Queueing Syst
DOI 10.1007/s11134-016-9513-y

The analysis of batch sojourn-times in polling systems

**Jelmer P. van der Gaast[1] · Ivo J. B. F. Adan[2] ·
René B. M. de Koster[1]**

Received: 11 November 2015 / Revised: 4 August 2016

Abstract We consider a cyclic polling system with general service times, general switch-over times, and simultaneous batch arrivals. This means that at an arrival epoch, a batch of customers may arrive simultaneously at the different queues of the system. For the exhaustive service discipline, we study the batch sojourn-time, which is defined as the time from an arrival epoch until service completion of the last customer in the batch. We obtain exact expressions for the Laplace–Stieltjes transform of the steady-state batch sojourn-time distribution, which can be used to determine the moments of the batch sojourn-time and, in particular, its mean. However, we also provide an alternative, more efficient way to determine the mean batch sojourn-time, using mean value analysis. We briefly show how our framework can be applied to other service disciplines: locally gated and globally gated. Finally, we compare the batch sojourn-times for different service disciplines in several numerical examples. Our results show that the best performing service discipline, in terms of minimizing the batch sojourn-time, depends on system characteristics.

Keywords Queueing models · Polling models · Mean value analysis

Electronic supplementary material The online version of this article (doi:10.1007/s11134-016-9513-y) contains supplementary material, which is available to authorized users.

✉ Jelmer P. van der Gaast
jgaast@rsm.nl

Ivo J. B. Adan
iadan@tue.nl

René B. M. de Koster
rkoster@rsm.nl

[1] Erasmus University Rotterdam, Rotterdam, The Netherlands

[2] Technical University Eindhoven, Eindhoven, The Netherlands

Published online: 12 January 2017

Mathematics Subject Classification 60K25 · 68M20

1 Introduction

Polling models are multi-queue systems in which a single server cyclically visits queues in order to serve waiting customers, typically incurring a switch-over time when moving to the next queue. Polling systems have been extensively used for decades to model a wide variety of applications in areas such as computer and communication systems, production systems, and traffic and transportation systems [1,19]. In the majority of the literature on polling systems, it is assumed that in each queue, new customers arrive via independent Poisson processes. However, in many applications, these arrival processes are not necessarily independent; customers arrive in batches, and batches of customers may arrive at different queues simultaneously [21]. It is important to consider the correlation structure in the arrival processes for these applications, because neglecting it may lead to strongly erroneous performance predictions and, consequently, to improper decisions about system performance. In this paper, we study the *batch sojourn-time* in polling systems with simultaneous arrivals, that is, the time until all the customers in a single batch are served after an arrival epoch.

Batch sojourn-times are of great interest in many applications of polling systems with simultaneous arrivals. Below we describe two examples in manufacturing and communication. The first example is the *stochastic economic lot scheduling problem*, which is used to study the production of multiple products on a single machine with limited capacity, under uncertain demands, production times, and setup times [9,24]. In the case of a cyclic policy, there is a fixed production sequence such that the order in which products are manufactured is always known to the manufacturer. Whenever a customer has placed an order for one or multiple products, the machine starts production. After the requested number of products has been produced, including possible demand for the same product from orders that just came in, the machine starts to process the next product in the sequence. In this way, the machine polls the buffers of the different product categories to check whether production is required. In this example, the server represents the machine, a customer represents a unit of demand for a given product, and a batch arrival corresponds to the order itself. The batch sojourn-time is defined as the total time required for manufacturing an entire order.

The second example from the area of computer communication systems is an *I/O subsystem* of a web server. Web servers are required to perform millions of transaction requests per day at an acceptable quality of service (QoS) level in terms of client response time and server throughput [22]. When a request for a web page from the server is made, several file-retrieval requests are made simultaneously (for example, text, images, and multimedia). In many implementations, these incoming file-retrieval requests are placed in separate I/O buffers. The I/O controller continuously polls, using a scheduling mechanism, the different buffers to check for pending file-retrieval requests to be executed. The web page will be fully loaded when all its file-retrieval requests are executed. In this application, the server represents the I/O controller, a customer represents an individual file-retrieval request, a batch of customers who arrive simultaneously corresponds to each web page request, and the batch sojourn-time is the time required to fully load a web page.

The objective of this paper is to analyze the batch sojourn-time in a cyclic polling system with simultaneous batch arrivals. The contribution of this paper is that we obtain exact expressions for the Laplace–Stieltjes transform of the steady-state batch sojourn-time distribution for exhaustive service, which can be used to determine the moments of the batch sojourn-time and, in particular, its mean. However, we provide an alternative, more efficient way to determine the mean batch sojourn-time by extending the mean value analysis (MVA) approach of Winands et al. [23]. We briefly show how our framework can be applied to other service disciplines that satisfy the branching property [16], i.e., locally gated and globally gated. We compare the batch sojourn-times for the different service disciplines in several numerical examples and show that the best performing service discipline, minimizing the batch sojourn-time, depends on system characteristics. From the results, we conclude that there is no unique best service discipline that minimizes the expected batch sojourn-time. As such, our results provide a starting point for a framework to minimize batch sojourn-times for a given polling system.

The organization of this paper is as follows. In Sect. 2, the literature review is given. In Sect. 3, a detailed description of the model and the corresponding notation used in this paper is given. Section 4 analyzes the batch sojourn-time for exhaustive service, the analysis for locally gated service and globally gated service is shown in the appendix. We extensively analyze the results of our model in Sect. 5 via computational experiments for a range of parameters. Finally, in Sect. 6, we conclude and suggest some further research topics.

2 Literature review

In the literature, polling systems with simultaneous arrivals have not been studied intensively. Shiozawa et al. [17] studies a two-queue polling system where customers arrive at each station according to an independent Poisson process and, in addition, customers can arrive in pairs at the system and each join a different queue. The authors derive the Laplace–Stieltjes transform of the waiting time distribution of an individual customer and the response time distribution of a pair of customers who arrive simultaneously. Levy and Sidi [14] studies polling models with simultaneous batch arrivals. For models with gated or exhaustive service, they derive a set of linear equations for the expected waiting time at each of the queues. They also provide a pseudo-conservation law for the system, i.e., an exact expression for a specific weighted sum of the expected waiting times at the different queues. Chiarawongse and Srinivasan [5] also derives pseudo-conservation laws, but in their model all customers in a batch join the same queue. Finally, Van der Mei [20] considers an asymmetric cyclic polling model with mixtures of gated and exhaustive service and general service time and switch-over time distributions and studies the heavy traffic behavior. The results were further generalized in [21].

3 Model description

Consider a polling system consisting of $N \geq 2$ infinite buffer queues $Q_1, \ldots, Q_N$ served by a single server that visits the queues in a fixed cyclic order. For ease of presentation, all references to queue indices greater than N or less than 1 are implicitly assumed to be modulo N, for example, Q_{N+1} is understood as Q_1. Assume that a new batch of customers arrives according to a Poisson process with rate λ. Each batch of customers is of size $\boldsymbol{K} = (K_1, \dots, K_N)$, where K_i represents the number of customers entering the system at Q_i, $i = 1, \dots, N$. The random vector $\boldsymbol{K}$ is assumed to be independent of past and future arrival epochs and at least one element of vector $\boldsymbol{K}$ is larger than 0 and the other elements are larger than or equal to 0, i.e., each batch contains at least one customer. The set of all possible realizations of $\boldsymbol{K}$ is denoted by $\mathcal{K}$, and let $\boldsymbol{k} = (k_1, \dots, k_N)$ be a realization of $\boldsymbol{K}$. The joint probability distribution of $\boldsymbol{K}, \pi(\boldsymbol{k}) = \mathbb{P}(K_1 = k_1, \dots, K_N = k_N)$ is arbitrary, and its corresponding probability generating function (PGF) is given by $\widetilde{K}(z) = E\left(z_1^{K_1} z_2^{K_2} \dots z_N^{K_N}\right)$. The PGF of the marginal batch size distribution at Q_i is denoted by $\widetilde{K}_i(z) = \widetilde{K}(1, \dots, 1, z, 1, \dots, 1)$, $|z| \leq 1$, where the z occurs at the ith entry. The arrival rate of customers to Q_i is $\lambda_i = \lambda E(K_i)$, and let $E(K_{ij}) = E(K_i K_j)$ for $i \neq j$ and $E(K_{ii}) = E(K_i^2) - E(K_i)$. The total arrival rate of customers arriving in the system is given by $\Lambda = \sum_{i=1}^N \lambda_i$.

The service time of a customer in Q_i is a generally distributed random variable B_i with Laplace–Stieltjes transform (LST) $\widetilde{B}_i(.)$, and with first and second moment $E(B_i)$ and $E(B_i^2)$, respectively. The workload at queue Q_i, $i = 1, \dots, N$, is defined by $\rho_i = \lambda_i E(B_i)$; the overall system load by $\rho = \sum_{i=1}^N \rho_i$. In order for the system to be stable, a necessary and sufficient condition is that $\rho < 1$ [18]. In the remainder of this paper, it is assumed that the condition for stability holds. When the server switches from Q_i to Q_{i+1}, it incurs a generally distributed switch-over time S_i with LST $\tilde{S}_i(.)$, and first and second moment $E(S_i)$ and $E(S_i^2)$. Let $E(S) = \sum_{i=1}^N E(S_i)$ be the mean total switch-over time in a cycle and $E(S^2) = \sum_{i=1}^N E(S_i^2) + \sum_{i \neq j} E(S_i) E(S_j)$ its second moment.

The cycle time C_i of Q_i is defined as the time between two successive visits of the server at this queue. A cycle consists of N visit periods each followed by a switch-over time; $V_i, S_i, V_{i+1}, \dots, V_{i+N-1}, S_{i+N-1}$ (see Fig. 1). A visit period, V_i, starts whenever there are customers waiting at Q_i with a service beginning and ends with a service completion. Its duration equals the sum of service times of the customers served during the current visit to Q_i. By definition, a visit beginning always corresponds to a switch-over completion, whereas a visit completion corresponds to a switch-over beginning. In the case where there are no customers waiting at Q_i, these two epochs coincide. It is well-known that the mean cycle length is independent of the queue involved (and the service discipline considered in this paper) and is given by (see, for example, [18]) $E(C) = E(S)/(1 - \rho)$.

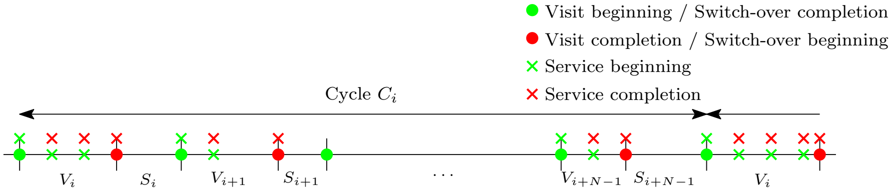

Fig. 1 Description of a cycle, visit periods, and switch-over times

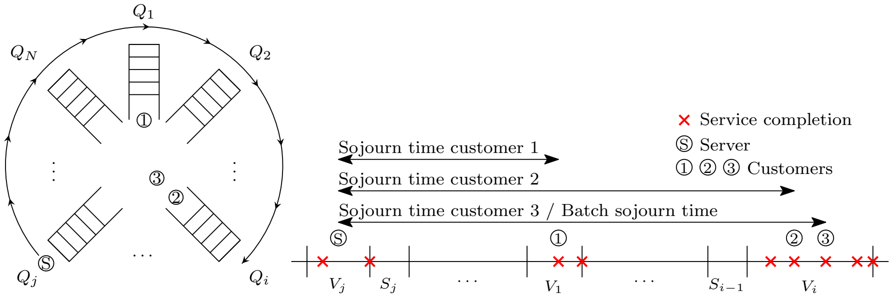

Fig. 2 Description of the batch sojourn-time

In this paper, three different service policies are considered that satisfy the branching property [16]. Under the *exhaustive policy*, when a visit beginning starts at Q_i, the server continues to work until the queue becomes empty. Any customer who arrives during the server's visit to Q_i is also served within the current visit. However, under the *locally gated policy*, the server only serves the customers who were present at Q_i at its visit beginning; all customers who arrive during the course of the visit are served in the next visit to Q_i. The final policy is the *globally gated policy*; according to this policy, the server will only serve the customers who were present at all queues at the visit beginning of a reference queue, which is normally assumed to be Q_1. Customers arriving after this visit beginning will only be served after the server has finished its current cycle. This policy strongly resembles the locally gated policy, except that all queues are gated at the same time instead of one per visit beginning.

The batch sojourn-time of a specific customer batch $\boldsymbol{k}$, denoted by $T_{\boldsymbol{k}}$ and its LST by $\widetilde{T}_{\boldsymbol{k}}(.)$, is defined as the time between its arrival epoch until the service completion of the last customer in the arrived batch; see Fig. 2. In this example, assume that when the server is in a visit period of Q_j, a batch of three customers arrives in Q_1 and Q_i. Then the batch sojourn-time of this batch equals the residual time in V_j, switch-over times $S_j, \ldots, S_{i-1}$, visit periods $V_{j+1}, \ldots, V_{i-1}$, and the time until service completion of the last customer of the batch in V_i. By definition, the batch sojourn-time corresponds to the sojourn-time of the last customer who is served within the batch. It is important to realize that the queue where the batch finishes service *depends* on the location of the server on the arrival of the batch, and there is no fixed order in which the customers need to be served. The order in which the customers are served in this example is the same for the three service policies, but varies between disciplines depending on the location of the server. Finally, the batch sojourn-time of an arbitrary customer batch is denoted by T and its corresponding LST by $\widetilde{T}(.)$.

Throughout this paper, we make references to the server path from Q_i to Q_j, which should be understood in a cyclic sense, for example, $Q_i, Q_{i+1}, \ldots, Q_j$ if $i \leq j$, and otherwise $Q_i, Q_{i+1}, \ldots, Q_N, Q_1, \ldots, Q_j$ if $i > j$. For ease of notation, we define a *cyclic sum* and, analogously, a *cyclic product* as [3]

$$\sum_{l=i}^{j}{}' x_l := \begin{cases} \sum\limits_{l=i}^{j} x_l, & \text{if } i \leq j, \\ \sum\limits_{l=i}^{N} x_l + \sum\limits_{l=1}^{j} x_l, & \text{if } i > j, \end{cases} \quad \prod_{l=i}^{j}{}' x_l := \begin{cases} \prod\limits_{l=i}^{j} x_l, & \text{if } i \leq j, \\ \prod\limits_{l=i}^{N} x_l \times \prod\limits_{l=1}^{j} x_l, & \text{if } i > j, \end{cases}$$

and alternatively,

$$\sum_{l=0}^{j-i}{}' x_{i+l} := \begin{cases} \sum\limits_{l=0}^{j-i} x_{i+l}, & \text{if } i \leq j, \\ \sum\limits_{l=0}^{j+N-i} x_{i+l}, & \text{if } i > j, \end{cases} \quad \prod_{l=0}^{j-i}{}' x_l := \begin{cases} \prod\limits_{l=0}^{j-i} x_{i+l}, & \text{if } i \leq j, \\ \prod\limits_{l=0}^{j+N-i} x_{i+l}, & \text{if } i > j. \end{cases}$$

Finally, let $\mathcal{K}_{i,j}$ be a subset of $\mathcal{K}$ where the last customer of an arbitrary arriving customer batch is served in Q_j and all its other customers are served in $Q_i, \ldots, Q_j$. By definition, a batch will complete its service in one of the queues, such that $\bigcup_{j=1}^{N} \mathcal{K}_{i,j} = \mathcal{K}, i = 1, \ldots, N$. The corresponding probability of subset $\mathcal{K}_{i,j}$ is given by

$$\pi\left(\mathcal{K}_{i,j}\right) = \begin{cases} \mathbb{P}\left(K_j > 0, K_{j+1} = 0, \ldots, K_{i-1} = 0\right), & j = 1, \ldots, N, \ i \neq j+1, \\ \mathbb{P}\left(K_j > 0\right), & \text{otherwise.} \end{cases}$$

In addition, let $E\left(K_l|\mathcal{K}_{i,j}\right)$ be the conditional expected number of customers who have arrived in $Q_l, l = 1, \ldots, N$, given subset $\mathcal{K}_{i,j}$. We define $\widetilde{K}\left(z|\mathcal{K}_{i,j}\right)$ as the conditional PGF of the distribution of the number of customers who arrive in $Q_i, \ldots, Q_j$ given $\mathcal{K}_{i,j}$,

$$\widetilde{K}\left(z|\mathcal{K}_{i,j}\right) = \sum_{\boldsymbol{k} \in \mathcal{K}_{i,j}} \frac{\pi\left(\boldsymbol{k}\right)}{\pi\left(\mathcal{K}_{i,j}\right)} \prod_{l=i}^{j}{}' z_l^{k_l}, \tag{1}$$

such that $\widetilde{K}\left(z\right) = \sum_{j=1}^{N} \pi\left(\mathcal{K}_{i,j}\right) \widetilde{K}\left(z|\mathcal{K}_{i,j}\right), i = 1, \ldots, N$.

4 Exhaustive service

In this section, we start by deriving the LST of the batch sojourn-time distribution of a specific batch of customers in the case of exhaustive service. The batch sojourn-time distribution is found by conditioning on the numbers of customers present in each queue at an arrival epoch and then studying the evolution of the system until all customers within the batch have been served. For this analysis, we first study the joint queue-length distribution at several embedded epochs in Sect. 4.1. We use these results to determine the LST of the batch sojourn-time distribution for both a specific and an arbitrary batch of arriving customers in Sect. 4.2, and present a MVA to calculate the mean batch sojourn-time in Sect. 4.3.

4.1 The joint queue-length distribution

In the polling literature, the probability generating function (PGF) of the joint queue-length distribution at various epochs is extensively studied (for example., [11,13,18]). Let $\widetilde{LB}^{(V_i)}(z)$ and $\widetilde{LC}^{(V_i)}(z)$ be the joint queue-length PGF at *visit* beginnings and completions at Q_i, where $z = (z_1, \ldots, z_N)$ is an N-dimensional vector with $|z_i| \leq 1$. Similarly, let $\widetilde{LB}^{(S_i)}(z)$ and $\widetilde{LC}^{(S_i)}(z)$ be the joint queue-length PGFs at *switch-over* beginnings and completions at Q_i, respectively. Because of the branching property [16], these PGFs can be related to each other as follows:

$$\widetilde{LC}^{(V_i)}(z) = \widetilde{LB}^{(V_i)}\Big(z_1, \ldots, z_{i-1}, \widetilde{BP}_i\left(\lambda - \lambda\widetilde{K}\left(z_1, \ldots, z_{i-1}, 1, z_{i+1}, \ldots, z_N\right)\right), z_{i+1}, \ldots, z_N\Big), \tag{2}$$

$$\widetilde{LB}^{(S_i)}(z) = \widetilde{LC}^{(V_i)}(z), \tag{3}$$

$$\widetilde{LC}^{(S_i)}(z) = \widetilde{LB}^{(S_i)}(z)\,\widetilde{S}_i\left(\lambda - \lambda\widetilde{K}(z)\right), \tag{4}$$

$$\widetilde{LB}^{(V_{i+1})}(z) = \widetilde{LC}^{(S_i)}(z), \tag{5}$$

where $i = 1, \ldots, N$ and $\widetilde{BP}_i(.)$ is the LST of a busy period in Q_i, equals that of an $M^X/G/1$ queue initiated by the service of a customer and is given by

$$\widetilde{BP}_i(\omega) = \widetilde{B}_i\left(\omega + \lambda - \lambda\widetilde{K}_i\left(\widetilde{BP}_i(\omega)\right)\right). \tag{6}$$

Equations (2)–(5) are referred to in the polling literature as the *laws of motion*. The interpretation of (2) is that the queue-length in Q_j, $j \neq i$, at the end of visit period V_i is given by the number of customers already at Q_j at the visit beginning plus all the customers who arrive in the system during visit period V_i. For Q_i, all customers who are already in Q_i or arrive during V_i will be served before the end of the visit completion, and therefore, Q_i will contain no customers at the end of the visit period. Equation (3) simply states that the PGF of a visit completion corresponds to the PGF of the next switch-over beginning (see also Fig. 1). Finally, the queue-length vector at a switch-over completion corresponds to the sum of customers already present at the switch-over beginning plus all the customers who arrive during this switch-over period (4), and by definition the queue-length vector at a switch-over completion is the same for the next visit beginning (5). Note that Eqs. (2)–(5) can be differentiated with respect to $z_1, \ldots, z_N$ to compute moments of the queue-length distributions on embedded points [14] or numerically inverted for the queue-length probability distributions (for example, [6] for the case for non-simultaneous arrivals).

Let $\widetilde{LB}^{(B_i)}(z)$ and $\widetilde{LC}^{(B_i)}(z)$ be the joint queue-length PGFs at *service* beginnings and completions at Q_i. Eisenberg [8] proved that besides the laws of motion, there exists a simple relation between the joint queue-length distributions at *visit*- and *service* beginnings and completions. He observed that each visit beginning either starts with a service beginning, or with a visit completion in the case where there are no customers at the queue. Similarly, each visit completion coincides with either a visit beginning or a service completion. Eisenberg [8] only considered polling systems either with exhaustive or gated service at all queues and individual arriving customers, but [4] has proven that the relation is not restricted to a particular service discipline and also holds for general branching-type service disciplines. In this section, we generalize this result for the case of simultaneous batch arrivals. Similarly to [8], the four PGFs are related as follows:

$$\widetilde{LB}^{(V_i)}(z) + \lambda_i E(C)\,\widetilde{LC}^{(B_i)}(z) = \lambda_i E(C)\,\widetilde{LB}^{(B_i)}(z) + \widetilde{LC}^{(V_i)}(z), \tag{7}$$

where the term $1/(\lambda_i E(C))$ is the long-run ratio between the number of service beginnings/completions and visit beginnings/completions in Q_i, for every $i = 1, \ldots, N$.

Furthermore, the joint queue-length distribution at service beginnings and completions are related via

$$\widetilde{LC}^{(B_i)}(z) = \widetilde{LB}^{(B_i)}(z)\left[\widetilde{B}_i\left(\lambda - \lambda\widetilde{K}(z)\right)/z_i\right]. \tag{8}$$

Substituting (8) in (7) and rearranging terms, the joint queue-length distribution at a service beginning can be written as

$$\widetilde{LB}^{(B_i)}(z) = \frac{z_i\left(\widetilde{LC}^{(V_i)}(z) - \widetilde{LB}^{(V_i)}(z)\right)}{\lambda_i E(C)\left(\widetilde{B}_i\left(\lambda - \lambda\widetilde{K}(z)\right) - z_i\right)}. \tag{9}$$

Next, we can find the PGFs of the joint queue-length distributions at an arbitrary moment during V_i and S_i, denoted by $\tilde{L}^{(V_i)}(z)$ and $\tilde{L}^{(S_i)}(z)$, by noticing that the queue-length at an arbitrary moment in V_i or S_i is equal to the queue-length at service/switch-over beginning plus the number of customers who arrived in the past service/switch-over time,

$$\tilde{L}^{(V_i)}(z) = \widetilde{LB}^{(B_i)}(z)\,\frac{1 - \widetilde{B}_i\left(\lambda - \lambda\widetilde{K}(z)\right)}{E(B_i)\left(\lambda - \lambda\widetilde{K}(z)\right)}, \tag{10}$$

$$\tilde{L}^{(S_i)}(z) = \widetilde{LB}^{(S_i)}(z)\,\frac{1 - \widetilde{S}_i\left(\lambda - \lambda\widetilde{K}(z)\right)}{E(S_i)\left(\lambda - \lambda\widetilde{K}(z)\right)}. \tag{11}$$

Using these results, $\tilde{L}(z)$, which is the PGF of the joint queue-length distribution at an arbitrary moment, can be obtained. By conditioning on periods $V_1, S_1, \ldots, V_N, S_N$ and using (10) and (11) $\tilde{L}(z)$ can be written as

$$\tilde{L}(z) = \frac{1}{E(C)}\sum_{i=1}^{N}\left(E(V_i)\,\tilde{L}^{(V_i)}(z) + E(S_i)\,\tilde{L}^{(S_i)}(z)\right), \tag{12}$$

with $E(V_i) = \rho_i E(C)$ as the expected visit time to Q_i.

4.2 Batch sojourn-time distribution

In order to determine the LST of the steady-state batch sojourn-time distribution, we follow the method of Boon et al. [2] by conditioning on the location of the server and determining the time it takes until the last customer in a specific batch is served. These results are then used to determine the batch sojourn-time distribution of an arbitrary batch. Boon et al. [2] developed this method to study the steady-state waiting time distribution for polling systems with rerouting. For these kinds of models, the distributional form of Little's Law [10] cannot be applied, since the combined processes of internal and external arrivals do not necessarily form a Poisson process. However, by studying the evolution of the system after a customer arrival, this problem can be avoided and the waiting time distribution can be obtained. Important in their analysis is the concept of *descendants* from the theory of branching processes, which are defined as all the customers who arrive during the service of a tagged customer, plus the customers who arrive during the service of those customers, etc. (i.e., the total progeny of the tagged customer).

The approach of Boon et al. [2] is suitable to determine the steady-state batch sojourn-time distribution, since for a specific customer batch the location where the last customer in the batch will be served varies with the location of the server at the arrival of the batch (for example, in Fig. 2 depending of the location of the server the batch is either fully served in Q_1 or Q_i). We explicitly condition on the location of the server; the LST of the batch sojourn-time distribution of a specific customer batch $\boldsymbol{k}$ can be written as

$$\widetilde{T}_{\boldsymbol{k}}(\omega) = \frac{1}{E(C)} \sum_{j=1}^{N} \Big(E\left(V_j\right) \widetilde{T}_{\boldsymbol{k}}^{(V_j)}(\omega) + E\left(S_j\right) \widetilde{T}_{\boldsymbol{k}}^{(S_j)}(\omega) \Big), \tag{13}$$

where $\widetilde{T}_{\boldsymbol{k}}^{(V_j)}(.)$ is the LST of the batch sojourn-time for customer batch $\boldsymbol{k}$ *given* that the batch arrived during V_j, and where $\widetilde{T}_{\boldsymbol{k}}^{(S_j)}(.)$ is *given* that the customer batch arrived during S_j. The remainder of this section will focus on how to determine $\widetilde{T}_{\boldsymbol{k}}^{(V_j)}(.)$, $\widetilde{T}_{\boldsymbol{k}}^{(S_j)}(.)$, and the LST of an arbitrary batch $\widetilde{T}(.)$.

From the theory of branching processes, we denote $B_{j,i},\ i, j = 1, \ldots, N$, as the service of a tagged customer in Q_j plus all its descendants that will be served before or during the next visit to Q_i. Combining this gives the following recursive function:

$$B_{j,i} = \begin{cases} BP_j, & \text{if } i = j, \\ BP_j + \sum_{l=j+1}^{i}{}' \sum_{m=1}^{N_l(BP_j)} B_{l_m,i}, & \text{otherwise}, \end{cases} \tag{14}$$

where BP_j is the busy period initiated by the tagged customer in Q_j, $N_l\left(BP_j\right)$ denotes the number of customers who arrive in Q_l during this busy period in Q_j, and $B_{l_m,i}$ is a sequence of (independent) $B_{l,i}$'s. Let $\widetilde{B}_{j,i}(.)$ be the LST of $B_{j,i}$, which is given by

$$\widetilde{B}_{j,i}(\omega) = \widetilde{BP}_j\left(\omega + \lambda(1 - \widetilde{K}(\boldsymbol{B_{j+1,i}}))\right), \tag{15}$$

where $\boldsymbol{B_{j+1,i}}$ is an N-dimensional vector defined as follows:

$$(\boldsymbol{B_{j,i}})_l = \begin{cases} \widetilde{B}_{l,i}(\omega), & \text{if } l = j, \dots, i, \text{ and } j \neq i+1, \\ 1, & \text{otherwise.} \end{cases} \tag{16}$$

A similar LST can also be formulated for a switch-over time S_j and the service of all its descendants that will be served before the end of the visit to S_i,

$$\widetilde{S}_{j,i}(\omega) = \widetilde{S}_j\left(\omega + \lambda(1 - \widetilde{K}(\boldsymbol{B_{j+1,i}})\right), \tag{17}$$

Finally, let $\boldsymbol{B^*_{j,i}}$ be an N-dimensional vector defined as

$$(\boldsymbol{B^*_{j,i}})_l = \begin{cases} \widetilde{B}_i(\omega), & \text{if } l = i, \\ (\boldsymbol{B_{j,i-1}})_l, & \text{otherwise.} \end{cases} \tag{18}$$

The key difference with (16) is that (18) excludes any new customer arrivals in Q_i. This is needed to omit customers who arrive in Q_i after the batch arrival; these customers do not influence the batch sojourn-time of the arriving customer batch since they will be served afterwards.

We first focus on the batch sojourn-time of a customer batch that arrives during a visit period. Assume than an arriving customer batch $\boldsymbol{k}$ enters the system while the server is currently within visit period V_j and the last customer in the batch will be served in Q_i. Formally, this means $k_i > 0$ and all the other customer arriving in the same batch should be served before the next visit to Q_i; $k_l \geq 0$, $l = j, \dots, i-1$, and $k_l = 0$ elsewhere. Whenever all the customers arrive in the same queue that is currently visited, then $k_i = k_j > 0$, and $k_l = 0$ elsewhere.

The batch sojourn-time of customer batch $\boldsymbol{k}$ consists of (i) the residual service time in Q_j, (ii) the service of all the customers already in the system in $Q_j, \dots, Q_i$, (iii) the service of all new customer arrivals that arrive after customer batch $\boldsymbol{k}$ in $Q_j, \dots, Q_{i-1}$ before the server reaches Q_i, (iv) the switch-over times $S_j, \dots, S_{i-1}$, and (v) the service of the customers in the customer batch $\boldsymbol{k}$. From (10), we know that at the arrival of the customer batch, the PGF of the joint queue-length distribution is the equal to the queue-lengths at a service beginning, $\widetilde{LB}^{(B_j)}(.)$, plus the number of customers who arrived in the elapsed part of the service time, $\widetilde{B}^P_j(.)$. On the other hand, we also need to consider the residual part of the service time, $\widetilde{B}^R_j(.)$, and if $i \neq j$ the arrivals that occur in $Q_j, \dots, Q_{i-1}$ during this period as well. Therefore, similarly to [2], we need to consider the PGF-LST of the joint queue-length distribution at an arrival epoch *and* the residual service time; $\widetilde{L}^{(V_i)}(\boldsymbol{z}, \omega)$. First, since the number of customers who arrive in the elapsed and residual part of the service time are independent of each other and from the queue-lengths at a service beginning, we can write the LST of the joint distribution of $\widetilde{B}^P_j(.)$ and $\widetilde{B}^R_j(.)$ as [7]

$$\widetilde{B}_j^{PR}(\omega_P, \omega_R) = \frac{\widetilde{B}_j(\omega_P) - \widetilde{B}_j(\omega_R)}{E\left(B_j\right)(\omega_R - \omega_P)},$$

Then, because of independence between $\widetilde{B}_j^{PR}(\omega_P, \omega_R)$ and $\widetilde{LB}^{(B_j)}(z)$, we have

$$\widetilde{L}^{(V_j)}(z, \omega) = \widetilde{LB}^{(B_j)}(z)\, \widetilde{B}_j^{PR}\left(\lambda - \lambda \widetilde{K}(z), \omega\right). \tag{19}$$

Proposition 1 *The LST of the batch sojourn-time distribution of batch* $\boldsymbol{k}$ *conditional on the server being in visit period* V_j *and the last customer in the batch being served in* Q_i *is given by*

$$\begin{aligned}\widetilde{T}_{\boldsymbol{k}}^{(V_j)}(\omega) = \widetilde{L}^{(V_j)}&\left(\boldsymbol{B}^*_{\boldsymbol{j,i}},\ \omega + \lambda(1 - \widetilde{K}(\boldsymbol{B_{j,i-1}}))\right)\\ &\times \prod_{l=1}^{i-j}{}' \widetilde{S}_{j+l-1,i-1}(\omega) \frac{1}{(\boldsymbol{B}^*_{\boldsymbol{j,i}})_j} \prod_{l=j}^{i}{}' (\boldsymbol{B}^*_{\boldsymbol{j,i}})_l^{k_l}.\end{aligned} \tag{20}$$

Proof Consider the system just before the arrival of the customer batch and assume that the batch does not finish service in the current visit period, i.e., $i \neq j$. Then, let $n_1, n_2, \ldots, n_N$ be the number of customers present in the system at the arrival epoch of the customer batch and $k_1, \ldots, k_N$ be the number of customers per queue that arrived in batch $\boldsymbol{k}$. Since the batch arrives in V_j, it first has to wait for the residual service time of the customer currently in service. During this period, new customers can arrive before the next visit to Q_i which bring in additional work with $\lambda(1 - \widetilde{K}(\boldsymbol{B_{j,i-1}}))$. Afterwards, each customer already in the system at the arrival of the customer batch in $Q_j, \ldots, Q_i$ and each customer in batch $\boldsymbol{k}$ will make a contribution of $(\boldsymbol{B}^*_{\boldsymbol{j,i}})_l$, $l = j, \ldots, i$, to the batch sojourn-time. Finally, in the switch-over periods between Q_j and Q_i, new customers can arrive who will be served before the service of the last customer in the batch. Combining this gives the LST of the batch sojourn-time distribution of batch $\boldsymbol{k}$ conditional on $n_1, n_2, \ldots, n_N$ customers being already present in the system, the server being in visit period V_j, and the last customer in the batch being served in Q_i:

$$\begin{aligned}E(e^{-\omega T_{\boldsymbol{k}}^{(V_j)}} | n_1, n_2, \ldots, n_N) = \widetilde{B}_j^R&\left(\omega + \lambda(1 - \widetilde{K}(\boldsymbol{B_{j,i-1}}))\right) \widetilde{B}_{j,i-1}(\omega)^{n_j - 1 + k_j}\\ &\times \prod_{l=j+1}^{i-1}{}' \widetilde{B}_{l,i-1}(\omega)^{n_l + k_l} \prod_{l=j}^{i-1}{}' \widetilde{S}_{l,i-1}(\omega)\, \widetilde{B}_i(\omega)^{n_i + k_i}.\end{aligned} \tag{21}$$

Unconditioning this equation gives (20). □

Now, consider a customer batch that arrives during a switch-over period. Assume an arriving customer batch $\boldsymbol{k}$ enters the system while the server is currently within switch-over period S_{j-1} and the last customer in the batch will be served in Q_i. The reason that we consider S_{j-1} is that batch $\boldsymbol{k}$ will finish service in the same queue had it arrived in V_j because of the exhaustive service discipline.

In this case, the batch sojourn-time consists of the same components (ii), (iii), (iv), and (v). Component (i) is however different and is now defined as the residual switch-over time between Q_{j-1} and Q_j. Similarly, we define $\widetilde{L}^{(S_{j-1})}(z, \omega)$ as the PGF-LST of the joint queue-length distribution of customers present in the system at an arbitrary moment during S_{j-1} *and* the residual switch-over time $\widetilde{S}_{j-1}^{R}(.)$. From (11), we have the joint queue-length distribution at a switch-over beginning, $\widetilde{LB}^{(S_{j-1})}(.)$, and the number of customers who arrived in the elapsed part of the switch-over time, $\widetilde{S}_{j-1}^{P}(.)$. Similarly to $\widetilde{B}_{j}^{PR}(.)$, we define $\widetilde{S}_{j-1}^{PR}(\omega_R, \omega_P)$ as the LST of the joint distribution of the elapsed and residual switch-over time S_{j-1} as

$$\widetilde{S}_{j-1}^{PR}(\omega_P, \omega_R) = \frac{\widetilde{S}_{j-1}(\omega_P) - \widetilde{S}_{j-1}(\omega_R)}{E\left(S_{j-1}\right)(\omega_R - \omega_P)}.$$

Then, due to independence, the PGF-LST of the joint queue-length distribution present at an arbitrary moment during S_{j-1} and the residual switch-over time is given by

$$\widetilde{L}^{(S_{j-1})}(z, \omega) = \widetilde{LB}^{(S_{j-1})}(z)\, \widetilde{S}_{j-1}^{PR}\left(\lambda - \lambda \widetilde{K}(z), \omega\right). \tag{22}$$

Proposition 2 *The LST of the batch sojourn-time distribution of batch* $\boldsymbol{k}$ *conditional on the server being in switch-over period* S_{j-1} *and the last customer in the batch being served in* Q_i *is given by*

$$\widetilde{T}_{\boldsymbol{k}}^{(S_{j-1})}(\omega) = \widetilde{L}^{(S_{j-1})}\left(\boldsymbol{B}^{*}_{\boldsymbol{j,i}},\ \omega + \lambda(1 - \widetilde{K}(\boldsymbol{B}_{\boldsymbol{j,i-1}}))\right) \times \prod_{l=1}^{i-j}{}' \widetilde{S}_{j+l-1,i-1}(\omega) \prod_{l=j}^{i}{}' (\boldsymbol{B}^{*}_{\boldsymbol{j,i}})_{l}^{k_l}. \tag{23}$$

Proof Similarly to Proposition 1, we condition on the number of customers present in the system before the arrival of batch $\boldsymbol{k}$ and the number of customer who enter the system per queue that arrived in batch $\boldsymbol{k}$. Then, studying the contribution of each customer to the batch sojourn-time, we obtain (23). □

From Propositions 1 and 2, it can be seen that the LST of the batch sojourn-time distribution of batch $\boldsymbol{k}$ conditioned on a visit/switch-over period is comprised of two terms: a term independent of batch $\boldsymbol{k}$ *and* a term that corresponds to the additional contribution batch $\boldsymbol{k}$ makes to the batch sojourn-time:

$$\widetilde{T}_{\boldsymbol{k}}^{(V_j)}(\omega) = \sum_{i=1}^{N} 1_{(\boldsymbol{k} \in \mathcal{K}_{j,i})} \widetilde{W}_{i}^{(V_j)}(\omega) \prod_{l=j}^{i}{}' (\boldsymbol{B}^{*}_{\boldsymbol{j,i}})_{l}^{k_l}, \tag{24}$$

$$\widetilde{T}_{\boldsymbol{k}}^{(S_{j-1})}(\omega) = \sum_{i=1}^{N} 1_{(\boldsymbol{k} \in \mathcal{K}_{j,i})} \widetilde{W}_{i}^{(S_{j-1})}(\omega) \prod_{l=j}^{i}{}' (\boldsymbol{B}^{*}_{\boldsymbol{j,i}})_{l}^{k_l}, \tag{25}$$

where $1_{(\boldsymbol{k}\in\mathcal{K}_{j,i})}$ is an indicator function that is equal to one if all customers in batch $\boldsymbol{k}$ are served in $Q_j, \ldots, Q_i$ and the last customer will be served in Q_i, and zero otherwise. The terms $\widetilde{W}_i^{(V_j)}(\omega)$ and $\widetilde{W}_i^{(S_{j-1})}(\omega)$ can be considered as the time between the batch arrival epoch and the service completion of the last customer in Q_i that was already in the system at the arrival of the customer batch, excluding batch $\boldsymbol{k}$ and any arrivals to Q_i after the arrival epoch, conditioned on the location of the server. In the case where there are only individually arriving customers, this would correspond to the LST of the waiting time distribution of a customer arriving in Q_i conditional on the server being in a visit or switch-over period. The LST of the batch sojourn-time distribution of a specific customer batch $\boldsymbol{k}$ can now be calculated using (13).

Finally, we focus on the LST of the batch sojourn-time of an arbitrary batch $\widetilde{T}(.)$.

Theorem 1 *The LST of the batch sojourn-time distribution of an arbitrary batch $\widetilde{T}(.)$ in the case of exhaustive service is given by*

$$\widetilde{T}(\omega) = \sum_{\boldsymbol{k}\in\mathcal{K}} \pi(\boldsymbol{k})\, \widetilde{T}_{\boldsymbol{k}}(\omega), \tag{26}$$

where $\widetilde{T}_{\boldsymbol{k}}(\omega)$ is given by (13). *Alternatively, we can write* (26) *as*

$$\widetilde{T}(\omega) = \frac{1}{E(C)} \sum_{j=1}^{N}\sum_{i=1}^{N} \left(E(V_j)\, \widetilde{W}_i^{(V_j)}(\omega) + E(S_{j-1})\, \widetilde{W}_i^{(S_{j-1})}(\omega)\right) \times \pi(\mathcal{K}_{j,i})\, \widetilde{K}\left(\boldsymbol{B}^*_{\boldsymbol{j,i}}|\mathcal{K}_{j,i}\right). \tag{27}$$

Proof It can be easily seen that (26) follows by enumerating all possible realizations of customer batches and the law of total probability.

Next, for (27), we can partition $\mathcal{K}$ into $\mathcal{K}_{j,i}$ and write (26) using (13) as

$$\widetilde{T}(\omega) = \frac{1}{E(C)} \sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{\boldsymbol{k}\in\mathcal{K}_{j,i}} \pi(\boldsymbol{k}) \left(E(V_j)\, \widetilde{T}_{\boldsymbol{k}}^{(V_j)}(\omega) + E(S_{j-1})\, \widetilde{T}_{\boldsymbol{k}}^{(S_{j-1})}(\omega)\right). \tag{28}$$

From (24) and (25), it can be seen that when the server is either in S_{j-1} or V_j, then for two different customer batches that both finish service in the same queue, their LST of the batch sojourn-time distribution only varies in the contribution the batch makes to the batch sojourn-time.

Then, by (26) and (1), we have by rearrangement

$$\sum_{\boldsymbol{k}\in\mathcal{K}_{j,i}} \pi(\boldsymbol{k}) \left(E(V_j)\, \widetilde{T}_{\boldsymbol{k}}^{(V_j)}(\omega) + E(S_{j-1})\, \widetilde{T}_{\boldsymbol{k}}^{(S_{j-1})}(\omega)\right)$$
$$= \left(E(V_j)\, \widetilde{W}_i^{(V_j)}(\omega) + E(S_{j-1})\, \widetilde{W}_i^{(S_{j-1})}(\omega)\right) \pi(\mathcal{K}_{j,i})$$

$$\times \sum_{\boldsymbol{k}\in\mathcal{K}_{j,i}} \frac{\pi\left(\boldsymbol{k}\right)}{\pi\left(\mathcal{K}_{j,i}\right)} \prod_{l=j}^{i}{}' \left(\boldsymbol{B^*_{j,i}}\right)_l^{k_l}$$

$$= \left(E\left(V_j\right)\widetilde{W}_i^{(V_j)}\left(\omega\right) + E\left(S_{j-1}\right)\widetilde{W}_i^{(S_{j-1})}\left(\omega\right)\right)\pi\left(\mathcal{K}_{j,i}\right)\widetilde{K}\left(\boldsymbol{B^*_{j,i}}|\mathcal{K}_{j,i}\right).$$

Substituting the last equation in (28) gives (27). □

Differentiating (27) will give the mean batch sojourn-time; however, in the next section, an alternative, more efficient way to determine the mean batch sojourn-time is presented.

4.3 Mean batch sojourn-time

In this section, we derive the mean batch sojourn-time of a specific batch and an arbitrary batch using *MVA*. MVA for polling systems was developed by Winands et al. [23] to study mean waiting times in systems with exhaustive, gated service, or mixed service. The main advantage of MVA is that it has a pure probabilistic interpretation and is based on standard queueing results, i.e., the Poisson arrivals see time averages (PASTA) property [25] and Little's Law [15]. Furthermore, MVA evaluates the polling system at arbitrary time periods and not on embedded points such as visit beginnings, like in the buffer occupancy method [18] and the descendant set approach [12].

Central in MVA [23] is the derivation of $E\left(\bar{L}_i^{(S_{j-1},V_j)}\right)$, the mean queue-length at Q_i (excluding the potential customer currently in service) at an arbitrary epoch within switch-over period S_{j-1} and visit period V_j :

$$E\left(\bar{L}_i^{(S_{j-1},V_j)}\right) = \frac{E\left(S_{j-1}\right)}{E\left(S_{j-1}\right)+E\left(V_j\right)}E\left(\bar{L}_i^{(S_{j-1})}\right) + \frac{E\left(V_j\right)}{E\left(S_{j-1}\right)+E\left(V_j\right)}E\left(\bar{L}_i^{(V_j)}\right), \tag{29}$$

where $E\left(\bar{L}_i^{(S_{j-1})}\right)$ and $E\left(\bar{L}_i^{(V_j)}\right)$ are the expected queue-length in Q_i during, respectively, a switch-over/visit period and $E\left(V_j\right) = \rho_j E\left(C\right)$. Subsequently, with $E\left(\bar{L}_i^{(S_{j-1};V_j)}\right)$ the mean queue-length $E\left(\bar{L}_i\right)$ in Q_i can be determined:

$$E\left(\bar{L}_i\right) = \sum_{j=1}^{N} \frac{E\left(S_{j-1}\right)+E\left(V_j\right)}{E\left(C\right)} E\left(\bar{L}_i^{(S_{j-1},V_j)}\right), \quad i=1,\ldots,N, \tag{30}$$

and by Little's law, also the mean waiting time $E\left(W_i\right)$ of a random customer in Q_i, which is defined as the time in steady state from the customer's arrival until the start of his/her service.

For notational purposes, we introduce θ_j as short-hand for the intervisit period $\left(S_{j-1}, V_j\right)$; the expected duration of this period $E\left(\theta_j\right)$ is given by

$$E\left(\theta_j\right) = E\left(S_{j-1}\right) + E\left(V_j\right), \quad j = 1, \ldots, N. \tag{31}$$

Notice that $\sum_{j=1}^{N} E\left(\theta_j\right) = E\left(C\right)$. In addition, we define $\theta_{j,i}$ as the duration of an intervisit period starting in θ_j and ending in θ_i, the expected duration of this period $E\left(\theta_{j,i}\right)$ is equal to

$$E\left(\theta_{j,i}\right) = \sum_{l=j}^{i}{}' E\left(\theta_l\right), \quad i = 1, \ldots, N, \ j = 1, \ldots, N, \tag{32}$$

and where $E\left(\theta_{j,i}^R\right) = E\left(\theta_{j,i}^2\right)/2E\left(\theta_{j,i}\right)$ is the mean residual duration of this period. However, $E\left(\theta_{j,i}^2\right)$ is unknown and not straightforward to derive directly. In the MVA, based on probabilistic arguments, $E\left(\theta_{j,i}^2\right)$ will be expressed in terms of $E\left(\bar{L}_i^{(\theta_j)}\right)$.

We denote $E\left(B_{j,i}\right)$ as the mean service of a customer in Q_j and all its descendants *before* the server starts serving Q_i. Let $E\left(B_{j,j}\right) = E\left(B_j\right)$ and $E\left(B_{j,j+1}\right) = E\left(B_j\right)/\left(1 - \rho_j\right)$ be the expected busy period initiated by a customer in Q_j. Then, $E\left(B_{j,j+2}\right)$ equals the busy period in Q_j plus all the customers who arrive during this busy period in Q_{j+1} and the busy periods that they trigger:

$$E\left(B_{j,j+2}\right) = \frac{E\left(B_j\right)}{1 - \rho_j}\left(1 + \frac{\rho_{j+1}}{1 - \rho_{j+1}}\right) = \frac{E\left(B_j\right)}{\left(1 - \rho_j\right)\left(1 - \rho_{j+1}\right)}.$$

In general, we can write $E\left(B_{j,i}\right)$ for $i \neq j$ as

$$E\left(B_{j,i}\right) = \frac{E\left(B_j\right)}{\prod_{l=j}^{\prime i-1}\left(1 - \rho_l\right)}, \quad i = 1, \ldots, N, \ j = 1, \ldots, N. \tag{33}$$

Also, let $E\left(S_{j,i}\right)$ denote the switch-over in Q_j and the service of all the customers who arrive during $E\left(S_j\right)$ and their descendants *before* the server starts serving Q_i. Then $E\left(S_{j,j+1}\right) = E\left(S_j\right)$ and, in general, for $i \neq j + 1$,

$$E\left(S_{j,i}\right) = \frac{E\left(S_j\right)}{\prod_{l=j+1}^{\prime i-1}\left(1 - \rho_l\right)}, \quad i = 1, \ldots, N, \ j = 1, \ldots, N. \tag{34}$$

Finally, $E\left(B_{j,i}^R\right)$ is the mean residual service of a customer in Q_j and all its descendants *before* the server starts serving Q_i and is given by replacing $E\left(B_j\right)$ by

$E\left(B_j^R\right) = E\left(B_j^2\right)/2E\left(B_j\right)$ in $E\left(B_{j,i}\right)$. In addition, $E\left(S_{j,i}^R\right)$ is defined as $E\left(S_{j,i}\right)$ and by replacing $E\left(S_j\right)$ by $E\left(S_j^R\right) = E\left(S_j^2\right)/2E\left(S_j\right)$.

In MVA, a set of N^2 linear equations is derived for $E\left(\bar{L}_i\right)$ in terms of unknowns $E\left(\bar{L}_i^{(\theta_j)}\right)$. For this, we have to consider the waiting time of an arbitrary customer and make use of the arrival relation and the PASTA property. Assume that an arbitrary customer enters the system in Q_i. The waiting time of the customer consists of (i) the service of $E\left(\bar{L}_i\right)$ customers already at Q_i upon its arrival to the system, (ii) the service of $E\left(K_{ii}\right)/2E\left(K_i\right)$ customers who arrived in the same customer batch, but are placed before the arbitrary customer in Q_i, (iii) if the server is currently in intervisit period θ_i, then the arbitrary customer has to wait with probability ρ_i for the residual service time $E\left(B_i^R\right)$ and with probability $E\left(S_{i-1}\right)/E\left(C\right)$ for the residual switch-over time $E\left(S_{i-1}^R\right)$. Finally, (iv) whenever the server is not in intervisit period θ_i, the arbitrary customer has to wait for the expected residual duration before the server returns at Q_i. Based on these components, the mean waiting time $E\left(W_i\right)$ of a customer in Q_i, $i = 1, \ldots, N$, is given by

$$E\left(W_i\right) = E\left(\bar{L}_i\right)E\left(B_i\right) + \frac{E\left(K_{ii}\right)}{2E\left(K_i\right)}E\left(B_i\right) + \rho_i E\left(B_i^R\right) + \frac{E\left(S_{i-1}\right)}{E\left(C\right)}E\left(S_{i-1}^R\right) + \left(1 - \frac{E\left(\theta_i\right)}{E\left(C\right)}\right)\left(E\left(\theta_{i+1,i-1}^R\right) + E\left(S_{i-1}\right)\right). \tag{35}$$

The next step to derive the equations is to relate the unknowns $E\left(\theta_{i+1,i-1}^R\right)$ to $E\left(\bar{L}_i^{(\theta_j)}\right)$. Consider $E\left(\theta_{j,i}^R\right)$, the expected residual duration of an intervisit period starting in θ_j and ending in θ_i given that an arbitrary customer batch just entered the system. Then with probability $E\left(\theta_l\right)/E\left(\theta_{j,i}\right)$, the server is during this period in intervisit period θ_l, $l = j, \ldots, i$, and the expected residual duration until the intervisit ending of θ_i, conditional on the server being in intervisit period θ_l, is defined as follows. First, with probability $E\left(V_l\right)/E\left(\theta_l\right)$, the server is busy serving a customer in Q_l and with probability $E\left(S_{l-1}\right)/E\left(\theta_l\right)$, the server is in switch-over period S_{l-1}. During the residual service/switch-over time, new customers can arrive who will be served before the intervisit ending in θ_i, which equals $E\left(B_{l,i+1}^R\right)$ and $E\left(S_{l-1,i+1}^R\right)$, respectively. In addition, the expected number of customers in Q_n given the server is in θ_l, $E\left(\bar{L}_n^{(\theta_l)}\right)$, and the expected number of customers $E\left(K_{nl}\right)/E\left(K_n\right)$ who arrived in Q_n in the arbitrary customer batch will increase the duration of $E\left(\theta_{j,i}^R\right)$ by $E\left(B_{n,i+1}\right)$. Finally, the customer also has to wait for all the switch-over times $E\left(S_{n,i+1}\right)$, $n = j, \ldots, i$, between Q_n to Q_{n+1} plus the customers who arrive during the switch-over times and their descendants that will be served before the end of $E\left(\theta_{j,i}^R\right)$. Combining this gives the following expression for $i \neq j - 1$:

$$E\left(\theta_{j,i}^R\right) = \sum_{l=j}^{i}{}' \frac{E\left(\theta_l\right)}{E\left(\theta_{j,i}\right)}\left(\frac{E\left(V_l\right)}{E\left(\theta_l\right)} E\left(B_{l,i+1}^R\right) + \frac{E\left(S_{l-1}\right)}{E\left(\theta_l\right)} E\left(S_{l-1,i+1}^R\right)\right.$$
$$\left. + \sum_{n=l}^{i}{}' \left[\frac{E\left(K_{nl}\right)}{E\left(K_n\right)} + E\left(\bar{L}_n^{(\theta_l)}\right)\right] E\left(B_{n,i+1}\right) + \sum_{n=1}^{i-l}{}' E\left(S_{l+n-1,i+1}\right)\right). \tag{36}$$

It is now possible to set up a set of N^2 linear equations. First, after the server has visited Q_i, there will be no customers present in the queue. Therefore, the number of customers in Q_i given an arbitrary moment in an intervisit period starting in θ_{i+1} and ending in θ_j equals the number of Poisson arrivals during the age of this period [23]. Because the age is equal to the residual time in distribution, we have, for $i = 1, \ldots, N,\ j = 1, \ldots, N$, and $i \neq j$,

$$\sum_{l=i+1}^{j}{}' \frac{E\left(\theta_l\right)}{E\left(\theta_{i+1,j}\right)} E\left(\bar{L}_i^{(\theta_l)}\right) = \lambda_i E\left(\theta_{i+1,j}^R\right). \tag{37}$$

Second, by (35) and using Little's Law, $\lambda_i E\left(W_i\right) = E\left(\bar{L}_i\right)$. Substituting this into (30) gives, for $i = 1, 2 \ldots, N$,

$$\sum_{j=1}^{N} \frac{E\left(\theta_j\right)}{E\left(C\right)} E\left(\bar{L}_i^{(\theta_j)}\right) = \frac{\lambda_i}{1-\rho_i}\left(\frac{E\left(K_{ii}\right)}{2E\left(K_i\right)} E\left(B_i\right) + \rho_i E\left(B_i^R\right) \frac{E\left(S_{i-1}\right)}{E\left(C\right)} E\left(S_{i-1}^R\right)\right.$$
$$\left. + \left(1 - \frac{E\left(\theta_i\right)}{E\left(C\right)}\right)\left(E\left(\theta_{i+1,i-1}^R\right) + E\left(S_{i-1}\right)\right)\right). \tag{38}$$

With (37) and (38), a set of N^2 linear equations for unknowns $E\left(\bar{L}_i^{(\theta_j)}\right)$ are now defined. Solving the set of linear equations and by (30) and (35) will give the expected queue-lengths and waiting times.

In order to derive the mean batch sojourn-time $E\left(T_{\boldsymbol{k}}\right)$ of customer batch $\boldsymbol{k}$, $E\left(\bar{L}_i^{(\theta_j)}\right)$ also plays an integral role. Similarly to (13), in order to calculate the expected batch sojourn-time distribution of a specific customer batch $\boldsymbol{k}$, we explicitly condition on the location on the server:

$$E\left(T_{\boldsymbol{k}}\right) = \frac{1}{E\left(C\right)} \sum_{j=1}^{N} E\left(\theta_j\right) E\left(T_{\boldsymbol{k}}^{(\theta_j)}\right), \tag{39}$$

where $E\left(T_{\boldsymbol{k}}^{(\theta_j)}\right)$ is the expected batch sojourn-time distribution of a specific customer batch $\boldsymbol{k}$ given that the server is in intervisit period θ_j. $E\left(T_{\boldsymbol{k}}^{(\theta_j)}\right)$ can be derived in a similar way to (36). This gives the following expression:

$$E\left(T_{\boldsymbol{k}}^{(\theta_j)}\right) = \frac{E\left(V_j\right)}{E\left(\theta_j\right)} E\left(B_{j,i}^R\right) + \frac{E\left(S_{j-1}\right)}{E\left(\theta_j\right)} E\left(S_{j-1,i}^R\right) + {\sum_{l=j}^{i}}' E\left(\bar{L}_l^{(\theta_j)}\right) E\left(B_{l,i}\right)$$
$$+ {\sum_{l=j}^{i}}' k_l E\left(B_{l,i}\right) + {\sum_{n=1}^{i-j}}' E\left(S_{j+n-1,i}\right). \tag{40}$$

Note that the same decomposition as (24) and (25) also holds for the expected batch sojourn-time:

$$E\left(T_{\boldsymbol{k}}^{(\theta_j)}\right) = \sum_{i=1}^{N} 1_{(\boldsymbol{k}\in\mathcal{K}_{j,i})} \left[E\left(W_i^{(\theta_j)}\right) + {\sum_{l=j}^{i}}' k_l E\left(B_{l,i}\right) \right],$$

where $E\left(W_i^{(\theta_j)}\right)$ is the expected time between the batch arrival epoch and the service completion of the last customer in Q_i that is already in the system, excluding any arrivals to Q_i after the arrival epoch. The term ${\sum'}_{l=j}^{i} k_l E\left(B_{l,i}\right)$ can be interpreted as the total contribution batch $\boldsymbol{k}$ makes to the batch sojourn-time.

Finally, the expected batch sojourn-time of an arbitrary customer batch is obtained by multiplying $E\left(T_{\boldsymbol{k}}\right)$ with the probability that a particular batch $\boldsymbol{k}$ enters the system:

$$E\left(T\right) = \sum_{\boldsymbol{k}\in\mathcal{K}} \pi\left(\boldsymbol{k}\right) E\left(T_{\boldsymbol{k}}\right). \tag{41}$$

However, if there are many different realizations of customer batches possible, (41) might not be computationally feasible, since for every $\boldsymbol{k}$ we have to determine the mean batch sojourn-time given that the server starts in intervisit period θ_j and ends in θ_i; in total, there are $|\mathcal{K}| \times N \times N$ combinations to consider, where $|\mathcal{K}|$ denotes the size of set $\mathcal{K}$. Instead, by using $E\left(K_l|\mathcal{K}_{j,i}\right)$, we can rewrite (41) as follows:

$$\begin{aligned}
E\left(T\right) &= \frac{1}{E\left(C\right)} \sum_{j=1}^{N}\sum_{i=1}^{N}\sum_{\boldsymbol{k}\in\mathcal{K}_{j,i}} \pi\left(\boldsymbol{k}\right) E\left(\theta_j\right) E\left(T_{\boldsymbol{k}}^{(\theta_j)}\right) \\
&= \frac{1}{E\left(C\right)} \sum_{j=1}^{N}\sum_{i=1}^{N} E\left(\theta_j\right) \sum_{\boldsymbol{k}\in\mathcal{K}_{j,i}} \pi\left(\boldsymbol{k}\right) \left(E\left(W_i^{(\theta_j)}\right) + {\sum_{l=j}^{i}}' k_l E\left(B_{l,i}\right) \right) \\
&= \frac{1}{E\left(C\right)} \sum_{j=1}^{N}\sum_{i=1}^{N} E\left(\theta_j\right) \pi\left(\mathcal{K}_{j,i}\right) \left(E\left(W_i^{(\theta_j)}\right) + {\sum_{l=j}^{i}}' E\left(K_l|\mathcal{K}_{j,i}\right) E\left(B_{l,i}\right) \right).
\end{aligned}$$

The advantage is that the number of combinations reduces to $N \times N$, and $\pi\left(\mathcal{K}_{j,i}\right)$ can be determined in $|\mathcal{K}|$ steps.

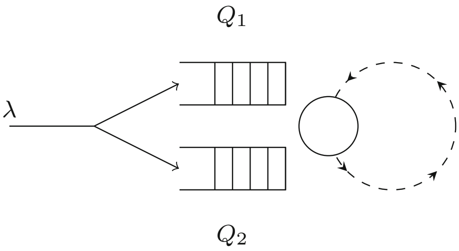

Fig. 3 A symmetrical polling system with two exponential queues

5 Numerical results

In this section we investigate the batch sojourn-times for the three server disciplines. In Sect. 5.1 we study a symmetrical polling system with two queues and derive a closed-form solution for the expected batch sojourn-times and show under which parameters settings, which service discipline has the smallest expected batch sojourn-time. In Sect. 5.2 we study asymmetrical systems and show that the service discipline that achieves the shortest expected batch sojourn-time depends on the system parameters.

5.1 A symmetrical polling system with two exponential queues

Consider a symmetrical polling system with two queues where all customers arrive in pairs and each of them joins another queue as shown in Fig. 3. Assume that the arrival rate is λ, the expected service time of a customer in Q_1 or Q_2 is $E(B_1) = E(B_2) = b$, and the expected switch-over time from Q_1 to Q_2 and vice versa is $E(S_1) = E(S_2) = s$. In addition, we make the assumption that both service times and switch-over times are exponentially distributed, i.e., $E\left(B_1^R\right) = E\left(B_2^R\right) = b$ and $E\left(S_1^R\right) = E\left(S_2^R\right) = s$. Since customers arrive in pairs, $E(K_1) = E(K_2) = 1$, and $E(K_{12}) = E(K_{21}) = 1$ and $E(K_{11}) = E(K_{22}) = 0$. Finally, the overall system load is $\rho = \rho_1 + \rho_2 = 2b\lambda$.

In Fig. 4, a comparison is made between the mean batch sojourn-time and its variance for exhaustive and locally gated service. We excluded the results for globally gated since in this case it is always dominated by locally gated. The mean batch sojourn-times are obtained from MVA, and using (41), the mean batch sojourn-time in the case of exhaustive service is given by

$$E\left(T^{\mathrm{EX}}\right) = \frac{0.25\rho^2 b - 0.25\rho^2 s - \rho s + 2b + 2s}{1-\rho}, \tag{42}$$

and in the case of locally gated service

$$E\left(T^{\mathrm{LG}}\right) = \frac{-0.125\rho^3 b + 0.125\rho^3 s + 0.25\rho^2 b - 0.5\rho^2 s + 0.5\rho b + \rho s + 2b + 2s}{(1+0.5\rho)(1-\rho)}. \tag{43}$$

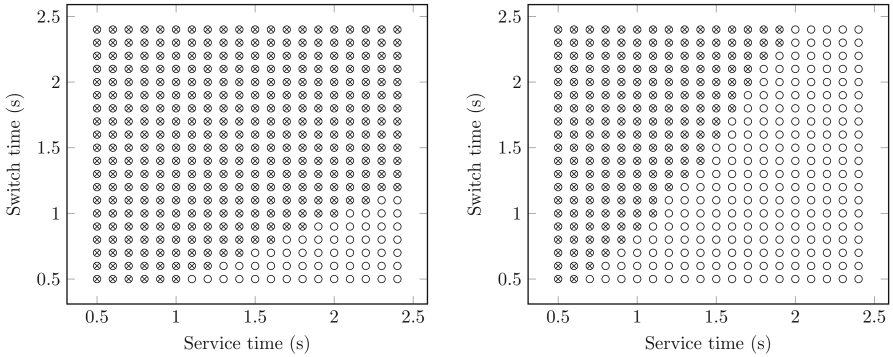

Fig. 4 Batch sojourn-time for the symmetrical polling system with two queues. ○ means locally gated is better, ⊗ means exhaustive is better. **a** Mean batch sojourn-time, **b** variance batch sojourn-time

Table 1 Parameters for three polling models

Q_i	*Model a*			*Model b*			*Model c*		
	1	2	3	1	2	3	1	2	3
$E(B_i)$	1.00	1.00	1.00	1.00	1.00	1.00	0.10	0.40	0.90
$E\left(B_i^{(2)}\right)$	2.00	2.00	2.00	2.00	2.00	2.00	1.00	1.00	1.00
$E(S_i)$	0.10	0.10	0.10	1.00	1.00	1.00	1.00	1.00	1.00
$E\left(S_i^{(2)}\right)$	0.02	0.02	0.02	2.00	2.00	2.00	1.00	1.00	1.00
$\boldsymbol{k} \in \mathcal{K}$	$\pi(1,1,0) = 1/4$ $\pi(3,0,1) = 3/4$			$\pi(1,0,0) = 1/3$ $\pi(0,1,0) = 1/3$ $\pi(0,0,1) = 1/3$			$\pi(1,1,0) = 4/5$ $\pi(1,0,3) = 1/5$		

In order to obtain the variance of the batch sojourn-time, we numerically invert (26) using the algorithm from Choudhury and Whitt [6], adapted for the case of batch arrivals.

Now, we can compare the batch sojourn-times for the symmetrical polling system and investigate under which parameter settings which service discipline achieves the smallest expected batch sojourn-time. Figure 4 shows the combinations of service and switch-over times where a specific service discipline achieves the smallest batch sojourn-time. It can be seen that when the switch-over times are longer compared to the service times, the exhaustive service discipline achieves the smallest expected batch sojourn-time, since it is more beneficial to serve all customers at the current queue first before moving to the other queue. However, if the service times are longer than the switch-over times, it is better to switch to the other queue more often, because otherwise the server will spend too much time serving customers in one queue and it will take a long time before a customer batch is completely served. In this case, locally gated performs better than exhaustive service. The same pattern can also be observed for the variance.

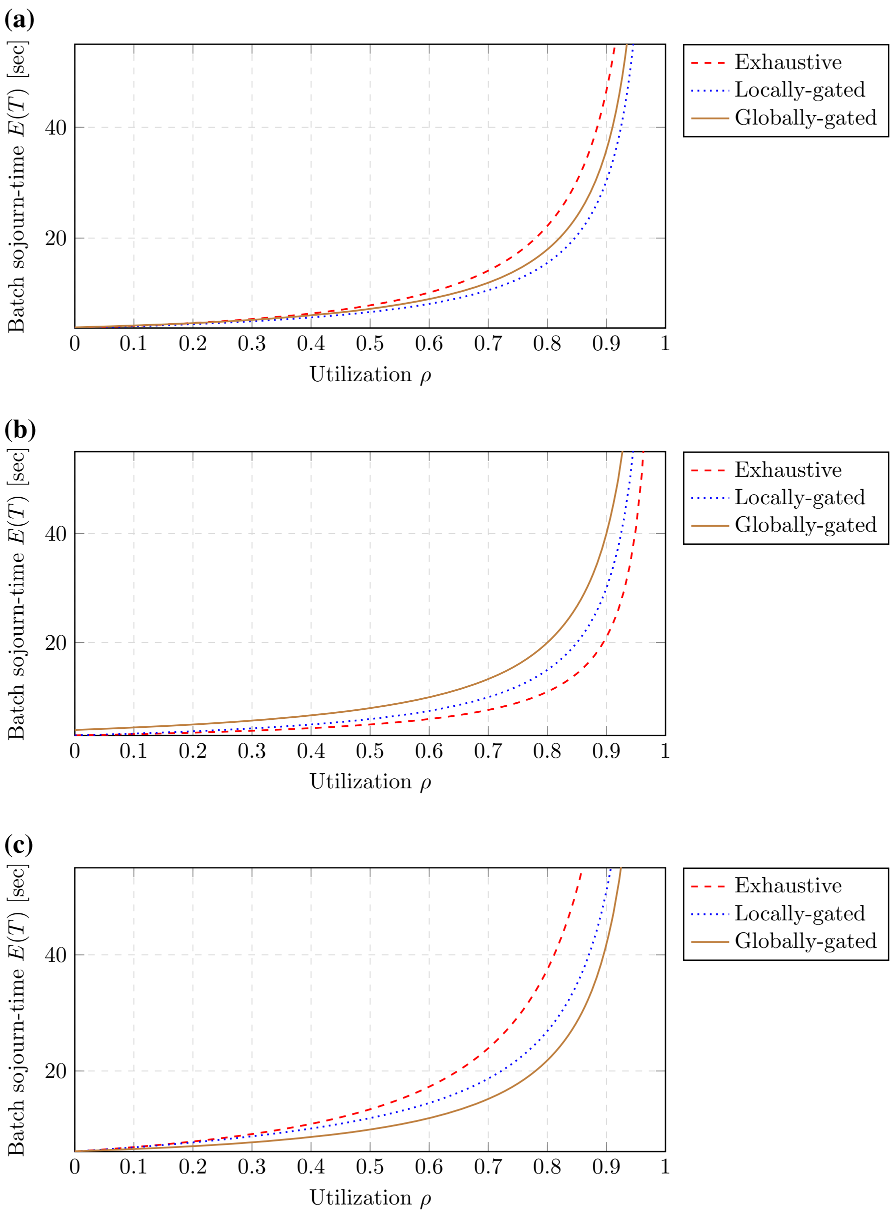

Fig. 5 Expected batch sojourn-time for various utilizations for three different systems. **a** Locally gated minimizes the expected batch sojourn-time, **b** exhaustive minimizes the expected batch sojourn-time, **c** globally gated minimizes the expected batch sojourn-time

5.2 Asymmetrical polling systems with multiple queues

In the previous section we have shown that depending on the system parameters, exhaustive service or locally gated service minimizes the expected batch sojourn-time. However, it can be shown that *any* of the three service disciplines studied in this paper can minimize the expected batch sojourn-time. In Table 1, the parameters of three systems with $N = 3$ are given. *Model a* has short switch-over times, *Model b* is a system with individual arriving customers and equal switch-over times and service times, and in *Model c* the last queue is the slowest and receives most of the work. Using the results of Sect. 4.3, and the online appendix the expected batch sojourn-times for the three different models can be calculated. The batch sojourn-times are shown in Fig. 5 for $0 \leq \rho < 1$. The results of *Model a* in Fig. 5a show that locally gated achieves the lowest expected batch sojourn-times, which is similar to Sect. 5.1 when the switch-over times were short. From the results of *Model b* shown in Fig. 5b, it can be seen that exhaustive service has the lowest expected batch sojourn-times. Here it is beneficial to serve a customer arriving to the same queue that is currently being served, since otherwise this customer has to wait a full cycle which increases the mean batch sojourn-time. Finally, *Model c* in Fig. 5c shows that globally gated service achieves the lowest expected batch sojourn-times, since for this policy the server will switch more often between the queues and finish service for all customers in a batch during one cycle, compared to the other disciplines.

6 Conclusion and further research

In this paper we analyzed the batch sojourn-time in a cyclic polling system with simultaneous batch arrivals and obtained exact expressions for the Laplace–Stieltjes transform of the steady-state batch sojourn-time distribution for the locally gated, globally gated, and exhaustive service disciplines. Also, we provided a more efficient way to determine the mean batch sojourn-time using MVA. We compared the batch sojourn-times for the different service disciplines in several numerical examples and showed that the best performing service discipline, minimizing the batch sojourn-time, depends on system characteristics.

A further research topic would be to determine, for each of the three policies, under what conditions on the system parameters its mean batch sojourn-time is smaller than that of the other two, and whether alternative service disciplines can achieve even lower batch sojourn-times. Another interesting further research topic would be to study how the customers of an arriving customer batch should be allocated over the various queues in order to minimize the batch sojourn-times.

References

1. Boon, M.A.A., Van der Mei, R.D., Winands, E.M.M.: Applications of polling systems. Surv. Oper. Res. Manag. Sci. **16**(2), 67–82 (2011)
2. Boon, M.A.A., Van der Mei, R.D., Winands, E.M.M.: Waiting times in queueing networks with a single shared server. Queueing Syst. **74**(4), 403–429 (2012). doi:10.1007/s11134-012-9334-6
3. Boxma, O.J., Groenendijk, W.P., Weststrate, J.A.: A pseudoconservation law for service systems with a polling table. IEEE Trans. Commun. **38**(10), 1865–1870 (1990). doi:10.1109/26.61458
4. Boxma, O.J., Kella, O., Kosinski, K.M.: Queue lengths and workloads in polling systems. Oper. Res. Lett. **39**(6), 401–405 (2011). doi:10.1016/j.orl.2011.10.006
5. Chiarawongse, J., Srinivasan, M.M.: On pseudo-conservation laws for the cyclic server system with compound Poisson arrivals. Oper. Res. Lett. **10**(8), 453–459 (1991). doi:10.1016/0167-6377(91)90022-H
6. Choudhury, G.L., Whitt, W.: Computing distributions and moments in polling models by numerical transform inversion. Perform. Eval. **25**(4), 267–292 (1996). doi:10.1016/0166-5316(95)00015-1
7. Cohen, J.W.: The Single Server Queue. North-Holland, Amsterdam (1982)
8. Eisenberg, M.: Queues with periodic service and changeover time. Oper. Res. **20**(2), 440–451 (1972). doi:10.1287/opre.20.2.440
9. Federgruen, A., Katalan, Z.: The impact of adding a make-to-order item to a make-to-stock production system. Manag. Sci. **45**(7), 980–994 (1999)
10. Keilson, J., Servi, L.D.: A distributional form of Little's law. Oper. Res. Lett. **7**(5), 223–227 (1988). doi:10.1016/0167-6377(88)90035-1
11. Kleinrock, L., Levy, H.: The analysis of random polling systems. Oper. Res. **36**(5), 716–732 (1988)
12. Konheim, A.G., Levy, H., Srinivasan, M.M.: Descendant set: an efficient approach for the analysis of polling systems. IEEE Trans. Commun. **42**(234), 1245–1253 (1994). doi:10.1109/TCOMM.1994.580233
13. Levy, H., Sidi, M.: Polling systems: applications, modeling, and optimization. IEEE Trans. Commun. **38**(10), 1750–1760 (1990)
14. Levy, H., Sidi, M.: Polling systems with simultaneous arrivals. IEEE Trans. Commun. **39**(6), 823–827 (1991). doi:10.1109/26.87170
15. Little, J.D.C.: A proof of the queuing formula $L = \lambda W$. Oper. Res. **9**(3), 383–387 (1961)
16. Resing, J.A.C.: Polling systems and multitype branching processes. Queueing Syst. **13**(4), 409–426 (1993). doi:10.1007/BF01149263
17. Shiozawa, Y., Takine, T., Takahashi, Y., Hasegawa, T.: Analysis of a polling system with correlated input. Comput. Netw. ISDN Syst. **20**(1–5), 297–308 (1990). doi:10.1016/0169-7552(90)90038-T
18. Takagi, H.: Analysis of Polling Systems. MIT Press, London (1986)
19. Takagi, H.: Analysis and application of polling models. In: Haring, G., Lindemann, C., Reiser, M. (eds.) Performance Evaluation: Origins and Directions. Lecture Notes in Computer Science, vol. 1769, pp. 424–442. Springer, Berlin (2000)
20. Van der Mei, R.D.: Polling systems with simultaneous batch arrivals. Stoch. Models **17**(3), 271–292 (2001). doi:10.1081/STM-100002274
21. Van der Mei, R.D.: Waiting-time distributions in polling systems with simultaneous batch arrivals. Ann. Oper. Res. **113**(1–4), 155–173 (2002). doi:10.1023/A:1020918230560
22. Van der Mei, R.D., Hariharan, R., Reeser, P.K.: Web server performance modeling. Telecommun. Syst. **16**(3–4), 361–378 (2001). doi:10.1023/A:1016667027983
23. Winands, E.M.M., Adan, I.J.B.F., Van Houtum, G.J.: Mean value analysis for polling systems. Queueing Syst. **54**(1), 35–44 (2006). doi:10.1007/s11134-006-7898-8
24. Winands, E.M.M., Adan, I.J.B.F., Van Houtum, G.J.: The stochastic economic lot scheduling problem: a survey. Eur. J. Oper. Res. **210**(1), 1–9 (2011). doi:10.1016/j.ejor.2010.06.011
25. Wolff, R.W.: Poisson arrivals see time averages. Oper. Res. **30**(2), 223–231 (1982). doi:10.1287/opre.30.2.223