

# Mean value analysis for polling systems

E.M.M. Winands<sup>1,2</sup>, I.J.B.F. Adan<sup>1</sup> and G.J. van Houtum<sup>2</sup>

<sup>1</sup>Department of Mathematics and Computer Science

<sup>2</sup>Department of Technology Management

Technische Universiteit Eindhoven

P.O. Box 513, 5600 MB Eindhoven, The Netherlands

{e.m.m.winands,i.j.b.f.adan,g.j.v.houtum}@tue.nl

September 16, 2005

## Abstract

The present paper deals with the problem of calculating mean delays in polling systems with either exhaustive or gated service. We develop a mean value analysis (MVA) to compute these delay figures. The merits of MVA are in its intrinsic simplicity and its intuitively appealing derivation. As a consequence, MVA may be applied, both in an exact and approximate manner, to a large variety of models.

**Keywords:** polling systems, mean value analysis, exhaustive service, gated service.

# 1 Introduction

A typical polling system consists of a number  $N$  of queues, attended by a single server in a fixed order. There is a huge body of literature on polling systems that has continued to grow since the late 1950s, when the papers of [15, 16] concerning a patrolling repairman model for the British cotton industry were published. Polling systems have a wide range of applications in communication, production, transportation and maintenance systems. Excellent surveys on polling systems and their applications may be found in [22, 23, 24] and in [13].

The present paper is concerned with the two service disciplines most commonly used in polling systems, the so-called *exhaustive* and *gated* policies. Exhaustive means that a queue must be empty before the server moves on, whereas in case of gated service only those customers in the queue at the polling start are served. The single most important performance measure for polling systems is, in many applications, the mean delay of a customer. Unfortunately, explicit closed-form expressions for the mean delays in polling systems with exhaustive-type or gated-type service are only known in very special cases. However, in the past several numerical approaches have been proposed for computing these mean delays in general continuous-time polling systems with either exhaustive or gated service.

One such method is the *buffer occupancy* method as developed by [3, 4, 5]. This method is based on the buffer occupancy variables  $X_{i,j}$ , which denote the queue length at queue  $j$  at a polling instant of queue  $i$ ,  $i, j = 1, 2, \dots, N$ . The buffer occupancy method requires the solution of  $N^3$  linear equations with unknowns  $\mathbb{E}[X_{ij}X_{jk}]$  to compute the mean delays in *all*  $N$  stations simultaneously. These equations may be efficiently solved in an iterative manner requiring  $O(N^3 \log_\rho \epsilon)$  operations (additions and multiplications), where  $\rho$  is the total occupation rate and  $\epsilon$  is the relative accuracy required [12]. Based on this buffer occupancy method, [11] developed the *descendant set* method; an iterative technique that computes the mean delay at each queue independently of the other queues. The descendant set approach is based on counting the number of descendants of each customer in the system. This method requires  $O(N \log_\rho \epsilon)$  operations for the computation of the mean delay in a *single* station. A second well-known method based on the buffer occupancy method is the *individual station* technique [20], which also allows, as the name suggests, the individual computation of the mean delay at each queue. The individual station technique is, however, not an iterative approach. The mean delay at a *single* queue is computed in  $O(N^2)$  operations, which obviously does not depend on the system utilization contrary to the computational complexities of the aforementioned methods.

Besides the techniques based on the buffer occupancy method, another school of approaches are the ones embroidering on the *station time* method [6]. In the station time approach, all mean delays are obtained simultaneously starting from the station time variables  $U_i$ ,  $i = 1, 2, \dots, N$ . The station time  $U_i$  is composed of the time the server spends servicing customers at queue  $i$  plus the *preceding* setup time in case of exhaustive service or plus the *succeeding* setup time in case of gated service. The station time technique induces a set of  $N^2$  linear equations with unknowns  $\mathbb{E}[U_i U_j]$ , which can be solved iteratively in  $O(N^2 \log_\rho \epsilon)$  operations leading to *all*  $N$  mean delays. An extension of the station time method is the approach developed by [19]. Their approach induces a set of only  $N$  linear equations, which is, however, less sparse. Solving this set of equations requires  $O(N^3)$  operations for *all*  $N$  delay figures.

Recently, *Hirayama et al.* [9] developed a third alternative method for obtaining the mean delays. The authors analyze first the mean delays conditioned on the state of the system at

an arrival epoch. Then, from the analysis of the system at polling instants, a set of linear functional equations for these conditional delays is obtained. By applying a limiting procedure, they derive a set of  $N(N + 1)$  linear equations for the unconditional mean delays, which can be solved in  $O(N^6)$  operations. With respect to this computational complexity, it should be noted that the method of [9] shows some similarities with the buffer occupancy approach and that it, therefore, may be possible to construct more efficient iterative algorithms to solve their set of equations.

With respect to the above-mentioned methods, two issues are noteworthy. Firstly, each of the approaches can be readily adapted to a discrete-time counterpart, apart from some occasional subtleties (see, e.g., [10, 18, 21]). The second important observation is that when comparing the use of the aforementioned approaches in the open literature over the recent years, it immediately strikes the eye that the buffer occupancy method and its variations can be - or at least have been - applied to the widest variety of polling systems. In fact, this method appears to be applicable to the complete class of service disciplines satisfying a well-known branching property (see [8] and [17]). However, the techniques based on the station time method and Hirayama's method have been applied to a restricted class of polling systems only. For example, it is known that the station time method cannot be used in polling systems with *mixed service*, where some of the queues are served according to the exhaustive policy and some by the gated strategy.

The objective of the present paper is the development of a novel approach to compute the mean delays for exhaustive-type or gated-type polling systems in a pure probabilistic manner. More specifically, we derive a set of  $N^2$  and  $N(N + 1)$  linear equations for these delay figures in case of exhaustive and gated service, respectively, with the help of the following two basic queueing results: (i) the PASTA property, i.e., Poisson arrivals see time averages [25] and (ii) Little's Law [14]. The unknowns in these equations are  $\mathbb{E}[L_{i,j}]$ , the mean queue length at queue  $i$  at an arbitrary epoch within a station time, also called a visit time, of queue  $j$ . The method of the present paper can be looked upon as a *mean value analysis* (MVA) for general polling systems with exhaustive or gated service. MVA is known as a powerful tool to determine mean performance measures in all kinds of queueing models, but it has never been applied to polling systems.

The main contribution of the present paper is two-fold. The first main contribution can be found in the set of equations itself. In contrast to most of the above-mentioned approaches, the unknowns in these equations are all *first* moments of random variables and, thus, no correlation terms are required. Furthermore, MVA evaluates the polling system at arbitrary epochs in time and not on embedded points such as polling instants. An additional merit of MVA is the fact that it allows for an evaluation of polling systems with mixed service. Finally, MVA results in the solution of no more than  $N^2$  and  $N(N + 1)$  linear equations for *all*  $N$  mean delays in case of exhaustive and gated service, respectively. In the past, various efficient algorithms have been developed based on the buffer occupancy method and the station time approach, which require systems of up to  $N^3$  equations. It may, therefore, be possible to construct just as efficient iterative algorithms based on MVA. We emphasize that the development of such an algorithm is not within the scope of the present paper.

The second main contribution, which is perhaps even more important, lies in the derivation of the set of equations by MVA. This derivation is based on standard queueing results and has a probabilistic interpretation all the way. Consequently, it is rather straightforward to apply MVA to variants of the considered polling systems: (i) systems with Poisson batch arrivals, (ii) systems with fixed polling tables and (iii) discrete-time polling systems. Finally, MVA

may open new ways for the evaluation, both in an exact and approximate manner, of other polling systems.

The structure of the present paper is as follows. Section 2 defines the model and introduces further notation. Section 3 presents the main results of the paper: the derivation of a set of equations for the mean delay in polling systems with either exhaustive or gated service. The paper is wound up in Section 4 with conclusions and a list of possible extensions.

## 2 Model description and notation

We consider a system with one single server for  $N \geq 1$  queues, in which there is infinite buffer capacity for each queue. The server visits and serves the queues in a fixed cyclic order. We index the queues by  $i$ ,  $i = 1, 2, \dots, N$ , in the order of the server movement. For compactness of presentation, all references to queue indices greater than  $N$  or less than 1 are implicitly assumed to be modulo  $N$ , e.g., queue  $N + 1$  actually refers to queue 1.

Service at each queue is according to one of the following service policies:

- *Exhaustive* policy: when the server polls a queue, he serves its customers until that queue is empty;
- *Gated* policy: when the server polls a queue, he serves all, and only, customers found at the polling instant.

Customers arrive at all queues according to independent Poisson processes with rates  $\lambda_i$ ,  $i = 1, 2, \dots, N$ . The service times at queue  $i$  are independent, identically distributed random variables with mean  $\mathbb{E}[B_i]$  and second moment  $\mathbb{E}[B_i^2]$ ,  $i = 1, 2, \dots, N$ . When the server starts service at queue  $i$ , a setup time is incurred of which the first and second moment are denoted by  $\mathbb{E}[S_i]$  and  $\mathbb{E}[S_i^2]$ ,  $i = 1, 2, \dots, N$ , respectively. These setup times are identically distributed random variables, independent of any other event involved. In particular, they are independent of the service times. The mean total setup time  $\mathbb{E}[S]$  in a cycle is given by

$$\mathbb{E}[S] = \sum_{i=1}^N \mathbb{E}[S_i].$$

For further reference, we introduce the mean residual service time and the mean residual setup time for queue  $i$ , which can be expressed as follows, respectively,

$$\mathbb{E}[R_{B_i}] = \frac{\mathbb{E}[B_i^2]}{2\mathbb{E}[B_i]}, \quad \mathbb{E}[R_{S_i}] = \frac{\mathbb{E}[S_i^2]}{2\mathbb{E}[S_i]}, \quad i = 1, 2, \dots, N.$$

The occupation rate  $\rho_i$  (excluding setups) at queue  $i$  is defined by  $\rho_i = \lambda_i \mathbb{E}[B_i]$  and the total occupation rate  $\rho$  is given by

$$\rho = \sum_{i=1}^N \rho_i.$$

A necessary and sufficient condition for the stability of this polling system is obviously  $\rho < 1$  (see, e.g., [22]). In the remainder of the present paper, this stability condition is assumed to hold as we restrict ourselves to steady-state behavior.

We continue this section with some further notation. The cycle length of queue  $i$ ,  $i = 1, 2, \dots, N$ , is defined as the time between two successive arrivals of the server at this queue.

It is well-known that the mean cycle length is independent of the queue involved and is given by (see, e.g., [22])

$$\mathbb{E}[C] = \frac{\mathbb{E}[S]}{1 - \rho}.$$

We note that although the first moments of the cycle lengths are identical for all queues, higher moments generally differ. For presentation reasons, the present paper focusses mainly on the case  $\mathbb{E}[S] > 0$ . When the total setup time is equal to zero, some subtleties appear due to the fact that the number of cycles with zero length tends to infinity. However, with some minor adjustments MVA can still be applied as elaborated on in Subsection 3.3.

The visit time  $\theta_i$  of queue  $i$ ,  $i = 1, 2, \dots, N$ , is composed of the service period of queue  $i$ , the time the server spends servicing customers at queue  $i$ , plus the *preceding* setup time in case of exhaustive service or plus the *succeeding* setup time in case of gated service. By virtue of these two different definitions, a queue is empty exactly at the end of its visit time in case of exhaustive service, while the queue before the gate is empty at the beginning of a visit time in case of gated service (all customers waiting for service are then placed behind the gate).

Since the server is working a fraction  $\rho_i$  of the time on queue  $i$ , the mean of a visit period of queue  $i$  reads, for exhaustive service,

$$\mathbb{E}[\theta_i] = \rho_i \mathbb{E}[C] + \mathbb{E}[S_i], \quad i = 1, 2, \dots, N,$$

and, for gated service,

$$\mathbb{E}[\theta_i] = \rho_i \mathbb{E}[C] + \mathbb{E}[S_{i+1}], \quad i = 1, 2, \dots, N.$$

We define an  $(i, j)$ -period  $\theta_{i,j}$  as the sum of  $j$  consecutive visit times starting in queue  $i$ ,  $j = 1, 2, \dots, N$ . The corresponding mean is given by

$$\mathbb{E}[\theta_{i,j}] = \sum_{n=i}^{i+j-1} \mathbb{E}[\theta_n], \quad i = 1, 2, \dots, N, \quad j = 1, 2, \dots, N.$$

Notice that in case  $j = 1$  and  $j = N$ ,  $\mathbb{E}[\theta_{i,j}]$  is equal to the mean visit period  $\mathbb{E}[\theta_i]$  of queue  $i$  and the mean cycle length  $\mathbb{E}[C]$ , respectively.

The fraction of the time  $q_{i,j}$  the system is in an  $(i, j)$ -period equals

$$q_{i,j} = \frac{\mathbb{E}[\theta_{i,j}]}{\mathbb{E}[C]}, \quad i = 1, 2, \dots, N, \quad j = 1, 2, \dots, N,$$

where, by definition,  $q_{i,N}$  equals 1. Moreover, the mean of a residual  $(i, j)$ -period is given by

$$\mathbb{E}[R_{\theta_{i,j}}] = \frac{\mathbb{E}[\theta_{i,j}^2]}{2\mathbb{E}[\theta_{i,j}]}, \quad i = 1, 2, \dots, N, \quad j = 1, 2, \dots, N, \quad (1)$$

with the remark that the second moments  $\mathbb{E}[\theta_{i,j}^2]$  are still unknown at this stage. Notice that since the successive  $(i, j)$ -periods are dependent, they do not form a renewal process. This means, among others, that Equation (1) does not directly follow from the theory of regenerative processes. For a proof why this result is nevertheless still valid see, e.g., [7].

Our main interest is in the mean delay  $\mathbb{E}[W_i]$  of a type- $i$  customer,  $i = 1, 2, \dots, N$ , which is defined as the time in steady state from a customer's arrival at queue  $i$  until the start of his

service. By Little's Law, these mean delays are obviously related to the mean queue lengths (excluding the customer possibly in service)  $\mathbb{E}[L_i]$ ,  $i = 1, 2, \dots, N$ . The analysis of the present paper is oriented towards the determination of  $\mathbb{E}[L_{i,j}]$ , the mean queue length at queue  $i$  at an arbitrary epoch within a visit time of queue  $j$ ,  $i, j = 1, 2, \dots, N$ . The corresponding unconditional mean queue length  $\mathbb{E}[L_i]$  can be expressed in terms of  $\mathbb{E}[L_{i,j}]$  as follows

$$\mathbb{E}[L_i] = \sum_{n=1}^N q_{n,1} \mathbb{E}[L_{i,n}], \quad i = 1, 2, \dots, N. \quad (2)$$

Finally, for a broad class of polling systems linear relationships among the  $\mathbb{E}[W_i]$ , the so-called *pseudoconservation* law, have been derived [2]. In case of exhaustive or gated service, this law reduces to

$$\sum_{i=1}^N \rho_i \mathbb{E}[W_i] = \rho \sum_{i=1}^N \frac{\lambda_i \mathbb{E}[B_i^2]}{2(1-\rho)} + \frac{\rho}{2\mathbb{E}[S]} \sum_{i=1}^N (\mathbb{E}[S_i^2] - \mathbb{E}[S_i]^2) + \frac{\mathbb{E}[S]}{2(1-\rho)} \sum_{i=1}^N \rho_i (1-\rho_i) + \frac{\mathbb{E}[S]}{1-\rho} \sum_{i \in G} \rho_i^2, \quad (3)$$

where  $G$  stands for the index set of queues with gated service. Although this pseudoconservation law does not give explicit expressions for the mean delays themselves, it appears to be a useful tool for developing approximations. Furthermore, it provides a relatively simple expression for the weighted sum of the mean delays, which may be used as a first indication of overall system performance.

### 3 Mean value analysis

In the present section, we derive sets of linear equations in case of either exhaustive or gated service to compute the mean delays involving  $\mathbb{E}[L_{i,j}]$  as unknowns. We present these derivations, for reasons of clarity, first for *pure* exhaustive (Subsection 3.1) and *pure* gated polling systems (Subsection 3.2) with nonzero setup times. In Subsection 3.3, it is then shown that the application of MVA to systems with mixed service and systems with zero setup times is rather straightforward.

#### 3.1 Exhaustive service

The goal of the present subsection is the derivation of the mean delays in exhaustive polling systems by using MVA. This typically starts with the derivation of a so-called *arrival relation* with the help of the PASTA property. Therefore, consider a tagged customer at the moment he arrives at queue  $i$ ,  $i = 1, 2, \dots, N$ . Based on PASTA, we know that the state distribution seen by this tagged customer is identical to the equilibrium distribution. That is, this customer has to wait for the servicing of all customers  $L_i$ , who were already waiting in this queue on his arrival. Further, with probability  $\rho_i$  the server is working at queue  $i$  on his arrival and the tagged customer has to wait for the residual service time of the customer in service as well. On the other hand, with probability  $\mathbb{E}[S_i]/\mathbb{E}[C]$  the server is in a setup phase for queue  $i$  and the delay of the customer is increased by a residual setup time. Finally, with probability  $1 - q_{i,1}$  the server is at one of the other queues and the service of the tagged customer is delayed until the server starts service again at queue  $i$ . The latter time period is obviously equal to the sum of a residual  $(i + 1, N - 1)$ -period and a setup time for queue  $i$ .

Hence, we have the following arrival relation for the mean delay  $\mathbb{E}[W_i]$  of a type- $i$  customer,  $i = 1, 2, \dots, N$ ,

$$\mathbb{E}[W_i] = \mathbb{E}[L_i]\mathbb{E}[B_i] + \rho_i\mathbb{E}[R_{B_i}] + \frac{\mathbb{E}[S_i]}{\mathbb{E}[C]}\mathbb{E}[R_{S_i}] + (1 - q_{i,1})\left(\mathbb{E}[R_{\theta_{i+1,N-1}}] + \mathbb{E}[S_i]\right).$$

Application of Little's Law,

$$\mathbb{E}[L_i] = \lambda_i\mathbb{E}[W_i], \quad i = 1, 2, \dots, N, \quad (4)$$

yields

$$\mathbb{E}[L_i] = \frac{\lambda_i}{1 - \rho_i} \left( \rho_i\mathbb{E}[R_{B_i}] + \frac{\mathbb{E}[S_i]}{\mathbb{E}[C]}\mathbb{E}[R_{S_i}] + (1 - q_{i,1})\left(\mathbb{E}[R_{\theta_{i+1,N-1}}] + \mathbb{E}[S_i]\right) \right). \quad (5)$$

Summarizing, we can say that Equation (5) has been derived by a standard application of MVA, i.e., combining the arrival relation with Little's Law. However, the unknown  $\mathbb{E}[R_{\theta_{i+1,N-1}}]$  form the stumbling block to the straightforward computation of the mean queue lengths via this equation. To obtain the unknowns  $\mathbb{E}[R_{\theta_{i+1,N-1}}]$ , we relate them to  $\mathbb{E}[L_{i,j}]$  and derive a set of equations for these quantities.

Firstly, as under the exhaustive policy no type- $i$  customers are left at the end of a visit time of queue  $i$ , the following property can be obtained. The number of type- $i$  customers present at an arbitrary moment within an  $(i+1, j)$ -period equals the number of Poisson arrivals during the age of this  $(i+1, j)$ -period. Since the age is in distribution equal to the residual time, the following equation holds

$$\sum_{n=i+1}^{i+j} \frac{q_{n,1}}{q_{i+1,j}} \mathbb{E}[L_{i,n}] = \lambda_i \mathbb{E}[R_{\theta_{i+1,j}}], \quad i = 1, 2, \dots, N, \quad j = 1, \dots, N-1. \quad (6)$$

Secondly, substitution of Identity (2) into Equation (5) yields, for  $i = 1, 2, \dots, N$ ,

$$\sum_{n=1}^N q_{n,1} \mathbb{E}[L_{i,n}] = \frac{\lambda_i}{1 - \rho_i} \left( \rho_i\mathbb{E}[R_{B_i}] + \frac{\mathbb{E}[S_i]}{\mathbb{E}[C]}\mathbb{E}[R_{S_i}] + (1 - q_{i,1})\left(\mathbb{E}[R_{\theta_{i+1,N-1}}] + \mathbb{E}[S_i]\right) \right). \quad (7)$$

It is easily seen that Equations (6) and (7) represent a set of  $N^2$  linear equations for the unknowns  $\mathbb{E}[L_{i,j}]$  and  $\mathbb{E}[R_{\theta_{i,j}}]$ . In the remainder of this subsection, we derive additional equations by expressing  $\mathbb{E}[R_{\theta_{i,j}}]$  in terms of  $\mathbb{E}[L_{i,j}]$ .

There to, we first focus on  $\mathbb{E}[R_{\theta_{i,1}}]$ . At an arbitrary moment within a visit time of queue  $i$ ,  $L_{i,i}$  type- $i$  customers are waiting, who all initiate a busy period with mean  $\mathbb{E}[B_i]/(1 - \rho_i)$ . Furthermore, with probabilities  $\rho_i\mathbb{E}[C]/\mathbb{E}[\theta_{i,1}]$  and  $\mathbb{E}[S_i]/\mathbb{E}[\theta_{i,1}]$  an additional busy period with mean  $\mathbb{E}[R_{B_i}]/(1 - \rho_i)$  and  $\mathbb{E}[R_{S_i}]/(1 - \rho_i)$  is induced, respectively. So, we have

$$\mathbb{E}[R_{\theta_{i,1}}] = \frac{1}{1 - \rho_i} \left( \mathbb{E}[L_{i,i}]\mathbb{E}[B_i] + \frac{\rho_i\mathbb{E}[C]}{\mathbb{E}[\theta_{i,1}]}\mathbb{E}[R_{B_i}] + \frac{\mathbb{E}[S_i]}{\mathbb{E}[\theta_{i,1}]}\mathbb{E}[R_{S_i}] \right), \quad i = 1, 2, \dots, N. \quad (8)$$

Next, we turn our attention to  $\mathbb{E}[R_{\theta_{i,2}}]$ . With probability  $\frac{q_{i+1,1}}{q_{i,2}}$ , the interval  $R_{\theta_{i,2}}$  is simply equal to  $R_{\theta_{i+1,1}}$ . On the other hand, with probability  $\frac{q_{i,1}}{q_{i,2}}$  this residual period equals  $R_{\theta_{i,1}} + S_{i+1}$  plus the busy periods initiated by the type- $(i+1)$  customers arriving during  $R_{\theta_{i,1}} + S_{i+1}$  and

by the type- $(i + 1)$  customers present at an arbitrary moment within a visit time of queue  $i$ . That is, we have, for  $i = 1, 2, \dots, N$ ,

$$\begin{aligned}\mathbb{E}[R_{\theta_{i,2}}] &= \frac{q_{i,1}}{q_{i,2}} \left( (\mathbb{E}[R_{\theta_{i,1}}] + \mathbb{E}[S_{i+1}])(1 + \frac{\lambda_{i+1}\mathbb{E}[B_{i+1}]}{1 - \rho_{i+1}}) + \frac{\mathbb{E}[L_{i+1,i}]\mathbb{E}[B_{i+1}]}{1 - \rho_{i+1}} \right) + (1 - \frac{q_{i,1}}{q_{i,2}})\mathbb{E}[R_{\theta_{i+1,1}}] \\ &= \frac{q_{i,1}}{q_{i,2}} \left( \frac{\mathbb{E}[R_{\theta_{i,1}}]}{1 - \rho_{i+1}} + \frac{\mathbb{E}[S_{i+1}] + \mathbb{E}[L_{i+1,i}]\mathbb{E}[B_{i+1}]}{1 - \rho_{i+1}} \right) + (1 - \frac{q_{i,1}}{q_{i,2}})\mathbb{E}[R_{\theta_{i+1,1}}].\end{aligned}$$

The derivation of  $\mathbb{E}[R_{\theta_{i,j}}]$  for general  $j$  proceeds along the same lines and is, therefore, omitted. The general expression looks as follows, for  $i = 1, 2, \dots, N$  and  $j = 2, 3, \dots, N - 1$ ,

$$\mathbb{E}[R_{\theta_{i,j}}] = \frac{q_{i,1}}{q_{i,j}} \left( \frac{\mathbb{E}[R_{\theta_{i,1}}]}{\prod_{n=1}^{j-1} (1 - \rho_{i+n})} + \sum_{n=1}^{j-1} \frac{\mathbb{E}[S_{i+n}] + \mathbb{E}[L_{i+n,i}]\mathbb{E}[B_{i+n}]}{\prod_{m=n}^{j-1} (1 - \rho_{i+m})} \right) + (1 - \frac{q_{i,1}}{q_{i,j}})\mathbb{E}[R_{\theta_{i+1,j-1}}]. \quad (9)$$

Finally, eliminating  $\mathbb{E}[R_{\theta_{i,j}}]$  from Equations (6) and (7) with the help of Equations (8) and (9) renders a set of  $N^2$  linear equations for equally many unknowns  $\mathbb{E}[L_{i,j}]$ . After solving these equations, the unconditional mean queue lengths and mean delays can be computed via Identity (2) and Little's Law (4).

It is noteworthy that the residual cycle lengths  $\mathbb{E}[R_{\theta_{i,N}}]$ ,  $i = 1, 2, \dots, N$ , which are not required for the computation of the mean delays, satisfy Equation (9) as well. An important observation is that whereas the mean cycle lengths do not depend on the queue at which the cycle starts, the mean *residual* cycle lengths generally differ. We close this subsection with an example.

**Example 3.1.** Consider a two-queue polling system with exhaustive service. Suppose that the service and setup times follow exponential distributions with means equal to 1 for both customer types. Further,  $\lambda_1 = 0.6$  and  $\lambda_2 = 0.2$  and, thus, the total occupation rate of the system equals 0.8. After some straightforward manipulations of Equations (6) and (7), we obtain the following set of 4 equations for  $\mathbb{E}[L_{i,j}]$ ,

$$\begin{aligned}165 &= 56\mathbb{E}[L_{1,1}] + 24\mathbb{E}[L_{1,2}] - 45\mathbb{E}[L_{2,2}], \\ 55 &= -35\mathbb{E}[L_{1,1}] + 56\mathbb{E}[L_{2,1}] + 24\mathbb{E}[L_{2,2}], \\ 3 &= 4\mathbb{E}[L_{1,2}] - 3\mathbb{E}[L_{2,2}], \\ 1 &= -\mathbb{E}[L_{1,1}] + 2\mathbb{E}[L_{2,1}],\end{aligned}$$

with solution,

$$\mathbb{E}[L_{1,1}] = \frac{129}{35}, \quad \mathbb{E}[L_{1,2}] = \frac{12}{5}, \quad \mathbb{E}[L_{2,1}] = \frac{82}{35}, \quad \mathbb{E}[L_{2,2}] = \frac{11}{5}.$$

By using Identity (2),

$$\mathbb{E}[L_i] = \frac{7}{10}\mathbb{E}[L_{i,1}] + \frac{3}{10}\mathbb{E}[L_{i,2}], \quad i = 1, 2,$$

in conjunction with Little's Law (4), the mean delays for both customer types readily follow, i.e.,

$$\mathbb{E}[W_1] = 5\frac{1}{2}, \quad \mathbb{E}[W_2] = 11\frac{1}{2}.$$

For the mean residual cycle lengths  $\mathbb{E}[R_{C_i}] = \mathbb{E}[R_{\theta_{i,N}}]$ , where a cycle obviously starts at a departure epoch of the server from queue  $i$ , we have

$$\mathbb{E}[R_{C_1}] = 13\frac{3}{4}, \quad \mathbb{E}[R_{C_2}] = 14\frac{3}{8}.$$

From these number, the following well-known linear relation between the mean delays and the mean residual cycle lengths in case of exhaustive service clearly emerges (see, e.g., [2])

$$\mathbb{E}[W_i] = (1 - \rho_i)\mathbb{E}[R_{C_i}], \quad i = 1, 2. \quad (10)$$

In the past, highly accurate approximations for the exhaustive service policy have developed based on the following idea [2]. Firstly, assume that the mean residual cycle lengths are identical for all queues, i.e.,  $\mathbb{E}[R_{C_i}] = \mathbb{E}[R_C]$ . Secondly, substitute the linear relations between  $\mathbb{E}[W_i]$  and  $\mathbb{E}[R_C]$  resulting from Equation (10) and this assumption in the pseudoconservation law (3) and solve for the single unknown  $\mathbb{E}[R_C]$ .

For the present example, this results into the following approximate figures,

$$\mathbb{E}[R_C] \approx 14, \quad \mathbb{E}[W_1] \approx 5\frac{3}{5}, \quad \mathbb{E}[W_2] \approx 11\frac{1}{5},$$

which are approximately equal to the exact numbers obtained by MVA. This observation clearly supports the developed approximation and, in particular, the assumption of equality of the mean residual cycle lengths.  $\square$

### 3.2 Gated service

In case of gated service, all customers waiting in queue at the start of a visit time of this queue are placed behind a gate meaning that they are served in the current cycle. However, customers arriving during a visit time of their queue are placed before this gate and are, thus, only served in the next cycle. With this difference understood, it is clear that, in case  $i = j$ ,  $L_{i,j}$  is the sum of two auxiliary variables, i.e.,

$$L_{i,i} = \bar{L}_{i,i} + \tilde{L}_{i,i}, \quad i = 1, 2, \dots, N,$$

where  $\bar{L}_{i,i}$  and  $\tilde{L}_{i,i}$  represent the queue length behind and before the gate, respectively. Recall that the customer in service is excluded. In case  $i \neq j$ , all customers in queue  $i$  are obviously located before the gate, i.e.,

$$L_{i,j} = \tilde{L}_{i,j}, \quad i \neq j = 1, 2, \dots, N.$$

The corresponding unconditional queue length  $L_i$  has mean

$$\mathbb{E}[L_i] = \mathbb{E}[\tilde{L}_i] + q_{i,1}\mathbb{E}[\bar{L}_{i,i}] = \sum_{n=1}^N q_{n,1}\mathbb{E}[\tilde{L}_{i,n}] + q_{i,1}\mathbb{E}[\bar{L}_{i,i}], \quad i = 1, 2, \dots, N. \quad (11)$$

Again, our analysis makes extensively use of Little's Law and the PASTA property. That is, we tag a customer at its arrival to queue  $i$ ,  $i = 1, 2, \dots, N$ . By the PASTA property, we know that this customer sees the system in equilibrium. So, the tagged customer has to wait for the service of all customers  $\tilde{L}_i$ , who were already waiting before the gate on his

arrival. Furthermore, he has to wait until the first polling instant of queue  $i$  equalling a residual  $(i, N)$ -period, i.e., a residual cycle. By definition of the gated policy, this extra delay is incurred even in case the tagged type- $i$  customer arrives in a visit time of queue  $i$ .

Consequently, the mean delay  $\mathbb{E}[W_i]$  of a type- $i$  customer is given by

$$\mathbb{E}[W_i] = \mathbb{E}[\tilde{L}_i]\mathbb{E}[B_i] + \mathbb{E}[R_{\theta_{i,N}}], \quad i = 1, 2, \dots, N,$$

which, in combination with Little's Law (4), gives us the following relation

$$\mathbb{E}[L_i] = \rho_i \mathbb{E}[\tilde{L}_i] + \lambda_i \mathbb{E}[R_{\theta_{i,N}}], \quad i = 1, 2, \dots, N. \quad (12)$$

Once more, the mean residual periods have to be obtained, where we choose the same solution approach as in the exhaustive case.

The gated policy, together with the definition of a visit time, clearly implies that the number of type- $i$  customers before the gate at an arbitrary moment within an  $(i, j)$ -period is equal to the number of Poisson arrivals during the age of an  $(i, j)$ -period, which is in distribution again equal to a residual  $(i, j)$ -period. That is,

$$\sum_{n=i}^{i+j-1} \frac{q_{n,1}}{q_{i,j}} \mathbb{E}[\tilde{L}_{i,n}] = \lambda_i \mathbb{E}[R_{\theta_{i,j}}], \quad i = 1, 2, \dots, N, \quad j = 1, 2, \dots, N. \quad (13)$$

Secondly, if we substitute Identity (11) into Equation (12), we get

$$(1 - \rho_i) \sum_{n=1}^N q_{n,1} \mathbb{E}[\tilde{L}_{i,n}] + q_{i,1} \mathbb{E}[\bar{L}_{i,i}] = \lambda_i \mathbb{E}[R_{\theta_{i,N}}], \quad i = 1, 2, \dots, N. \quad (14)$$

Now, we see that Equations (13) and (14) comprise a set of  $N(N+1)$  linear equations for  $\mathbb{E}[\bar{L}_{i,i}]$ ,  $\mathbb{E}[\tilde{L}_{i,j}]$  and  $\mathbb{E}[R_{\theta_{i,j}}]$ . To eliminate the unknown mean residual  $(i, j)$ -periods from this set, these quantities are rewritten in terms of  $\mathbb{E}[\bar{L}_{i,i}]$  and  $\mathbb{E}[\tilde{L}_{i,j}]$ .

Starting with  $\mathbb{E}[R_{\theta_{i,1}}]$ , we recognize that this period lasts at least the sum of the service times of the customers behind the gate. With probability  $\rho_i \mathbb{E}[C]/\mathbb{E}[\theta_{i,1}]$  a residual service time and a setup time for queue  $i+1$  is induced, while with probability  $\mathbb{E}[S_{i+1}]/\mathbb{E}[\theta_{i,1}]$  only a residual setup time for queue  $i+1$  is generated. Consequently, we have

$$\mathbb{E}[R_{\theta_{i,1}}] = \mathbb{E}[\bar{L}_{i,i}]\mathbb{E}[B_i] + \frac{\mathbb{E}[S_{i+1}]}{\mathbb{E}[\theta_{i,1}]} \mathbb{E}[R_{S_{i+1}}] + \frac{\rho_i \mathbb{E}[C]}{\mathbb{E}[\theta_{i,1}]} (\mathbb{E}[R_{B_i}] + \mathbb{E}[S_{i+1}]), \quad i = 1, 2, \dots, N. \quad (15)$$

In case of an  $(i, 2)$ -period,  $R_{\theta_{i,2}}$  equals  $R_{\theta_{i+1,1}}$  with probability  $\frac{q_{i+1,1}}{q_{i,2}}$ . With probability  $\frac{q_{i,1}}{q_{i,2}}$ , however, this residual period equals  $R_{\theta_{i,1}} + S_{i+2}$  plus the service times of the type- $(i+1)$  customers present at an arbitrary moment within a visit time of queue  $i$  and of the type- $(i+1)$  customers arriving during  $R_{\theta_{i,1}}$ . This yields, for  $i = 1, 2, \dots, N$ ,

$$\begin{aligned} \mathbb{E}[R_{\theta_{i,2}}] &= \frac{q_{i,1}}{q_{i,2}} \left( \mathbb{E}[R_{\theta_{i,1}}] + \mathbb{E}[S_{i+2}] + (\lambda_{i+1} \mathbb{E}[R_{\theta_{i,1}}] + \mathbb{E}[\tilde{L}_{i+1,i}]) \mathbb{E}[B_{i+1}] \right) + \left( 1 - \frac{q_{i,1}}{q_{i,2}} \right) \mathbb{E}[R_{\theta_{i+1,1}}] \\ &= \frac{q_{i,1}}{q_{i,2}} \left( \mathbb{E}[R_{\theta_{i,1}}] (1 + \rho_{i+1}) + \mathbb{E}[S_{i+2}] + \mathbb{E}[\tilde{L}_{i+1,i}] \mathbb{E}[B_{i+1}] \right) + \left( 1 - \frac{q_{i,1}}{q_{i,2}} \right) \mathbb{E}[R_{\theta_{i+1,1}}]. \end{aligned}$$

The derivation of  $\mathbb{E}[R_{\theta_{i,j}}]$  for general  $j$  is similar. After some straightforward calculations, the following expression can be derived for  $i = 1, 2, \dots, N$  and  $j = 2, 3, \dots, N$ ,

$$\begin{aligned} \mathbb{E}[R_{\theta_{i,j}}] &= \frac{q_{i,1}}{q_{i,j}} \left( \mathbb{E}[R_{\theta_{i,1}}] \prod_{n=1}^{j-1} (1 + \rho_{i+n}) + \sum_{n=1}^{j-1} (\mathbb{E}[S_{i+n+1}] + \mathbb{E}[\tilde{L}_{i+n,i}] \mathbb{E}[B_{i+n}]) \prod_{m=n+1}^{j-1} (1 + \rho_{i+m}) \right) \\ &\quad + (1 - \frac{q_{i,1}}{q_{i,j}}) \mathbb{E}[R_{\theta_{i+1,j-1}}]. \end{aligned} \quad (16)$$

To conclude, elimination of  $\mathbb{E}[R_{\theta_{i,j}}]$  from Equations (13) and (14) with the help of Equations (15) and (16) yields a set of  $N(N+1)$  linear equations for equally many unknowns  $\mathbb{E}[\tilde{L}_{i,i}]$  and  $\mathbb{E}[\tilde{L}_{i,j}]$ . Together with Identity (11) and Little's Law (4), the solution to these equations yields the unconditional mean queue lengths and mean delays. This subsection is closed with an example.

**Example 3.2.** All input parameters are taken the same as in Example 3.1, but the customers are now served according to the gated discipline. The following set of 6 equations can be obtained for this case,

$$\begin{aligned} 39 &= 35\mathbb{E}[\tilde{L}_{1,1}] - 21\mathbb{E}[\bar{L}_{1,1}], \\ 5 &= 15\mathbb{E}[\tilde{L}_{2,2}] - 3\mathbb{E}[\bar{L}_{2,2}], \\ 414 &= 70\mathbb{E}[\tilde{L}_{1,1}] + 30\mathbb{E}[\tilde{L}_{1,2}] + 49\mathbb{E}[\bar{L}_{1,1}] - 105\mathbb{E}[\tilde{L}_{2,1}] - 45\mathbb{E}[\bar{L}_{2,2}], \\ 120 &= -15\mathbb{E}[\tilde{L}_{1,2}] - 35\mathbb{E}[\bar{L}_{1,1}] + 140\mathbb{E}[\tilde{L}_{2,1}] + 60\mathbb{E}[\tilde{L}_{2,2}] + 51\mathbb{E}[\bar{L}_{2,2}], \\ 414 &= 175\mathbb{E}[\tilde{L}_{1,1}] + 75\mathbb{E}[\tilde{L}_{1,2}] - 126\mathbb{E}[\bar{L}_{1,1}] - 105\mathbb{E}[\tilde{L}_{2,1}] - 45\mathbb{E}[\bar{L}_{2,2}], \\ 120 &= -15\mathbb{E}[\tilde{L}_{1,2}] - 35\mathbb{E}[\bar{L}_{1,1}] + 175\mathbb{E}[\tilde{L}_{2,1}] + 75\mathbb{E}[\tilde{L}_{2,2}] - 24\mathbb{E}[\bar{L}_{2,2}]. \end{aligned}$$

The solution of this set reads

$$\begin{aligned} \mathbb{E}[\tilde{L}_{1,1}] &= \frac{23412}{6545}, & \mathbb{E}[\tilde{L}_{1,2}] &= \frac{7121}{935}, & \mathbb{E}[\bar{L}_{1,1}] &= \frac{5373}{1309}, \\ \mathbb{E}[\tilde{L}_{2,1}] &= \frac{13561}{6545}, & \mathbb{E}[\tilde{L}_{2,2}] &= \frac{513}{935}, & \mathbb{E}[\bar{L}_{2,2}] &= \frac{604}{561}. \end{aligned}$$

Applying Identity (11),

$$\begin{aligned} \mathbb{E}[L_1] &= \frac{7}{10}(\mathbb{E}[\tilde{L}_{1,1}] + \mathbb{E}[\bar{L}_{1,1}]) + \frac{3}{10}\mathbb{E}[\tilde{L}_{1,2}], \\ \mathbb{E}[L_2] &= \frac{7}{10}\mathbb{E}[\tilde{L}_{2,1}] + \frac{3}{10}(\mathbb{E}[\tilde{L}_{2,2}] + \mathbb{E}[\bar{L}_{2,2}]), \end{aligned}$$

together with Little's Law (4) give us the mean delays, i.e.,

$$\mathbb{E}[W_1] = 12\frac{144}{187}, \quad \mathbb{E}[W_2] = 9\frac{129}{187}.$$

When comparing these delays to the values in Example 3.1, we can make two observations. Firstly, the delay in queue 1 has increased, whereas the delay in the second queue has become smaller. Secondly, a well-known qualitative property of polling systems comes to light, i.e., in exhaustive systems heavily loaded queues experience lower delays than lightly loaded queues,

whereas in gated systems the opposite is true.

Again, the mean residual cycle lengths  $\mathbb{E}[R_{C_i}] = \mathbb{E}[R_{\theta_{i,N}}]$  can be easily computed as well,

$$\mathbb{E}[R_{C_1}] = 7\frac{367}{374}, \quad \mathbb{E}[R_{C_2}] = 8\frac{14}{187},$$

where a cycle for queue  $i$  starts with the arrival of the server at queue  $i$ . These values satisfy a well-known linear relation between the mean delays and the mean residual cycle lengths for gated service (see, e.g., [2]), i.e.,

$$\mathbb{E}[W_i] = (1 + \rho_i)\mathbb{E}[R_{C_i}], \quad i = 1, 2.$$

As in the exhaustive case, gated polling systems can also be approximated by assuming equal mean residual cycle lengths and using the pseudoconservation law (3). In the present example, we obtain the following approximate values,

$$\mathbb{E}[R_C] \approx 8, \quad \mathbb{E}[W_1] \approx 12\frac{4}{5}, \quad \mathbb{E}[W_2] \approx 9\frac{3}{5},$$

which closely resemble the exact values computed by MVA. □

### 3.3 Model variations

The present subsection shows that MVA can be generalized to two model variations of the basic system introduced in Section 2, i.e., systems with mixed service and systems with zero setup times.

#### 3.3.1 Systems with mixed service

To treat the case of mixed service polling systems, we first have to realize that the two different definitions of the visit times conflict. It is, therefore, necessary to do the conditioning of the queue lengths on the system state in a more detailed manner. That is, we introduce  $L'_{i,j}$  and  $L_{i,j}$  as the queue lengths at queue  $i$  at an arbitrary epoch within a setup time and a service period of queue  $j$ ,  $i, j = 1, 2, \dots, N$ , respectively. Recall that in the pure exhaustive and gated systems, we could aggregate a setup time and a service period in one single random variable, the visit period. It goes without saying that for the gated queues we again have to distinguish between the queue length behind and before the gate indicated by an additional bar and tilde on the corresponding random variables, respectively.

The total number of variables, and thus the total number of linear equations, is now equal to  $2N^2 + K$  with  $K \leq N$  the number of queues deploying the gated discipline. The derivation of these equations is comparable to the analyses of the preceding subsections and provides little additional insight. Instead, the application of MVA to mixed service polling systems can be exhibited in a more satisfactory and illuminating way by an example, which is done in the remainder of the present subsection.

**Example 3.3.** Consider again the input parameters of Example 3.1, but assume now that queue 1 is served exhaustively, whereas queue 2 is served according to the gated policy. The

following set of 9 equations for equally many unknowns holds

$$\begin{aligned}
198 &= 10\mathbb{E}[L'_{1,1}] + 60\mathbb{E}[L_{1,1}] + 10\mathbb{E}[L'_{1,2}] + 20\mathbb{E}[L_{1,2}] - 15\mathbb{E}[\tilde{L}'_{2,2}] - 30\mathbb{E}[\tilde{L}_{2,2}], \\
75 &= -5\mathbb{E}[L'_{1,1}] - 30\mathbb{E}[L_{1,1}] - 10\mathbb{E}[L_{1,2}] + 8\mathbb{E}[\tilde{L}'_{2,1}] + 48\mathbb{E}[\tilde{L}_{2,1}] + 8\mathbb{E}[\tilde{L}'_{2,2}] + 16\mathbb{E}[\tilde{L}_{2,2}] + \mathbb{E}[\bar{L}_{2,2}], \\
3 &= 5\mathbb{E}[L'_{1,2}], \\
2 &= \mathbb{E}[\tilde{L}'_{2,2}], \\
48 &= 25\mathbb{E}[L'_{1,2}] + 50\mathbb{E}[L_{1,2}] - 15\mathbb{E}[\tilde{L}'_{2,2}] - 30\mathbb{E}[\tilde{L}_{2,2}], \\
108 &= 25\mathbb{E}[L'_{1,1}] + 25\mathbb{E}[L'_{1,2}] + 50\mathbb{E}[L_{1,2}] - 15\mathbb{E}[\tilde{L}'_{2,2}] - 30\mathbb{E}[\tilde{L}_{2,2}], \\
1 &= 5\mathbb{E}[\tilde{L}_{2,2}] - \mathbb{E}[\bar{L}_{2,2}], \\
5 &= 5\mathbb{E}[\tilde{L}'_{2,1}] + 10\mathbb{E}[\tilde{L}_{2,2}] - 2\mathbb{E}[\bar{L}_{2,2}], \\
11 &= -\mathbb{E}[L'_{1,1}] - 6\mathbb{E}[L_{1,1}] - \mathbb{E}[L_{1,2}] + 2\mathbb{E}[\tilde{L}'_{2,1}] + 12\mathbb{E}[\tilde{L}_{2,1}] + 4\mathbb{E}[\tilde{L}_{2,2}] - \mathbb{E}[\bar{L}_{2,2}],
\end{aligned}$$

whence

$$\begin{aligned}
\mathbb{E}[L'_{1,1}] &= \frac{12}{5}, & \mathbb{E}[L_{1,1}] &= \frac{453}{125}, & \mathbb{E}[L'_{1,2}] &= \frac{3}{5}, & \mathbb{E}[L_{1,2}] &= \frac{687}{250}, \\
\mathbb{E}[\tilde{L}'_{2,1}] &= \frac{3}{5}, & \mathbb{E}[\tilde{L}_{2,1}] &= \frac{867}{250}, & \mathbb{E}[\tilde{L}'_{2,2}] &= 2, & \mathbb{E}[\tilde{L}_{2,2}] &= \frac{87}{125}, & \mathbb{E}[\bar{L}_{2,2}] &= \frac{62}{25}.
\end{aligned}$$

Introducing the following obvious relations, commensurable to the identities used in the pure exhaustive and gated analyses,

$$\begin{aligned}
\mathbb{E}[L_1] &= \frac{1}{10}\mathbb{E}[L'_{1,1}] + \frac{3}{5}\mathbb{E}[L_{1,1}] + \frac{1}{10}\mathbb{E}[L'_{1,2}] + \frac{1}{5}\mathbb{E}[L_{1,2}], \\
\mathbb{E}[L_2] &= \frac{1}{10}\mathbb{E}[\tilde{L}'_{2,1}] + \frac{3}{5}\mathbb{E}[\tilde{L}_{2,1}] + \frac{1}{10}\mathbb{E}[\tilde{L}'_{2,2}] + \frac{1}{5}(\mathbb{E}[\tilde{L}_{2,2}] + \mathbb{E}[\bar{L}_{2,2}]),
\end{aligned}$$

together with Little's Law (4) yields the mean delays,

$$\mathbb{E}[W_1] = 5\frac{1}{25}, \quad \mathbb{E}[W_2] = 14\frac{22}{25}.$$

When comparing these values with the ones in the pure exhaustive and gated systems, it can be seen that the mean delay for the customers in queue 1 is decreased at the expense of a lower quality of service level in queue 2.

As a spin-off of our analysis, we obtain the mean residual cycle lengths given by

$$\mathbb{E}[R_{C_1}] = 12\frac{3}{5}, \quad \mathbb{E}[R_{C_2}] = 12\frac{2}{5},$$

where the cycle for queue 1 starts when the server leaves this queue. However, for queue 2 the cycle commences at the arrival epoch of the server at queue 2. As in the pure exhaustive and gated cases, again relations between the mean delay and the mean residual cycles exist, i.e.,

$$\mathbb{E}[W_1] = (1 - \rho_1)\mathbb{E}[R_{C_1}], \quad \mathbb{E}[W_2] = (1 + \rho_2)\mathbb{E}[R_{C_2}].$$

By ignoring the slight difference in definition of the starting point of cycles, the approximation technique proposed in [2] can also be applied in a polling with mixed service resulting in the following approximations for the present example,

$$\mathbb{E}[R_C] \approx 12\frac{1}{2}, \quad \mathbb{E}[W_1] \approx 5, \quad \mathbb{E}[W_2] \approx 15,$$

which almost coincide with the exact values obtained by MVA.  $\square$

A minor but interesting variant of the gated discipline is the so-called fully gated strategy [1], also called the reserved gated strategy [2]. Under this fully gated policy, all, and only, customers found by the server at the start of the setup time are served. For this fully gated strategy, the definition of the visit time can be identically chosen to the definition for exhaustive service, which simplifies the analysis of a mixed exhaustive/fully gated system considerably (the conditioning of the queue lengths on the system state can obviously be done in the standard manner of Section 2).

### 3.3.2 Systems with zero setup times

In Section 2, we touched upon the difference between systems with *nonzero* and *zero* setup times. In the latter system, the mean cycle length and mean visit times both tend to zero, which causes problems in the definition of the probabilities  $q_{i,j}$ . To circumvent these difficulties, we should modify the definition for a mean  $(i, j)$ -period as follows

$$\mathbb{E}[\theta_{i,j}] = \sum_{n=i}^{i+j-1} \rho_n \mathbb{E}[C], \quad i = 1, 2, \dots, N, \quad j = 1, 2, \dots, N, \quad (17)$$

where we can leave the value of  $\mathbb{E}[C]$  unspecified, since it appears that, in case of zero setup times, this quantity cancels out in all steps of the analysis. Then, the probabilities  $q_{i,j}$  are again well defined and change accordingly to

$$q_{i,j} = \sum_{n=i}^{i+j-1} \rho_n, \quad i = 1, 2, \dots, N, \quad j = 1, 2, \dots, N.$$

By replacing all variables related to setup times by zeros in the analysis of Section 3 and applying Definition (17), the mean delay in polling systems with zero setup times can be computed as well.

## 4 Conclusion and possible extensions

In the present paper, we have studied the so-called *mean value analysis* (MVA) for the computation of the mean delays in exhaustive-type or gated-type polling systems. The most significant benefits of MVA are in its intrinsic simplicity and intuitively appealing derivation. Hence, MVA can be, without seriously complicating the analysis, carried over to variants of the considered polling systems: (i) systems with Poisson batch arrivals, (ii) systems with fixed polling tables and (iii) discrete-time polling systems. Extensions to other polling systems, either in an exact or approximate way, is not inconceivable as well.

In particular, [8] and [17] have independently illuminated a striking dichotomy in complexity between different polling systems. If the service discipline satisfies a certain branching property, the polling system allows for an exact analysis by rather standard methods. Possibly, MVA may be applied for the exact evaluation of these systems. If this branching property is, however, violated, the corresponding polling systems can not be analyzed exactly in the general setting, but MVA may in these cases be a useful tool for the development of approximations.

## Acknowledgment

The authors would like to thank Onno Boxma for valuable discussions related to this paper.

## References

- [1] Bertsekas, D., Gallager, R., (1987). *Data Networks* (Prentice-Hall, New Jersey).
- [2] Boxma, O.J., (1989). *Workloads and waiting times in single-server systems with multiple customer classes* (Queueing Systems, vol. 5, pp. 185-214).
- [3] Cooper, R.B., Murray, G., (1969). *Queues served in cyclic order* (The Bell System Technical Journal, vol. 48, pp. 675-689).
- [4] Cooper, R.B., (1970). *Queues served in cyclic order: waiting times* (The Bell System Technical Journal, vol. 49, 399-413).
- [5] Eisenberg, M., (1972). *Queues with periodic service and changeover time* (Operations Research, vol. 20, no. 2, pp. 440-451).
- [6] Ferguson, M.J., Aminetzah, Y., (1985). *Exact results for nonsymmetric token ring systems* (IEEE Transactions on Communications, vol. COM-33, pp. 223-231).
- [7] Franken, P., Koenig, D., Arndt, W., Schmidt, F., (1982). *Queues and Point Processes* (John Wiley, New York).
- [8] Fuhrmann, S.W., (1981). *Performance analysis of a class of cyclic schedules* (Bell Laboratories Technical Memorandum 81-59531-1).
- [9] Hirayama, T., Hong, S.J., Krunk, M., (2004). *A new approach to analysis of polling systems* (Queueing Systems, vol. 48, no. 1-2, pp. 135-158).
- [10] Konheim, A.G., Meister, B., (1974). *Waiting lines and times in a system with polling* (Journal of the Association for Computing Machinery, vol. 21, no. 3, pp. 470-490).
- [11] Konheim, A.G., Levy, H., Srinivasan, M.M., (1994). *Descendant set: an efficient approach for the analysis of polling systems* (IEEE Transactions on Communications, vol. 42, no. 2/3/4, pp. 1245-1253).
- [12] Levy, H., (1989). *Delay computation and dynamic behavior of non-symmetric polling systems* (Performance Evaluation, vol. 10, no. 1, pp. 35-51).
- [13] Levy, H., Sidi, M., (1990). *Polling systems: applications, modeling and optimization* (IEEE Transactions on Communications, vol. COM-38, no. 10, pp. 1750-1760).
- [14] Little, J.D.C., (1961). *A proof of the queueing formula  $L = \lambda W$*  (Operations Research, vol. 9, pp. 383-387).
- [15] Mack, C., Murphy, T., Webb, N.L., (1957). *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, vol. 19, no. 1, pp. 166-172).
- [16] Mack, C., (1957). *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, vol. 19, no. 1, pp. 173-178).
- [17] Resing, J.A.C., (1993). *Polling systems and multitype branching processes* (Queueing Systems, vol. 13, pp. 409-426).
- [18] Rubin, I., De Moraes, L.F.M., (1983). *Message delay analysis for polling and token multiple-access schemes for local communication networks* (IEEE Journal on Selected Areas in Communications, vol. SAC-1, no. 5, pp. 935-947).

- [19] Sarkar, D., Zangwill, W.I., (1989). *Expected waiting time for nonsymmetric cyclic queueing systems - exact results and applications* (Management Science, vol. 35, pp. 1463-1474).
- [20] Srinivasan, M.M., Levy, H., Konheim, A.G., (1996). *The individual station technique for the analysis of polling systems* (Naval Research Logistics, vol. 43, no. 1, pp. 79-101).
- [21] Swartz, G.B., (1980). *Polling in a loop system* (Journal of the Association for Computing Machinery, vol. 27, no. 1, pp. 42-59).
- [22] Takagi, H., (1990). *Queueing analysis of polling models: an update* (In Stochastic Analysis of Computer and Communication Systems, H. Takagi (ed.), North-Holland, Amsterdam, pp. 267-318).
- [23] Takagi, H., (1997). *Queueing analysis of polling models: progress in 1990-1994* (In Frontiers in Queueing: Models, Methods and Problems, J.H. Dshalalow (ed.), CRC Press, Boca Raton, pp. 119-146).
- [24] Takagi, H., (2000). *Analysis and application of polling models* (In Performance Evaluation: Origins and Directions, G. Haring, C. Lindemann and M. Reiser (eds.), Lecture Notes in Computer Science, vol. 1769, Springer, Berlin, pp. 423-442).
- [25] Wolff, R.W., (1982). *Poisson arrivals see time averages* (Operations Research, vol. 30, no. 2, pp. 223-231).