Mass.), 1986.
- [149] H. Takagi. Queueing analysis of polling models: an update. In: H. Takagi (Ed.), *Stochastic Analysis of Computer and Communication Systems*, pages 267–318. North-Holland, Amsterdam, 1990.
- [150] H. Takagi. Analysis of finite-capacity polling systems. *Advances in Applied Probability*, 23(2):373–387, 1991.
- [151] H. Takagi. Queueing analysis of polling models: progress in 1990–1994. In: J. H. Dshalalow (Ed.), *Frontiers in Queueing: Models, Methods and Problems*, pages 119–146. CRC Press, Boca Raton, 1997.
- [152] H. Takagi. Analysis and application of polling models. In: G. Haring, C. Lindemann, and M. Reiser (Eds.), *Performance Evaluation: Origins and Directions*, volume 1769 of *Lecture Notes in Computer Science*, pages 424–442. Springer, Berlin, 2000.

- [153] B. van Arem. *Queueing Network Models for Slotted Transmission Systems*. PhD thesis, University of Twente, The Netherlands, 1990.
- [154] J. L. van den Berg, R. Litjens, and J. Laverman. HSDPA flow level performance: the impact of key system and traffic aspects. In: *Proceedings MSWiM '04 (Venice, Italy)*, 2004.
- [155] M. S. van den Broek, J. S. H. van Leeuwaarden, I. J. B. F. Adan, and O. J. Boxma. Bounds and approximations for the fixed-cycle traffic-light queue. *Transportation Science*, 40(4):484–496, 2006.
- [156] M. van der Heijden, A. van Harten, and M. Ebben. Waiting times at periodically switched one-way traffic lanes. *Probability in the Engineering and Informational Sciences*, 15:495–517, 2001.
- [157] R. D. van der Mei and S. C. Borst. Analysis of multiple-server polling systems by means of the power-series algorithm. *Communications in Statistics – Stochastic Models*, 13(2):339–369, 1997.
- [158] R. D. van der Mei, R. Hariharan, and P. K. Reeser. Web server performance modeling. *Telecommunication Systems*, 16:361–378, 2001.
- [159] R. D. van der Mei and J. A. C. Resing. Analysis of polling models with two-stage gated service: fairness versus efficiency. In: L. Mason, T. Drwiega, and J. Yan (Eds.), *Managing Traffic Performance in Converged Networks - the Interplay between Convergent and Divergent Forces*, ITC2007, Lecture Notes in Computer Science 4516, pages 544–555, 2007.
- [160] R. D. van der Mei and J. A. C. Resing. Polling models with two-phase gated service: heavy-traffic results for the waiting-time distributions. *Probability in the Engineering and Informational Sciences*, 22:623–651, 2008.
- [161] R. D. van der Mei and A. Roubos. Polling models with multi-phase gated service. *To appear in Annals of Operations Research*, 2011.
- [162] R. D. van der Mei and E. M. M. Winands. Polling models with renewal arrivals: A new method to derive heavy-traffic asymptotics. *Performance Evaluation*, 64(9–12):1029–1040, 2007.
- [163] R. D. van der Mei and E. M. M. Winands. A note on polling models with renewal arrivals and nonzero switch-over times. *Operations Research Letters*, 36:500–505, 2008.
- [164] J. van der Wal and U. Yechiali. Dynamic visit-order rules for batch-service polling. *Probability in the Engineering and Informational Sciences*, 17:351–367, 2003.
- [165] B. van Houdt. Numerical solution of polling systems for analyzing networks on chips. In: *Proceedings of NSMC 2010, Williamsburg (USA)*, pages 90–93, 2010.
- [166] J. S. H. van Leeuwaarden. Delay analysis for the fixed-cycle traffic-light queue. *Transportation Science*, 40(2):189–199, 2006.
- [167] M. van Vuuren and E. M. M. Winands. Iterative approximation of k -limited polling systems. *Queueing Systems*, 55(3):161–178, 2007.
- [168] V. M. Vishnevskii and O. V. Semenova. Mathematical methods to study the polling systems. *Automation and Remote Control*, 67(2):173–220, 2006.
- [169] M. Vlasiou, I. J. B. F. Adan, and O. J. Boxma. A two-station queue with dependent preparation and service times. *European Journal of Operations Research*, 195:104–116, 2009.
- [170] M. Vlasiou and U. Yechiali. $M/G/\infty$ polling systems with random visit times. *Probability in the Engineering and Informational Sciences*, 22:81–106, 2008.

- [171] F. V. Webster. Traffic signal settings. Technical Paper 39, Road Research Laboratory, 1958.
- [172] J. A. Weststrate. *Analysis and Optimization of Polling Systems*. PhD thesis, Tilburg University, 1992.
- [173] A. Wierman, E. M. M. Winands, and O. J. Boxma. Scheduling in polling systems. *Performance Evaluation*, 64:1009–1028, 2007.
- [174] E. M. M. Winands. *Polling, Production & Priorities*. PhD thesis, Eindhoven University of Technology, 2007.
- [175] E. M. M. Winands. Branching-type polling systems with large setups. *OR Spectrum*, 33(1):77–97, 2011.
- [176] E. M. M. Winands, I. J. B. F. Adan, and G.-J. van Houtum. Mean value analysis for polling systems. *Queueing Systems*, 54:35–44, 2006.
- [177] E. M. M. Winands, I. J. B. F. Adan, G. J. van Houtum, and D. G. Down. A state-dependent polling model with k -limited service. *Probability in the Engineering and Informational Sciences*, 23(2):385–408, 2009.
- [178] E. M. M. Winands, A. G. de Kok, and C. Timpe. Case study of a batch-production/inventory system. *Interfaces*, 39(6):552–554, 2009.
- [179] D. Xu, N. Shi, and R. K. Cheung. Heavy traffic analysis of a single vehicle loop in an automated storage and retrieval system. *OR Spectrum*, 29(3):489–512, 2007.
- [180] H. Yamashita, Y. Ishizuka, and S. Suzuki. Mean and variance of waiting time and their optimization for alternating traffic control systems. *Mathematical Programming*, 108(2–3):419–433, 2006.
- [181] U. Yechiali. Optimal dynamic control of polling systems. In: J. W. Cohen and C. D. Pack (Eds.), *Proceedings ITC13, Workshop: Queueing, Performance and Control in ATM*, pages 205–218. North-Holland, Amsterdam, 1991.
- [182] U. Yechiali. Analysis and control of poling systems. In: L. Donatiello and R. Nelson (Eds.), *Performance Evaluation of Computer and Communication Systems*, pages 630–650. Springer Verlag, 1993.
- [183] W. I. Zangwill. The limits of Japanese production theory. *Interfaces*, 22(5):14–25, 1992.
- [184] W. I. Zangwill. Response to comments on our work by Duenyas, by Gerchak and Zhang, and by McIntyre. *Interfaces*, 24(5):90–94, 1994.
- [185] G. Zussman, A. Segall, and U. Yechiali. On the analysis of the bluetooth time division duplex mechanism. *IEEE Transactions on Wireless Communications*, 6(6):2149–2161, 2007.