



Analysis of polling models with a self-ruling server

Jan-Kees van Ommeren¹  · Ahmad Al Hanbali² · Richard J. Boucherie¹

Received: 15 November 2018 / Revised: 18 October 2019 / Published online: 22 November 2019
© The Author(s) 2019

Abstract

Polling systems are systems consisting of multiple queues served by a single server. In this paper, we analyze polling systems with a server that is self-ruling, i.e., the server can decide to leave a queue, independent of the queue length and the number of served customers, or stay longer at a queue even if there is no customer waiting in the queue. The server decides during a service whether this is the last service of the visit and to leave the queue afterward, or it is a regular service followed, possibly, by other services. The characteristics of the last service may be different from the other services. For these polling systems, we derive a relation between the joint probability generating functions of the number of customers at the start of a server visit and, respectively, at the end of a server visit. We use these key relations to derive the joint probability generating function of the number of customers and the Laplace transform of the workload in the queues at an arbitrary time. Our analysis in this paper is a generalization of several models including the exponential time-limited model with preemptive-repeat-random service, the exponential time-limited model with non-preemptive service, the gated time-limited model, the Bernoulli time-limited model, the 1-limited discipline, the binomial gated discipline, and the binomial exhaustive discipline. Finally, we apply our results on an example of a new polling discipline, called the 1 + 1 self-ruling server, with Poisson batch arrivals. For this example, we compute numerically the expected sojourn time of an arbitrary customer in the queues.

Keywords Queueing · Polling systems · Limited visits

Mathematics Subject Classification 60K25 · 90B22

✉ Jan-Kees van Ommeren
j.c.w.vanommeren@utwente.nl

¹ Stochastic Operations Research, University of Twente, Postbox 217, 7500 AE Enschede, The Netherlands

² Systems Engineering, King Fahd University of Petroleum and Minerals, Dhahran, Kingdom of Saudi Arabia

1 Introduction

Polling systems are systems consisting of multiple queues served by shared servers. In recent years, polling models were used to model many real-life systems. For instance, traffic light systems, product-assembly systems, and wireless communication systems were modeled as polling systems. Good surveys on a broad class of polling models and their analysis can be found in, for example, [2,14–16].

In the analysis of polling systems, the standard method models the system at specific time points as a Markov chain and then relates the state space at these points; see [7]. The kernel relation within this method is the joint length of the queues in the system at the end of a server visit to queue i , denoted by $Q^{(i)}$, as function of the joint length of the queues at the start of the visit to $Q^{(i)}$. This relation can be written in the following general form:

$$\beta^{(i)}(\mathbf{z}) = \mathcal{F}_i(\alpha^{(i)})(\mathbf{z}), \quad (1)$$

where $\beta^{(i)}(\mathbf{z})$ is the joint probability generating function (p.g.f.) of the queue lengths at the end of a server visit to $Q^{(i)}$, $\alpha^{(i)}(\mathbf{z})$ is the joint p.g.f. of the queue lengths at the start of a server visit to $Q^{(i)}$, and \mathcal{F}_i is an operator representing the mapping between the queue lengths at these time points which depends on the assumed service discipline.

The next step in the analysis is to relate the joint p.g.f. of the queue lengths at the start of a server visit to $Q^{(i)}$ to the joint p.g.f. of the queue lengths at the end of a server visit to $Q^{(j)}$, $j = 1, \dots, M$, where M denotes the number of queues in the system, for example,

$$\alpha^{(i)}(\mathbf{z}) = \mathcal{G}_i(\beta^{(1)}, \dots, \beta^{(M)})(\mathbf{z}), \quad (2)$$

where \mathcal{G}_i is an operator representing the mapping between the queue lengths at the end of a visit to a queue and at the beginning of a visit to $Q^{(i)}$ which incorporates the effect of the switchover times and the routing of the server. We refer to [1,6] for the incorporation of Eqs. (1) and (2) into a numerical iterative framework to compute the joint queue length probabilities. However, in this paper, we do not consider the complete polling system and focus only on the relation in Eq. (1). We assume throughout the paper that the polling system is stable, i.e., we assume that all the processes under consideration have a proper limiting distribution. Here, we remark that stating and proving sufficient and/or necessary conditions for stability of the polling systems considered in this paper is a study on its own.

In this paper, we concentrate on $\mathcal{F}_i(\alpha^{(i)})(\cdot)$, which relates the joint p.g.f. of the queue lengths at the end of a server visit to $Q^{(i)}$, to the joint p.g.f. of the queue lengths at the start of the visit to this queue for queues with a self-ruling server. A self-ruling server can decide to leave a queue, independent of the queue length and the number of served customers, and can stay longer at a queue even if there is no customer waiting in the queue. During a service, the server decides with probability $p^{(i)}$ that this is the final service of the visit and to leave the queue after this service; otherwise, it is called a regular service which is possibly followed by other services during the same server visit. After service, customers can join another queue or leave the system; moreover,

new customers are added to the queues, independent of other customers. These new customers are called replacements. Regular customers are replaced, stochastically, in the same way. The replacement of the final customer can have a different distribution, for example, due to being interrupted in time-limited systems. This is the reason we assume that the server decides during the service whether it will be the final one. When the queue empties before the server decided for the final service, extra customers might be added to the queues. The server decides either to leave with probability $p_I^{(i)}$ or to stay and serve more customers during the same visit. We assume that $p^{(i)} + p_I^{(i)} > 0$, which implies that a server visit to $Q^{(i)}$ always ends. The distribution of the extra customers in the queues depends on the choice of the server to leave or stay after being idle.

We find the relation $\mathcal{F}_i(\alpha^{(i)})(\cdot)$ for service disciplines of the branching type when the server never decides that the ongoing service is the final one, i.e., $p^{(i)} = 0$. The so-called branching property, see [11, 12], plays an important role in the analysis of polling systems. Polling systems with server disciplines satisfying this property, such as exhaustive, gated or Bernoulli disciplines, can be analyzed exactly and have results in explicit form. The analysis of disciplines which do not satisfy the branching property, such as the time-limited and K -limited disciplines, is usually restricted to special cases, approximations, or numerical methods.

In the description of the self-ruling server, we did not focus on how the customers are generated. In the following, we describe the various models of generating customers in more detail. In the most general case, we assume that the server will serve the customers who are present at the beginning of the visit unless he decides to leave before serving them all. After the service of a customer, new customers are put in the queues, the so-called indirect replacements. We do not specify how these indirect replacements are generated, we only know the joint distribution of the number of indirect replacements at the queues. These indirect replacements have to wait for a new visit of the server before they are served. As a second model, we assume that if the queue becomes empty before the server decided to leave, a number of new customers are added to the queue, some of which might be served during the same visit, the so-called direct replacements; these potential customers to be served are treated as if they were there at the beginning of the visit. Again, we do not specify how these new customers are generated after the queue emptied; we only assume that we know the joint distribution of the number of new customers at every queue. In the next model, the so-called service-based discipline, we assume that, after the service of a customer, we do not only have indirect replacements but also direct replacements, which are customers who are directly served after the service ends during the same visit. Note that after the final service, there are no direct replacements. Also in this model, we do not specify how the new customers are generated, we only assume that the joint distribution of the number of direct and indirect replacements is known. As a final model for the replacements, we focus on the service-based discipline, where we now assume that customers arrive according to a batch Poisson process and that we know the service-time distributions. In this paper, we derive the relation between the joint p.g.f. of the number of customers in the queues at the start and the end of a visit to a queue. Based on these relations, we also find the joint Laplace transform of

the workload at an arbitrary time in the queues and the expected sojourn time of an arbitrary customer in the queues.

Many polling models found in the literature can be reformulated as an SRS polling model, for example, the 1-limited discipline (take $p^{(i)} = 1$), the binomial gated discipline (take the SRS discipline with $p^{(i)} = 0$ and no direct replacements), and the binomial exhaustive discipline (take the SRS discipline with $p^{(i)} = 0$ where the server leaves immediately without any replacements after becoming idle). The examples also extend to exponential time-limited models (where $p^{(i)}$ is the probability that the service of a customer is interrupted) such as the exponential time-limited model with preemptive-repeat-random service, the exponential time-limited model with non-preemptive service, the gated time-limited model, and the Bernoulli time-limited models.

The paper is structured as follows: In Sect. 2, we explain the general polling model with a self-ruling server. In Sect. 3, we analyze this general model and describe and analyze some more specified systems. Section 4 relates our work to the so-called exponentially time-limited polling models which were studied in [1,5,6,8] and their extensions in [9]. This latter is done in Sect. 5. In Sect. 6, we use the key relations derived in the previous sections to find expressions for the joint p.g.f. of the queue length and the joint Laplace transform of the workload at an arbitrary time in the queues. Finally, in Sect. 7, we apply our results in a numerical example of the 1+1 self-ruling discipline with Poisson batch arrivals. For this example, we compute numerically the expected queue length at various epochs and the expected sojourn time of an arbitrary customer.

2 Model and notation

Let us consider a polling system with a single server and $M \geq 1$ queues. The server visits the queues according to a routing schedule which we do not further specify, since our focus is on the relation in Eq. (1). During a visit to a queue, the server may decide to interrupt or prolong this visit independent of the queue length. More specifically, the server may decide, with probability $p^{(i)}$, during the service of a customer whether this is the last customer to be served during this visit. When the queue becomes empty before the server decides to leave, the server may serve some extra customers using the same discipline. Furthermore, we assume that the underlying service discipline, that is the service discipline when we assume that $p^{(i)} = 0$, is of branching type (see Sect. 2.1). We call this the self-ruling server (SRS) discipline.

We start this section with a short introduction to the branching-type service discipline. In the last part of this section, we discuss the SRS discipline in more details.

2.1 Branching-type service discipline

In this subsection, we consider the branching-type discipline. The standard definition of a branching-type service discipline is as follows (cf. [11,12]):

If there are $N_{si}^{(i)}$ customers present at $Q^{(i)}$ at the start of a visit, then during the course of the visit, each of these $N_{si}^{(i)}$ customers will be replaced in an i.i.d. manner by a random population.

Denote the number of indirect replacements for a customer in the queues by $\mathbf{R}^{(i)} = (R_1^{(i)}, \dots, R_M^{(i)})$, which has an M -dimensional p.g.f. $H^{(i)}(\mathbf{z}) := \mathbb{E}[\mathbf{z}^{\mathbf{R}^{(i)}}]$, where $\mathbf{z} = (z_1, \dots, z_M)$ with $|z_j| \leq 1$. Denote by $\mathbf{N}_s^{(i)} = (N_{s1}^{(i)}, \dots, N_{sM}^{(i)})$ the number of customers present in the queues at the start of a visit to $Q^{(i)}$ and define its p.g.f. $\alpha^{(i)}(\mathbf{z}) := \mathbb{E}[\mathbf{z}^{\mathbf{N}_s^{(i)}}]$. The number of customers in the queues after a server visit to $Q^{(i)}$ is given by $\mathbf{N}_e^{(i)} = \mathbf{N}_{s0}^{(i)} + \sum_{k=1}^{N_{si}^{(i)}} \mathbf{R}_k^{(i)}$, where $\mathbf{N}_{s0}^{(i)} = (N_{s1}^{(i)}, \dots, N_{si}^{(i)} = 0, \dots, N_{sM}^{(i)})$. Then, the joint p.g.f. of $\mathbf{N}_e^{(i)}$ is of the simple form

$$\beta^{(i)}(\mathbf{z}) := \mathbb{E}[\mathbf{z}^{\mathbf{N}_e^{(i)}}] = \alpha^{(i)}(\mathbf{z}_i^*),$$

(see Eq. (1)), where $\mathbf{z}_i^* := (z_1, \dots, z_{i-1}, H^{(i)}(\mathbf{z}), z_{i+1}, \dots, z_M)$. In Sect. 3, we first concentrate on this general form of branching where $H^{(i)}(\mathbf{z})$ is not specified.

2.2 The self-ruling server

Our focus is on the behavior of queues with the self-ruling server discipline. According to this SRS discipline, the server decides during a service, with probability $p^{(i)}$, whether this is the final customer to be served during this visit to. The customers that are served before the final customer are called regular customers. Let $N_V^{(i)}$ denotes the number of regular customers that are served during the server visit. In a queue with the SRS discipline, at most a geometrically distributed number of customers, with mean $1/p^{(i)}$, is served. After (or during) the service of a customer, the length of all the queues can increase by a random amount. The extra customers are called the indirect replacement of a customer. The indirect replacements for regular customers are stochastically equivalent; the indirect replacement of a final customer may have a different distribution. The indirect replacements of all the customers are independent. Furthermore, the indirect replacements at $Q^{(i)}$ are not served during the ongoing server visit and have to wait for a new server visit. Let $\mathbb{1}_E$ denotes the indicator function of an event E and let F denote the event that the customer being served is the final customer. Define the p.g.f. of \mathbf{R} on the event it is a final customer by

$$H_-^{(i)}(\mathbf{z}) := \mathbb{E}[\mathbf{z}^{\mathbf{R}^{(i)}} \mathbb{1}_F], \tag{3}$$

the p.g.f. of \mathbf{R} on the event it is a regular customer by

$$H_+^{(i)}(\mathbf{z}) := \mathbb{E}[\mathbf{z}^{\mathbf{R}^{(i)}} (1 - \mathbb{1}_F)],$$

and the p.g.f. of \mathbf{R} by

$$H^{(i)}(\mathbf{z}) := H_-^{(i)}(\mathbf{z}) + H_+^{(i)}(\mathbf{z}).$$

Note that $H_-^{(i)}(\mathbf{z})$ and $H_+^{(i)}(\mathbf{z})$ are not conditional p.g.f.'s and that $p^{(i)} = H_-^{(i)}(\mathbf{1})$, where $\mathbf{1} := (1, \dots, 1)$.

When the server becomes idle before serving the final customer, the server will serve $S_X^{(i)}$ extra customers during the same visit as if they were present at the start of the visit and $\mathbf{R}_X^{(i)}$ additional customers are added to all the queues. If no extra customers are served, i.e., $S_X^{(i)} = 0$, the server leaves $Q^{(i)}$. The joint p.g.f. of the extra customers is denoted by $H_X^{(i)}(z, \mathbf{z}) := \mathbb{E} \left[z^{S_X^{(i)}} \mathbf{z}^{\mathbf{R}_X^{(i)}} \right]$. Define

$$H_{X-}^{(i)}(\mathbf{z}) := H_X^{(i)}(0, \mathbf{z}) \text{ and } H_{X+}^{(i)}(z, \mathbf{z}) := H_X^{(i)}(z, \mathbf{z}) - H_{X-}^{(i)}(\mathbf{z}), \quad (4)$$

the joint p.g.f. of $\mathbf{R}_X^{(i)}$, $S_X^{(i)}$ on the events, respectively, that the server leaves $Q^{(i)}$ and the server starts another service.

For convenience, we introduce the exhaustive self-ruling server (E-SRS) discipline. This E-SRS discipline is similar to the SRS discipline but the server will immediately end a visit as soon as the server becomes idle. So, in this E-SRS discipline, $H_X^{(i)}(z, \mathbf{z}) = 1$, which states that there are neither extra customers to be served nor additional customers at the queues after the server becomes idle.

In the following section, we analyze polling systems with the self-ruling server discipline. Note that we will not specify how exactly the indirect replacement of a customer is implemented except they are independent. A specific form of indirect replacements is considered in Sect. 3.3.

3 Analysis of polling systems under the SRS discipline

In this section, we focus on Eq. (1), which relates the joint p.g.f. of the queue lengths at the end of a visit to the joint p.g.f. of the queue lengths at the start of a visit. We present results for the self-ruling server discipline with the general branching-type discipline in Sect. 3.1. In Sect. 3.2, we specify the replacement process of a customer by considering the so-called service-based branching-type service discipline. In this discipline, a customer does not only have an indirect replacement population $\mathbf{R}_X^{(i)}$ but also a direct replacement by customers that may be served during the ongoing server visit. As an example of such a system, consider a task at a certain queue. During the service of a task, a new task may be generated, either at the same queue or at other queues. Some of the tasks, generated at the queue where the server is, have to be done during the same visit; others can wait for a next visit. In the previous models, we did not focus on how replacements are generated. In Sect. 3.3, a further specification of this generating process is given where replacements arrive during the service of a customer according to a batch Poisson process.

3.1 General branching-type discipline

In this part, we focus on the general form of branching where we assume that the replacement p.g.f.'s $H_+^{(i)}(\mathbf{z})$ and $H_-^{(i)}(\mathbf{z})$ are given.

Lemma 1 *For a polling system operating under an exhaustive self-ruling single server discipline, the relation between the joint p.g.f. of the queue length at the start and the end of a server visit to $Q^{(i)}$ reads*

$$\beta^{(i)}(\mathbf{z}) = \gamma_-^{(i)}(\mathbf{z})(\alpha^{(i)}(\mathbf{z}) - \alpha^{(i)}(\mathbf{z}_\star^{(i)})) + \alpha^{(i)}(\mathbf{z}_\star^{(i)}), \tag{5}$$

where $\mathbf{z}_\star^{(i)} := (z_1, \dots, z_{i-1}, H_+^{(i)}(\mathbf{z}), z_{i+1}, \dots, z_M)$, and

$$\gamma_-^{(i)}(\mathbf{z}) := \frac{H_-^{(i)}(\mathbf{z})}{z_i - H_+^{(i)}(\mathbf{z})}. \tag{6}$$

Proof Consider a server visit to $Q^{(i)}$. Denote the total indirect replacement at the end of the server visit by $\mathbf{R}_T^{(i)}$; the T is added here to indicate that it is the total indirect replacement for all customers that are present at the start of the visit, as opposed to the indirect replacement of a single customer in Eq. (3). It is easily seen that

$$\mathbf{R}_T^{(i)} = \sum_{k=1}^{N_V^{(i)}} \mathbf{R}_k^{(i)} + \mathbb{1}_{\{N_V^{(i)} < N_{si}^{(i)}\}} \mathbf{R}_-^{(i)}$$

where $\mathbf{R}_-^{(i)}$ denotes the indirect replacement of the final customer. Therefore

$$\begin{aligned} \mathbb{E}[\mathbf{z}^{\mathbf{R}_T} | N_{si}^{(i)} = n] &= \sum_{k=0}^n \mathbb{E}[\mathbf{z}^{\mathbf{R}_T} \mathbb{1}_{\{N_V^{(i)}=k\}} | N_{si}^{(i)} = n] \\ &= \sum_{k=0}^{n-1} (H_+^{(i)}(\mathbf{z}))^k H_-^{(i)}(\mathbf{z}) z_i^{n-k-1} + (H_+^{(i)}(\mathbf{z}))^n \\ &= H_-^{(i)}(\mathbf{z}) \frac{z_i^n - (H_+^{(i)}(\mathbf{z}))^n}{z_i - H_+^{(i)}(\mathbf{z})} + (H_+^{(i)}(\mathbf{z}))^n, \end{aligned}$$

where $N_{si}^{(i)}$ is the queue length of $Q^{(i)}$ at the start of a server visit to this queue. By unconditioning on $\mathbf{N}_s^{(i)}$, we find Eq. (5). □

In contrast with the E-SRS discipline, under the general SRS discipline the server may still serve extra customers at $Q^{(i)}$ even after the server becomes idle.

Theorem 1 *For a polling system operating under a self-ruling server discipline, the relation between the joint p.g.f. of the queue lengths at the start and the end of a server visit to $Q^{(i)}$ reads*

$$\beta^{(i)}(\mathbf{z}) = \gamma_-^{(i)}(\mathbf{z})(\alpha^{(i)}(\mathbf{z}) - \alpha^{(i)}(\mathbf{z}_\star^{(i)})) + \gamma_+^{(i)}(\mathbf{z})\alpha^{(i)}(\mathbf{z}_\star^{(i)}),$$

where $\mathbf{z}_\star^{(i)} := (z_1, \dots, z_{i-1}, H_+^{(i)}(\mathbf{z}), z_{i+1}, \dots, z_M)$,

$$\gamma_-^{(i)}(\mathbf{z}) = \frac{H_-^{(i)}(\mathbf{z})}{z_i - H_+^{(i)}(\mathbf{z})},$$

and

$$\gamma_+^{(i)}(\mathbf{z}) := \frac{H_{X_-}^{(i)}(\mathbf{z}) + \gamma_-^{(i)}(\mathbf{z}) \left(H_{X_+}^{(i)}(z_i, \mathbf{z}) - H_{X_+}^{(i)}(H_+^{(i)}(\mathbf{z}), \mathbf{z}) \right)}{1 - H_{X_+}^{(i)}(H_+^{(i)}(\mathbf{z}), \mathbf{z})}. \quad (7)$$

Proof Consider a visit of the server to $Q^{(i)}$ with initially $N_{si}^{(i)}$ customers. During this visit, it may occur that (i) the server decides that one of the $N_{si}^{(i)}$ customers is the final customer or (ii) the server serves all $N_{si}^{(i)}$ customers as regular customers. In case (i), it is readily seen that the indirect replacement process, that is $\mathbf{R}_T^{(i)}$, the number of additional customers in the queues, is identical for both the E-SRS and the SRS discipline. However, in case (ii), the indirect replacement process is different for each discipline. Under the E-SRS discipline, the server immediately leaves when the queue becomes empty, say this occurs at time t_0 . Under the general SRS discipline, at time t_0 the server may remain at the queue and a sequence of idle and busy periods will follow until eventually the server decides on a final customer or decides to serve no extra customers after an idle time ($S_X^{(i)} = 0$). This latter contribution (after t_0) to the indirect replacement process is represented in the term $\gamma_+^{(i)}(\mathbf{z})$.

Observe that an idle server, will leave a) without serving extra customers, b) during the busy period serving the extra customers, or c) after this subsequent busy period. This process is regenerative in the sense that if the server does not leave before the end of the first busy period following the idle period, then the process starts anew at that specific time instant. Then, we have the following relation for $\gamma_+^{(i)}(\mathbf{z})$:

$$\begin{aligned} \gamma_+^{(i)}(\mathbf{z}) &= H_{X_-}^{(i)}(\mathbf{z}) + \gamma_-^{(i)}(\mathbf{z}) \left(H_{X_+}^{(i)}(z_i, \mathbf{z}) - H_{X_+}^{(i)}(H_+^{(i)}(\mathbf{z}), \mathbf{z}) \right) \\ &\quad + H_{X_+}^{(i)}(H_+^{(i)}(\mathbf{z}), \mathbf{z})\gamma_+^{(i)}(\mathbf{z}), \end{aligned}$$

where $H_{X_-}^{(i)}(\mathbf{z})$ is the p.g.f. of the indirect replacements on the event that the server leaves; see Eq. (4). The other terms on the RHS are similar to Eq. (5), with the joint p.g.f. at the start of the busy period given by $H_{X_+}^{(i)}(z, \mathbf{z})$ instead of $\alpha_i(\mathbf{z})$. This leads to Eq. (7). \square

3.2 Service-based branching-type discipline

A special subclass of the general branching-type service discipline is the service-based branching-type discipline, which can be described as follows: At the start of a service period at $Q^{(i)}$, there are $N_{si}^{(i)}$ customers at this queue; these customers are

called zeroth-generation customers. With probability $q_1^{(i)}$, such a customer goes to the server, independent of the other customers, to receive service and becomes a first-generation customer; otherwise, this customer waits for the next visit of the server. We assume that every customer of the first generation, immediately after being served, is replaced in a stochastically identical way and independently of the other customers. The replacing customers are split into two parts, namely direct replacements, which are customers that are served at $Q^{(i)}$ in the same visit of the server, and indirect replacements, additional customers at all queues, including at $Q^{(i)}$, which are served in the subsequent visits. The direct replacements of the first-generation customers are called second-generation customers. After being served, a second-generation customer is replaced, in an i.i.d. fashion, by third-generation customers that are served at $Q^{(i)}$ during the same visit of the server and additional customers to all the queues, including $Q^{(i)}$ to be served in subsequent visits. We can continue this construction for further generations. All customers linked to the same first-generation customer are called a family. Denote the number of additional customers that arrived during the service of an n th generation customer at $Q^{(i)}$ and are served during the ongoing server visit by $S_n^{(i)}$ and denote the additional customers that are not served during this visit by $\mathbf{R}_n^{(i)} = (R_{n1}^{(i)}, \dots, R_{nM}^{(i)})$. Note that the order of service of customers does not influence the distribution of $\mathbf{R}^{(i)} = (R_1^{(i)}, \dots, R_M^{(i)})$, the total indirect replacement during the visit to $Q^{(i)}$. Furthermore, remark the difference between the total indirect replacement $\mathbf{R}^{(i)}$ of a family and the indirect replacement of an n th generation customer $\mathbf{R}_n^{(i)}$.

In the case of this service-based branching-type discipline together with a self-ruling server, we assume that regular customers and the final customers may have different joint distributions of the number of direct replacements and the indirect replacement population. We also make the assumption that, whether we consider a final or regular customer, the total of new customers at the queues per service, that is the sum of the indirect replacement and direct replacements, has the same distribution for every generation. By this assumption, the order in which customers are handled does not matter for the indirect replacement population when the server leaves, which can be seen as follows: Suppose that customers are served in one specific order. Now assume that the server leaves without serving a final customer. Then, in any generation, for all orders, the same number of customers have been fully served, so the indirect replacement population is the same. Next assume that the server leaves after a final customer; in this case, in all orders, the same number of customers will be served. The indirect replacement population added to the population in the queues at the start of the visit consists of all customers that were present at the start, plus new customers after the regular services and the final service, minus the number of served customers. By assumption, the total number of new customers does not depend on the generation number of the served customers, so also in this case, the indirect replacement population does not depend on the order of the customers.

In the case where we have a service-based branching-type discipline, we can further specify $H_-^{(i)}(\mathbf{z})$ and $H_+^{(i)}(\mathbf{z})$. We can then use these specified p.g.f.'s to modify Th. 1. As remarked before, a customer present at the start of a visit is replaced by a set of customers at every queue in the system. Let us introduce the following joint p.g.f.'s:

$$H_{n-}^{(i)}(\mathbf{z}) := \mathbb{E}\left[\mathbf{z}^{\mathbf{R}^{(i)}} \mathbb{1}_F\right] \text{ and } H_{n+}^{(i)}(z, \mathbf{z}) := \mathbb{E}\left[z^{S_n^{(i)}} \mathbf{z}^{\mathbf{R}_n^{(i)}} (1 - \mathbb{1}_F)\right],$$

where we used that after a final customer, no other customers are served at $Q^{(i)}$ during the same visit. Note that by the assumption that, per service, the total number of new customers per queue does not depend on the generation n , both $H_{*-}^{(i)} := H_{n-}^{(i)}(\mathbf{z})$ and $H_{*+}^{(i)} := H_{n+}^{(i)}(z_i, \mathbf{z})$ do not depend on the generation n . Also by this assumption, the order in which we serve the customers is not important for the indirect replacement population and we can assume that we serve a first-generation customer and its complete family consecutively. Within a family, we handle the customers generation by generation.

Let $G_{n+}^{(i)}(z, \mathbf{z})$ denote the joint p.g.f. of the direct and indirect replacements after serving the n th generation on the event that the visit still continues in the next generation. By the assumption that a zeroth-generation customer visits the server with probability $q_1^{(i)}$, we have that $G_{0+}^{(i)}(z, \mathbf{z}) = q_1^{(i)}z + (1 - q_1^{(i)})z_i$. For $n = 1, 2, \dots$, we see that

$$G_{n+}^{(i)}(z, \mathbf{z}) = G_{(n-1)+}^{(i)}(H_{n+}^{(i)}(z, \mathbf{z}), \mathbf{z}),$$

where we use that the direct replacements of the n th generation are served as the $(n + 1)$ th generation of a family.

Let $G_{n-}^{(i)}(z, \mathbf{z})$ denote the p.g.f. of the total population at the queues that will be served in later visits on the event that the visit to $Q^{(i)}$ was interrupted during the service of an n th generation customer. We then find

$$G_{n-}^{(i)}(\mathbf{z}) = \frac{H_{*-}^{(i)}(\mathbf{z})}{z_i - H_{*+}^{(i)}(\mathbf{z})} \left(G_{(n-1)+}^{(i)}(z_i, \mathbf{z}) - G_{n-1+}^{(i)}(H_{*+}^{(i)}(\mathbf{z}), \mathbf{z}) \right). \tag{8}$$

Since the system is stable, $\lim_{n \rightarrow \infty} \mathbb{P}(S_n^{(i)} = 0) = 1$ and we get the joint p.g.f. of $\mathbf{R}^{(i)}$ for a customer with an offspring that is fully served (cf. Eq. (3)):

$$H_+^{(i)}(\mathbf{z}) := \mathbb{E}\left[\mathbf{z}^{\mathbf{R}} (1 - \mathbb{1}_F)\right] = \lim_{n \rightarrow \infty} G_{n+}^{(i)}(z, \mathbf{z}). \tag{9}$$

Note that the limit on the RHS of Eq. (9) does not depend on z .

Because an interrupted family is interrupted in exactly one generation, and $G_{0+}^{(i)}(z, \mathbf{z}) = q_1^{(i)}z + (1 - q_1^{(i)})z_i$, we find

$$\begin{aligned} H_-^{(i)}(\mathbf{z}) &= \sum_{n=1}^{\infty} G_{n-}^{(i)}(\mathbf{z}) \\ &= \sum_{n=1}^{\infty} \frac{H_{*-}^{(i)}(\mathbf{z})}{z_i - H_{*+}^{(i)}(\mathbf{z})} \left(G_{(n-1)+}^{(i)}(z_i, \mathbf{z}) - G_{n-1+}^{(i)}(H_{*+}^{(i)}(\mathbf{z}), \mathbf{z}) \right) \end{aligned}$$

$$= \frac{H_{*-}^{(i)}(\mathbf{z})}{z_i - H_{*+}^{(i)}(\mathbf{z})} \left(z_i - H_{+}^{(i)}(\mathbf{z}) \right), \tag{10}$$

where we used Eq. (9) and the assumption that both $H_{*-}^{(i)}(\mathbf{z})$ and $H_{*+}^{(i)}(\mathbf{z})$ do not depend on the generation.

Theorem 2 *For a polling system operating under a service-based self-ruling server discipline where $H_{n-}^{(i)}(\mathbf{z})$ and $H_{n+}^{(i)}(z_i, \mathbf{z})$ do not depend on the generation, the relation between the joint p.g.f. of the queue lengths at the start and the end of a server visit to $Q^{(i)}$ reads*

$$\beta^{(i)}(\mathbf{z}) = \gamma_-^{(i)}(\mathbf{z})(\alpha^{(i)}(\mathbf{z}) - \alpha^{(i)}(\mathbf{z}_\star^{(i)})) + \gamma_+^{(i)}(\mathbf{z})\alpha^{(i)}(\mathbf{z}_\star^{(i)}),$$

where $\mathbf{z}_\star^{(i)} = (z_1, \dots, z_{i-1}, H_{+}^{(i)}(\mathbf{z}), z_{i+1}, \dots, z_M)$,

$$\gamma_-^{(i)}(\mathbf{z}) := \frac{H_{*-}^{(i)}(\mathbf{z})}{z_i - H_{*+}^{(i)}(\mathbf{z})} \tag{11}$$

and

$$\gamma_+^{(i)}(\mathbf{z}) := \frac{H_{X-}^{(i)}(\mathbf{z}) + \gamma_-^{(i)}(\mathbf{z}) \left(H_{X+}^{(i)}(z_i, \mathbf{z}) - H_{X+}^{(i)}(H_{+}^{(i)}(\mathbf{z}), \mathbf{z}) \right)}{1 - H_{X+}^{(i)}(H_{+}^{(i)}(\mathbf{z}), \mathbf{z})}.$$

Proof Using Eq. (10), we can directly rewrite Eq. (6) to Eq. (11). Now apply Th. 1 to prove this theorem. □

In Th. 2, we have the unknown function $H_{+}^{(i)}(\mathbf{z})$ as defined in Eq. (9). In the remainder of this section, we present some examples where we actually can specify $H_{+}^{(i)}(\mathbf{z})$.

Example 1 In this example, we introduce a special branching-type discipline, the 1+1 service discipline. In this discipline, all first-generation customers are served during a visit of the server. From all the customers which arrive during the service time of the same first-generation customer, at most one is served during the same visit. All other second- or third-generation customers, if any, have to wait until the next server visit. By considering different scenarios, namely whether the first-generation customer was a final customer or not and, in the latter case, whether there was no arrival or at least one during its service time, we find

$$H_{+}^{(i)}(\mathbf{z}) = H_{1+}^{(i)}(0, \mathbf{z}) + \frac{H_{*+}^{(i)}(\mathbf{z}) - H_{1+}^{(i)}(0, \mathbf{z})}{z_i} H_{*+}^{(i)}(\mathbf{z}) \tag{12}$$

and

$$H_{-}^{(i)}(\mathbf{z}) = H_{*-}^{(i)}(\mathbf{z}) + \frac{H_{*+}^{(i)}(\mathbf{z}) - H_{1+}^{(i)}(0, \mathbf{z})}{z_i} H_{*-}^{(i)}(\mathbf{z}).$$

It is easily verified that this expression for $H_-^{(i)}(\mathbf{z})$ also can be derived from Eqs. (10) and (12) by observing that

$$z_i - H_+^{(i)}(\mathbf{z}) = \frac{H_{1+}^{(i)}(0, \mathbf{z})(H_{*+}^{(i)}(\mathbf{z}) - z_i)}{z_i} + \frac{z_i^2 - (H_{*+}^{(i)}(\mathbf{z}))^2}{z_i}.$$

Example 2 Assume that, for all $|z| \leq 1$, the function $H_{n+}^{(i)}(z, \mathbf{z})$ does not depend on the generation number n . Then, by taking a different order of serving the customers, namely by serving its direct replacements directly after a customer, we find that a customer of the second generation has, stochastically, the same indirect replacement as a customer of the first generation. We then see that $H_+^{(i)}(\mathbf{z})$, see Eq. (9), also satisfies $H_+^{(i)}(\mathbf{z}) = H_{1+}^{(i)}(H_+^{(i)}(\mathbf{z}), \mathbf{z})$. Since we have a stable system $\frac{\partial}{\partial z} H_{1+}^{(i)}(z, \mathbf{z})|_{z=1} < 1$, and we can prove, by using Rouché’s theorem, that this equation for $H_+^{(i)}(\mathbf{z})$ has exactly one solution for all \mathbf{z} with $z_j \leq 1$ for $j = 1, \dots, M$.

3.3 Poisson arrivals

We will go one step further in specifying the replacement p.g.f. $H_+^{(i)}(z, \mathbf{z})$, by assuming that customers arrive to the system according to a batch Poisson processes with rate λ , where the batch may split over several queues, and that the service times of individual customers follow a general distribution. In this system, a customer is replaced by customers that arrive during its service. For a final customer, the service-time distribution may be different for the other customers. For the final customer, the Laplace–Stieltjes transform (LST) of the service time at $Q^{(i)}$ is denoted by $\tilde{T}_{S-}^{(i)}(s)$ and for the other customers by $\tilde{T}_{S+}^{(i)}(s)$.

We assume a Bernoulli discipline where a customer of the n th generation, if any, visits the server to be served with probability $q_n^{(i)}$ independent of other customers. A customer that does not visit the server can not be the final customer in the ongoing server visit. These customers are their own indirect replacement at $Q^{(i)}$ and will be served in a future server visit. Note that if $q_n^{(i)} = 0$ then, effectively, there is no n th generation. The customer that is served at $Q^{(i)}$ leaves the system or joins $Q^{(j)}$; see, for example, [13]. For a final customer, the routing probabilities may differ from the other customers and are denoted by $r_{j-}^{(i)}$. For the other customers, the routing probabilities are denoted by $r_{j+}^{(i)}$. The p.g.f.’s of the routing probabilities are

$$r_-^{(i)}(\mathbf{z}) := r_{0-}^{(i)} + \sum_{j=1}^M r_{j-}^{(i)} z_j \quad \text{and}$$

$$r_+^{(i)}(\mathbf{z}) := r_{0+}^{(i)} + \sum_{j=1}^M r_{j+}^{(i)} z_j.$$

As before, we assume that the service processes, the arrival process, the batch sizes, and the routing of customers are independent.

Denote the numbers of simultaneously arriving customers at $Q^{(j)}$ by $A^{(j)}$ for $j = 1, \dots, M$, and their joint p.g.f. by $\widehat{A}(\mathbf{z}) := \mathbb{E}[\mathbf{z}^{\mathbf{A}}]$, where $\mathbf{A} = (A^{(1)}, \dots, A^{(M)})$. Denote the number of customers that arrived to $Q^{(j)}$ during the service of a customer at $Q^{(i)}$ by $A_{Sj}^{(i)}$. Then the indirect replacement population is $R_{nj}^{(i)} = A_{Sj}^{(i)}$ with $j = 1, \dots, M$ and $j \neq i$ and $R_{ni}^{(i)} = A_{Si}^{(i)} - S_n^{(i)}$, where the number of direct replacements $S_n^{(i)}$ has, given $A_{Sj}^{(i)}$, a binomial distribution with parameters $A_{Sj}^{(i)}$ and $q_{n+1}^{(i)}$ by the assumption of the Bernoulli discipline. By conditioning on the service times, we can calculate that (cf. Eqs. (9) and (10))

$$H_{n+}^{(i)}(z, \mathbf{z}) = (1 - p^{(i)})\widetilde{T}_{S+}^{(i)}\left(\lambda\left(1 - \widehat{A}_n^{(i)}(z, \mathbf{z})\right)\right)r_+^{(i)}(\mathbf{z}),$$

and

$$H_{*-}^{(i)}(\mathbf{z}) = p^{(i)}\widetilde{T}_{S-}^{(i)}(\Lambda(\mathbf{z}))r_-^{(i)}(\mathbf{z}),$$

where

$$\widehat{A}_n^{(i)}(z, \mathbf{z}) := \widehat{A}\left(z_1, \dots, z_{i-1}, q_{n+1}^{(i)}z + (1 - q_{n+1}^{(i)})z_i, z_{i+1}, \dots, z_M\right). \tag{13}$$

In the following, we use several times that $\widehat{A}(\mathbf{z}) = \widehat{A}_n^{(i)}(z_i, \mathbf{z})$. For convenience, we also define $\Lambda(\mathbf{z}) := \lambda(1 - \widehat{A}(\mathbf{z}))$.

Under the SRS discipline, the server will always move to a next queue when the final customer is served. However, when the queue becomes empty and the last customer was not the final one, the server remains idle at the queue for some random time. If, during this idle time, no new customers arrive at $Q^{(i)}$, the server will move to another queue. Otherwise, the server will immediately start serving again under the same SRS discipline; if the server started working at the same queue and becomes idle for a second time, the server will behave, stochastically, the same as the first time the server became idle. Note that by these assumptions, Eq. (4) does not hold; if a batch arrives at $Q^{(i)}$ and all its customers are not served, which occurs with probability $\widehat{A}_1^{(i)}(0, \mathbf{1})$, then the server does not leave but will stay idle at $Q^{(i)}$ for another period.

Now, we can also specify $H_{X+}^{(i)}(z, \mathbf{z})$ and $H_{X-}^{(i)}(\mathbf{z})$ (see Eq. (4)), and the joint generating functions for the direct and indirect replacements after a period in which the server is idle, where the LST of the idle time is denoted by $\widetilde{W}_I^{(i)}(s)$. We get

$$H_{X-}^{(i)}(\mathbf{z}) = \widetilde{W}_I^{(i)}\left(\lambda\left(1 - \widehat{A}_1^{(i)}\left(0, \mathbf{z}_0^{(i)}\right)\right)\right)$$

and

$$H_{X+}^{(i)}(z, \mathbf{z}) = \left(1 - \widetilde{W}_I^{(i)}\left(\lambda\left(1 - \widehat{A}_1^{(i)}\left(0, \mathbf{z}_0^{(i)}\right)\right)\right)\right) \frac{\widehat{A}_1^{(i)}(z, \mathbf{z}) - \widehat{A}_1^{(i)}\left(0, \mathbf{z}_0^{(i)}\right)}{1 - \widehat{A}_1^{(i)}\left(0, \mathbf{z}_0^{(i)}\right)},$$

where $\mathbf{z}_0^{(i)} := (z_1, \dots, z_{i-1}, 0, z_{i+1}, \dots, z_M)$. Combining the above observations with Th. 2 leads to the following theorem.

Theorem 3 *For a polling system with batch Poisson arrivals operating under a service-based self-ruling server discipline, the relation between the joint p.g.f. of the queue lengths at the start and the end of a server visit to $Q^{(i)}$ reads*

$$\beta^{(i)}(\mathbf{z}) = \gamma_-^{(i)}(\mathbf{z})(\alpha^{(i)}(\mathbf{z}) - \alpha^{(i)}(\mathbf{z}_\star^{(i)})) + \gamma_+^{(i)}(\mathbf{z})\alpha^{(i)}(\mathbf{z}_\star^{(i)}),$$

where $\mathbf{z}_\star^{(i)} = (z_1, \dots, z_{i-1}, H_+^{(i)}(\mathbf{z}), z_{i+1}, \dots, z_M)$,

$$\gamma_-^{(i)}(\mathbf{z}) = \frac{p^{(i)}\tilde{T}_{S_-}^{(i)}(\Lambda(\mathbf{z}))r_-^{(i)}(\mathbf{z})}{z_i - (1 - p^{(i)})\tilde{T}_{S_+}^{(i)}(\Lambda(\mathbf{z}))r_+^{(i)}(\mathbf{z})}$$

and

$$\gamma_+^{(i)}(\mathbf{z}) = \frac{\tilde{W}_I^{(i)}\left(\lambda\left(1 - \widehat{A}_1^{(i)}\left(0, \mathbf{z}_0^{(i)}\right)\right)\right)\left(1 - \widehat{A}_1^{(i)}\left(0, \mathbf{z}_0^{(i)}\right)\right)}{D(z, \mathbf{z})} + \frac{\gamma_-^{(i)}(\mathbf{z})\left(1 - \tilde{W}_I^{(i)}\left(\lambda\left(1 - \widehat{A}_1^{(i)}\left(0, \mathbf{z}_0^{(i)}\right)\right)\right)\right)\left(\widehat{A}(\mathbf{z}) - \widehat{A}_1^{(i)}\left(H_+^{(i)}(\mathbf{z}), \mathbf{z}\right)\right)}{D(z, \mathbf{z})},$$

with

$$D(z, \mathbf{z}) := 1 - \widehat{A}_1^{(i)}\left(H_+^{(i)}(\mathbf{z}), \mathbf{z}\right) + \tilde{W}_I^{(i)}\left(\lambda\left(1 - \widehat{A}_1^{(i)}\left(0, \mathbf{z}_0^{(i)}\right)\right)\right)\left(\widehat{A}_1^{(i)}\left(H_+^{(i)}(\mathbf{z}), \mathbf{z}\right) - \widehat{A}_1^{(i)}\left(0, \mathbf{z}_0^{(i)}\right)\right).$$

Remark 1 If the idle time has an exponential distribution with rate $\xi^{(i)}$, that is $\tilde{W}_I^{(i)}(s) = \xi^{(i)} / (\xi^{(i)} + s)$, we get

$$\gamma_+^{(i)}(\mathbf{z}) = \frac{\xi^{(i)} + \lambda\gamma_-^{(i)}(\mathbf{z})\left(\widehat{A}(\mathbf{z}) - \widehat{A}_1^{(i)}\left(H_+^{(i)}(\mathbf{z}), \mathbf{z}\right)\right)}{\xi^{(i)} + \lambda\left(1 - \widehat{A}_1^{(i)}\left(H_+^{(i)}(\mathbf{z}), \mathbf{z}\right)\right)}.$$

Remark 2 If we use the definition in Eq. (4), that is, the server will either leave after the extra time or (really) start a new service, we would have to ignore the arrivals of the not served batches at $Q^{(i)}$ since they do not end the extra time. This gives

$$H_{X_-}^{(i)}(\mathbf{z}) := \tilde{W}_I^{(i)}\left(\lambda\left(1 - \widehat{A}_1^{(i)}(0, \mathbf{z})\right)\right)$$

and

$$H_{X_+}^{(i)}(z, \mathbf{z}) = \left(1 - \tilde{W}_I^{(i)}\left(\lambda\left(1 - \widehat{A}_1^{(i)}(0, \mathbf{z})\right)\right)\right) \frac{\widehat{A}_1^{(i)}(z, \mathbf{z}) - \widehat{A}_1^{(i)}(0, \mathbf{z})}{1 - \widehat{A}_1^{(i)}(0, \mathbf{z})}.$$

Remark 3 In many papers on polling systems, it is assumed that customers arrive to the system according to M independent Poisson processes, to $Q^{(j)}$ with rate λ_j , $j = 1, \dots, M$. The generating function of the batch size at $Q^{(j)}$ is denoted by $\widehat{B}^{(j)}(z) := \mathbb{E}\left[z^{B^{(j)}}\right]$. In our framework, we get $\lambda = \sum_{j=1}^M \lambda_j$ and

$$\widehat{A}(\mathbf{z}) = \sum_{j=1}^M \frac{\lambda_j}{\lambda} \widehat{B}^{(j)}(z_j).$$

4 Exponential time-limited polling systems

In this section, we study exponential time-limited polling systems with Poisson arrivals where the total visit time of a server to a queue is at most an exponentially distributed time. If the timer expires during a service, the ongoing service is interrupted. We consider two types of time-limited disciplines: (a) the pure time-limited case (P-TL) in which the server visits a queue for an exponentially distributed time and (b) the exhaustive time-limited discipline (E-TL) in which the server visits a queue for at most an exponentially distributed time, but also leaves when the queue becomes empty. In the case of the P-TL, customers that arrive at $Q^{(i)}$ at the end of an idle period are served as if they were present at the start of the visit. We model these time-limited systems in the framework of the customer-limited system. To do so, we first specify $p^{(i)}$ and $\widetilde{T}_{S+}^{(i)}(\cdot)$ since these quantities do not depend on the effect of the interruption or on the underlying branching-type discipline. Then, we also focus on $\widetilde{T}_{S-}^{(i)}(\cdot)$, $\gamma_+^{(i)}(\cdot)$ and $H_+^{(i)}(\cdot)$. We apply the results from Sect. 3.3 to obtain the relation between $\beta^{(i)}(\mathbf{z})$ and $\alpha^{(i)}(\mathbf{z})$.

Let $T_S^{(i)}$ denote the service time of a customer at $Q^{(i)}$ with distribution function $\widehat{T}_S^{(i)}(t)$ and LST $\widetilde{T}_S^{(i)}(s)$, let $q_n^{(i)}$ denote the probability that a customer of the n th generation goes to the server for service, and let $T_V^{(i)}$ denote the (maximal) visit time of the server to $Q^{(i)}$. We assume that $T_V^{(i)}$ is exponentially distributed with rate $\xi^{(i)}$. Note that it is assumed that the service-time distribution $\widehat{T}_S^{(i)}(\cdot)$ does not depend on the generation.

It is readily seen that the probability that the next customer is the final one is given by

$$p^{(i)} = \mathbb{P}(T_V^{(i)} \leq T_S^{(i)}) = \int_0^\infty 1 - e^{-\xi^{(i)}t} d\widehat{T}_S^{(i)}(t) = 1 - \widetilde{T}_S^{(i)}(\xi^{(i)}), \tag{14}$$

and that the LST of the service time of a regular customer is given by

$$\begin{aligned} \widetilde{T}_{S+}^{(i)}(s) &= \mathbb{E}\left[e^{-s \min(T_S^{(i)}, T_V^{(i)})} | T_S^{(i)} < T_V^{(i)}\right] = \frac{\int_0^\infty e^{-\xi^{(i)}t} e^{-st} d\widehat{T}_S^{(i)}(t)}{1 - p^{(i)}} \\ &= \frac{\widetilde{T}_S^{(i)}(s + \xi^{(i)})}{\widetilde{T}_S^{(i)}(\xi^{(i)})}. \end{aligned} \tag{15}$$

By combining the formulas above with Th. 3 and Remark 1, we obtain the following result.

Theorem 4 *For a polling system with batch Poisson arrivals operating under a branching-type service discipline combined with the exhaustive time-limited regime with the preemptive-repeat-random strategy, the relation between the joint p.g.f. of the queue lengths at the start and the end of a server visit to $Q^{(i)}$ reads*

$$\beta^{(i)}(\mathbf{z}) = \gamma_-^{(i)}(\mathbf{z})(\alpha^{(i)}(\mathbf{z}) - \alpha^{(i)}(\mathbf{z}_\star^{(i)})) + \gamma_+^{(i)}(\mathbf{z})\alpha^{(i)}(\mathbf{z}_\star^{(i)}),$$

with $\mathbf{z}_\star^{(i)} = (z_1, \dots, z_{i-1}, H_+^{(i)}(\mathbf{z}), z_{i+1}, \dots, z_M)$ and

$$\gamma_-^{(i)}(\mathbf{z}) = \frac{\left(1 - \tilde{T}_S^{(i)}(\xi^{(i)})\right) \tilde{T}_{S^-}^{(i)}(\Lambda(\mathbf{z}))r_-^{(i)}(\mathbf{z})}{z_i - \tilde{T}_S^{(i)}(\xi^{(i)} + \Lambda(\mathbf{z}))r_+^{(i)}(\mathbf{z})},$$

where, for the E-TL case,

$$\gamma_+^{(i)}(\mathbf{z}) = 1,$$

and for the P-TL case,

$$\gamma_+^{(i)}(\mathbf{z}) = \frac{\xi^{(i)} + \lambda\gamma_-^{(i)}(\mathbf{z})\left(\widehat{A}(\mathbf{z}) - \widehat{A}_1^{(i)}\left(H_+^{(i)}(\mathbf{z}), \mathbf{z}\right)\right)}{\xi^{(i)} + \lambda\left(1 - \widehat{A}_1^{(i)}\left(H_+^{(i)}(\mathbf{z}), \mathbf{z}\right)\right)}.$$

The function $\tilde{T}_{S^-}^{(i)}(s)$ depends on the interruption rule and the function $H_+^{(i)}(\mathbf{z})$ on the underlying branching-type discipline.

Remark 4 To apply Th. 4 to specific models, we need to find the functions $\tilde{T}_{S^-}^{(i)}(s)$ and $H_+^{(i)}(\mathbf{z})$. As examples, we will specify these functions for two interruption rules and three branching-type strategies. For any of the six combinations of strategy and interruption rule, we can then easily find $\gamma_-^{(i)}(\mathbf{z})$ and $\gamma_+^{(i)}(\mathbf{z})$, with which we specify the relation between $\alpha^{(i)}(\mathbf{z})$ and $\beta^{(i)}(\mathbf{z})$.

Interruption rules We consider two interruption rules, namely the *preemptive-repeat-random* strategy, i.e., the server immediately leaves after the timer expires, and at the next server visit a new service time will be drawn from the original service-time distribution for the interrupted service; and the *non-preemptive* strategy where the server finishes the ongoing service.

For the preemptive rule, the LST of the service time of the final customer is given by

$$\begin{aligned} \tilde{T}_{S-}^{(i)}(s) &= \mathbb{E}\left[e^{-sT_V^{(i)}} \mid T_V^{(i)} \leq T_S^{(i)}\right] = \frac{\int_0^\infty \int_0^t \xi^{(i)} e^{-\xi^{(i)}v} e^{-sv} dv d\widehat{T}_S^{(i)}(t)}{p^{(i)}} \\ &= \frac{\xi^{(i)}}{\xi^{(i)} + s} \frac{1 - \tilde{T}_S^{(i)}(s + \xi^{(i)})}{1 - \tilde{T}_S^{(i)}(\xi^{(i)})}, \end{aligned} \tag{16}$$

and the p.g.f. of the routing probabilities by $r_-^{(i)}(\mathbf{z}) = z_i$. For the non-preemptive rule, the LST of the service time of the final customer is given by

$$\begin{aligned} \tilde{T}_{S-}^{(i)}(s) &= \mathbb{E}\left[e^{-sT_S^{(i)}} \mid T_V^{(i)} \leq T_S^{(i)}\right] = \frac{\int_0^\infty (1 - e^{-\xi^{(i)}t}) e^{-st} d\widehat{T}_S^{(i)}(t)}{p^{(i)}} \\ &= \frac{\tilde{T}_S^{(i)}(s) - \tilde{T}_S^{(i)}(s + \xi^{(i)})}{1 - \tilde{T}_S^{(i)}(\xi^{(i)})}, \end{aligned} \tag{17}$$

and the p.g.f. of the routing probabilities by $r_-^{(i)}(\mathbf{z}) = r_+^{(i)}(\mathbf{z})$.

Remark 5 Conjecture 5.13 in [4], page 112, is a special case of Th. 4 combined with the preemptive-repeat-random strategy and independent Poisson arrival streams at all queues (see Remark 3). This is readily seen, since for the preemptive-repeat-random strategy the LST of a non-preempted service time is given by Eq. (16) and we can write

$$\gamma_-^{(i)}(\mathbf{z}) = \frac{\xi^{(i)}}{\xi^{(i)} + \Lambda(\mathbf{z})} \frac{\left(1 - \tilde{T}_S^{(i)}(\xi^{(i)} + \Lambda(\mathbf{z}))\right) z_i}{z_i - \tilde{T}_S^{(i)}(\xi^{(i)} + \Lambda(\mathbf{z}))r_+^{(i)}(\mathbf{z})}$$

and, for the P-TL case,

$$\gamma_+^{(i)}(\mathbf{z}) = \gamma_-^{(i)}(\mathbf{z}) + \xi^{(i)} \frac{\left(z_i - r_+^{(i)}(\mathbf{z})\right) \tilde{T}_S^{(i)}(\xi^{(i)} + \Lambda(\mathbf{z}))}{z_i - r_+^{(i)}(\mathbf{z}) \tilde{T}_S^{(i)}(\xi^{(i)} + \Lambda(\mathbf{z}))}.$$

Branching-type strategy We can also vary the underlying branching-type service discipline. We focus on four disciplines, two well-known ones, the *Bernoulli gated* discipline and the *Bernoulli exhaustive* discipline, and two new ones, the *1+1* discipline introduced in Ex. 1 and the so-called *2G-gated* discipline, where only first- and second-generation customers are served. Similar observations can be made for other branching-type service disciplines. First, consider the Bernoulli gated discipline, in which only first-generation customers are served. Second-generation customers have to wait until the next server visit. In this case, finding the total indirect replacement

of a customer is relatively simple; all the customers who arrive during its service are indirect replacements, so

$$H_+^{(i)}(\mathbf{z}) = q_1^{(i)} \tilde{T}_{S_+}^{(i)}(\Lambda(\mathbf{z})) r_+^{(i)}(\mathbf{z}) + (1 - q_1^{(i)}) z_i.$$

Secondly, consider the Bernoulli exhaustive discipline, where every customer that arrives during a visit to $Q^{(i)}$ is also served during the same visit with some fixed probability $q^{(i)}$ for all generations. For the exhaustive case, the assumption that $q^{(i)}$ does not depend on the generation, means that $\widehat{A}_n^{(i)}(z, \mathbf{z})$ defined in Eq. (13) is independent of the generation too. We, therefore, omit the index for the generation. Following the same steps as in Example 2, we find that $H_+^{(i)}(\mathbf{z}) = q^{(i)} H_{++}^{(i)}(\mathbf{z}) + (1 - q^{(i)}) z_i$, where $H_{++}^{(i)}(\mathbf{z})$ satisfies

$$H_{++}^{(i)}(\mathbf{z}) = \tilde{T}_{S_+}^{(i)}\left(\lambda \left(1 - \widehat{A}_1^{(i)}\left(H_{++}^{(i)}(\mathbf{z}), \mathbf{z}\right)\right)\right) r_+^{(i)}(\mathbf{z}).$$

The third example we focus on is the 1 + 1 discipline (see Ex. 1). We find, by distinguishing the cases where a first-generation customer has direct replacements or not, that

$$\begin{aligned} H_+^{(i)}(\mathbf{z}) &= \tilde{T}_{S_+}^{(i)}\left(\lambda \left(1 - \widehat{A}_1^{(i)}(0, \mathbf{z})\right)\right) r_+^{(i)}(\mathbf{z}) \\ &\quad + \frac{\tilde{T}_{S_+}^{(i)}(\Lambda(\mathbf{z})) - \tilde{T}_{S_+}^{(i)}\left(\lambda \left(1 - \widehat{A}_1^{(i)}(0, \mathbf{z})\right)\right)}{z_i} \\ &\quad \times \tilde{T}_{S_+}^{(i)}(\Lambda(\mathbf{z})) \left(r_+^{(i)}(\mathbf{z})\right)^2. \end{aligned}$$

As a last example for a branching-type discipline, we focus on the so-called Bernoulli 2G-gated discipline in which only first- and second-generation customers are served; in the terminology of this paper, a second-generation customer only has indirect replacements. In this discipline, the joint p.g.f. of the total indirect replacement of a family is given by

$$H_+^{(i)}(\mathbf{z}) = \tilde{T}_{S_+}^{(i)}\left(\lambda \left(1 - \widehat{A}_1^{(i)}\left(\tilde{T}_{S_+}^{(i)}(\Lambda(\mathbf{z}), \mathbf{z})\right)\right)\right).$$

5 Extension

The SRS discipline discussed in this paper can be extended to cover a variant of the exponential time-limited queues introduced in Eliazar and Yechiali [9]. In this paper, the authors study exponential time-limited queues with the exhaustive service discipline with preemptive-repeat interruptions and with non-preemptive interruptions. In addition, they consider the case where customers who are present in $Q^{(i)}$ at the time of the interruption are served during the ongoing server visit but all customers who

arrive after that time have to wait for a new visit. The interrupted service is normally finished.

In the setting of the SRS, we would not only have regular customers and a final customer but also the so-called tail customers who have indirect replacements which may be, stochastically different from the regular and the final customers. To indicate that a function is related to a tail customer, we use the subscript “=”. We again give first the results for the general SRS model, then for the case of service-based indirect replacements and, finally, the results for the exponentially time-limited systems.

For the general case, it is easily verified that Eqs. (5) and (6) transform into

$$\beta^{(i)}(\mathbf{z}) = \gamma_-^{(i)}(\mathbf{z})(\alpha^{(i)}(\mathbf{z}_=^{(i)}) - \alpha^{(i)}(\mathbf{z}_\star^{(i)})) + \alpha^{(i)}(\mathbf{z}_\star^{(i)}),$$

with $\mathbf{z}_=^{(i)} = (z_1, \dots, z_{i-1}, H_-^{(i)}(\mathbf{z}), z_{i+1}, \dots, z_M)$ and

$$\gamma_-^{(i)}(\mathbf{z}) := \frac{H_-^{(i)}(\mathbf{z})}{H_-^{(i)}(\mathbf{z}) - H_+^{(i)}(\mathbf{z})}.$$

Note that for the original variant we studied, the indirect replacement of a tail customer is just itself, so $H_-^{(i)}(\mathbf{z}) = z_i$.

For the service-based indirect replacement discipline, it is a bit more complicated, since a final customer might have direct replacements that are served during the same visit, so we need to redefine $H_{n-}^{(i)}(z, \mathbf{z}) := \mathbb{E}\left[z^{S_n^{(i)}} \mathbf{z}^{\mathbf{R}_n^{(i)}} \mathbb{1}_F\right]$. With this notation, we can rewrite Eq. (8) as follows:

$$G_{n-}^{(i)}(\mathbf{z}) = \frac{H_{n-}^{(i)}(\mathbf{z})}{H_{n=}^{(i)}(\mathbf{z}) - H_{n+}^{(i)}(H_{n+1=}^{(i)}(\mathbf{z}), \mathbf{z})} \times \left(G_{n-1+}^{(i)}(H_{n=}^{(i)}(\mathbf{z}), \mathbf{z}) - G_{n-1+}^{(i)}(H_{n+}^{(i)}(H_{n+1=}^{(i)}(\mathbf{z}), \mathbf{z}), \mathbf{z})\right).$$

Furthermore, we have to assume that the p.g.f.’s $H_{n=}^{(i)}(\mathbf{z})$ do not depend on the generation n ; under this assumption, the results and proofs are similar to those of the theorems in Sect. 3.2.

Finally, we consider the exponential time-limited system with the Bernoulli exhaustive discipline where the probability $q_n^{(i)}$ that a customer will visit the server does not depend on the generation of the customer. In this system, a timer interrupts the normal service. After the timer expires, all customers already present at $Q^{(i)}$ will be served (with probability $q_n^{(i)}$) whereas customers that arrive during the visit after the interruption have to wait for the next visit of the server. In this case, the probability that the next, non-tail, customer is interrupted is given by Eq. (14) and the LST of the service time of a regular customer by Eq. (15). The service-time LST of a tail customer is given by $\tilde{T}_S^{(i)}(s)$ and therefore $H_-^{(i)}(\mathbf{z}) = q_n^{(i)} \tilde{T}_S^{(i)}(\Lambda(\mathbf{z}) + (1 - q_n^{(i)})z_i)$. The LST of the final customer is given by Eq. (17); however, we have to distinguish between the time of service before (denoted by $T_{SB}^{(i)}$) and after (denoted by $T_{SA}^{(i)}$) the interruption

(timer expiration). In this case, we find that

$$\begin{aligned} & \mathbb{E}\left[e^{-(s_1 T_{SB}^{(i)} + s_2 T_{SA}^{(i)})} \mathbb{1}_{\{T_V^{(i)} \leq T_{SB}^{(i)} + T_{SA}^{(i)}\}}\right] \\ &= \int_0^\infty \int_0^t \xi^{(i)} e^{-\xi^{(i)} v} e^{-s_1 v} e^{-s_2(t-v)} dv d\widehat{T}_S^{(i)}(t) \\ &= \int_0^\infty e^{-s_2 t} \int_0^t \xi^{(i)} e^{-(\xi^{(i)} + s_1 - s_2)v} dv d\widehat{T}_S^{(i)}(t) \\ &= \int_0^\infty e^{-s_2 t} \frac{\xi^{(i)}}{\xi^{(i)} + s_1 - s_2} (1 - e^{-(\xi^{(i)} + s_1 - s_2)t}) d\widehat{T}_S^{(i)}(t) \\ &= \frac{\xi^{(i)}}{\xi^{(i)} + s_1 - s_2} \left(\widetilde{T}_S^{(i)}(s_2) - \widetilde{T}_S^{(i)}(\xi^{(i)} + s_1) \right). \end{aligned}$$

By setting $s_1 = \lambda \left(1 - \widehat{A}_1^{(i)}(H_{\equiv}^{(i)}(\mathbf{z}, \mathbf{z}))\right)$ and $s_2 = \Lambda(\mathbf{z})$, we find that the joint p.g.f. of the numbers of indirect replacements of an interrupted customer is given by

$$H_{n-}^{(i)}(\mathbf{z}) = \frac{\xi^{(i)} \left(\widetilde{T}_S^{(i)}(\Lambda(\mathbf{z})) - \widetilde{T}_S^{(i)}\left(\xi^{(i)} + \lambda \left(1 - \widehat{A}_1^{(i)}(H_{\equiv}^{(i)}(\mathbf{z}, \mathbf{z}))\right)\right) \right)}{\xi^{(i)} + \lambda \left(\widehat{A}(\mathbf{z}) - \widehat{A}_1^{(i)}(H_{\equiv}^{(i)}(\mathbf{z}, \mathbf{z})) \right)}.$$

With these observations, we can find similar results as in Sect. 4.

6 Queue length at the start of a service and customer sojourn time

In this section, we focus on a practical application of the results developed in this paper. We first focus on the p.g.f. of the joint queue length distribution at the start of an arbitrary service for the system with general branching-type discipline (see Sect. 3.1). We also derive this joint p.g.f. for continuous-time system with service-based branching-type discipline (see Sect. 3.3). Once we have obtained this p.g.f., we can use the techniques in [3] to find, for example, the expected sojourn time of a customer at $Q^{(i)}$.

6.1 The queue length at the start of a service

In this section, we consider the system of Sect. 3.1. To find the p.g.f. of the joint queue length distribution at the start of an arbitrary visit, we need to specify the routing of the server and the switchover times of the server between queues. For convenience, we assume a cyclic routing of the server, i.e., the server goes from $Q^{(i)}$ to $Q^{(i+1)}$, where $Q^{(M+1)}$ denotes $Q^{(1)}$. The number of new customers during switchovers form an independent process. The joint p.g.f. of the number of new customers during the switchover from $Q^{(i)}$ to $Q^{(i+1)}$ is denoted by $U^{(i,i+1)}(\mathbf{z})$. We can now specify key equation (2) to

$$\alpha^{(i+1)}(\mathbf{z}) = \beta^{(i)}(\mathbf{z})U^{(i,i+1)}(\mathbf{z}).$$

Together with the relation between $\alpha^{(i)}(\mathbf{z})$ and $\beta^{(i)}(\mathbf{z})$, found in Sect. 3 for several variants of the SRS polling systems, we can compute in a numerical experiment using an iterative algorithm $\alpha^{(i)}(\mathbf{z})$; see, for example, [1].

Before we focus on the derivation details of the p.g.f. of the joint queue length distribution, we first determine the expectation of both $N_V^{(j)}$, the number of served customers during a visit to $Q^{(j)}$, and $N_X^{(j)}$, the number of times the server becomes idle during that visit. To compute $\mathbb{E}[N_X^{(j)}]$, observe that once the server becomes idle during a visit to $Q^{(j)}$, the probability that it becomes idle again in the same visit equals $H_{X+}^{(j)}(H_+^{(j)}(\mathbf{1}), \mathbf{1})$, where $\mathbf{1}$ is a vector of size M with all entries equal to one. Moreover, the probability that the server becomes free at least once during this visit equals $\alpha^{(j)}(\mathbf{1}_\star^{(j)})$, where $\mathbf{1}_\star^{(j)}$ is a vector of size M with entries equal to one except for the j th entry which equals $H_{j+}^{(j)}(\mathbf{1})$. Therefore, it is easy to find

$$\mathbb{E}[N_X^{(j)}] = \frac{\alpha^{(j)}(\mathbf{1}_\star^{(j)})}{1 - H_{X+}^{(j)}(H_+^{(j)}(\mathbf{1}), \mathbf{1})}.$$

Next, since we have a stable system, $\mathbb{E}[N_V^{(j)}]$ equals the expected number of customers that arrive to $Q^{(j)}$ between the start of two consecutive server visits to $Q^{(j)}$. Customers enter the system either during a service, an extra time, or during a switchover time. This gives us the following set of equations:

$$\mathbb{E}[N_V^{(i)}] = \sum_{j=1}^M \left\{ \mathbb{E}[N_V^{(j)}] \frac{\partial H^{(j)}}{\partial z_i}(\mathbf{1}) + \mathbb{E}[N_X^{(j)}] \frac{\partial H_X^{(j)}}{\partial z_i}(\mathbf{1}) + \frac{\partial U^{(j,j+1)}}{\partial z_i}(\mathbf{1}) \right\}. \tag{18}$$

Note that this system has a unique solution due to the stability assumption.

To find the p.g.f. of the joint queue length distribution at the start of an arbitrary service in Th. 5 below, we closely follow the arguments of Eisenberg in [7] and of de Haan in [4]. Define the following events for $i = 1, \dots, M$:

- $S_s^{(i)}$: service starts at $Q^{(i)}$;
- $S_e^{(i)}$: service ends at $Q^{(i)}$;
- $V_s^{(i)}$: visit starts at $Q^{(i)}$;
- $V_e^{(i)}$: visit ends at $Q^{(i)}$;
- $X_s^{(i)}$: start of extra time at $Q^{(i)}$;
- $X_e^{(i)}$: end of extra time at $Q^{(i)}$.

It is readily seen that

$$\begin{aligned} S_s^{(i)} &= (S_s^{(i)} \cap V_s^{(i)}) \cup (S_s^{(i)} \cap X_e^{(i)}) \cup (S_s^{(i)} \cap S_e^{(i)}); \\ S_e^{(i)} &= (S_e^{(i)} \cap V_e^{(i)}) \cup (S_e^{(i)} \cap X_s^{(i)}) \cup (S_e^{(i)} \cap S_s^{(i)}); \end{aligned}$$

$$\begin{aligned} V_s^{(i)} &= (V_s^{(i)} \cap S_s^{(i)}) \cup (V_s^{(i)} \cap X_s^{(i)}); \\ V_e^{(i)} &= (V_e^{(i)} \cap S_e^{(i)}) \cup (V_e^{(i)} \cap X_e^{(i)}); \\ X_s^{(i)} &= (X_s^{(i)} \cap S_e^{(i)}) \cup (X_s^{(i)} \cap V_s^{(i)}); \\ X_e^{(i)} &= (X_e^{(i)} \cap S_s^{(i)}) \cup (X_e^{(i)} \cap V_e^{(i)}). \end{aligned}$$

Therefore, $S_s^{(i)} \cup V_e^{(i)} \cup X_s^{(i)} = S_e^{(i)} \cup V_s^{(i)} \cup X_e^{(i)}$. Next consider the average of the generating functions of the joint queue length over the first k events that occur; by the assumption that the joint queue length is ergodic, we can take the limit $k \rightarrow \infty$ to find that

$$\widehat{S}_s^{(i)}(\mathbf{z}) + \widehat{V}_e^{(i)}(\mathbf{z}) + \widehat{X}_s^{(i)}(\mathbf{z}) = \widehat{S}_e^{(i)}(\mathbf{z}) + \widehat{V}_s^{(i)}(\mathbf{z}) + \widehat{X}_e^{(i)}(\mathbf{z}),$$

where, for example, $\widehat{S}_s^{(i)}(\mathbf{z})$ denotes the p.g.f. of the joint queue length at an arbitrary service start at $Q^{(i)}$ and a service starts at $Q^{(i)}$. This can be rewritten as

$$\widehat{S}_s^{(i)}(\mathbf{z}) - \widehat{S}_e^{(i)}(\mathbf{z}) = \widehat{V}_s^{(i)}(\mathbf{z}) - \widehat{V}_e^{(i)}(\mathbf{z}) - \left(\widehat{X}_s^{(i)}(\mathbf{z}) - \widehat{X}_e^{(i)}(\mathbf{z}) \right). \tag{19}$$

Note that $\widehat{S}_s^{(i)}(\mathbf{1})$ equals the fraction of events that are a service start at $Q^{(i)}$. So $\widetilde{S}_s^{(i)}(\mathbf{z}) := \widehat{S}_s^{(i)}(\mathbf{z})/\widehat{S}_s^{(i)}(\mathbf{1})$ represents the joint conditional p.g.f. of the queue length at an arbitrary service start at $Q^{(i)}$ given a service starts at $Q^{(i)}$. Analogously, define the conditional p.g.f.’s

$$\begin{aligned} \widetilde{S}_s^{(i)}(\mathbf{z}) &:= \widehat{S}_s^{(i)}(\mathbf{z})/\widehat{S}_s^{(i)}(\mathbf{1}), \\ \widetilde{X}_s^{(i)}(\mathbf{z}) &:= \widehat{X}_s^{(i)}(\mathbf{z})/\widehat{X}_s^{(i)}(\mathbf{1}), \end{aligned} \tag{20}$$

and

$$\widetilde{X}_e^{(i)}(\mathbf{z}) := \widehat{X}_e^{(i)}(\mathbf{z})/\widehat{X}_e^{(i)}(\mathbf{1}),$$

and note that $\alpha^{(i)}(\mathbf{z}) = \widehat{V}_s^{(i)}(\mathbf{z})/\widehat{V}_s^{(i)}(\mathbf{1})$ and $\beta^{(i)}(\mathbf{z}) = \widehat{V}_e^{(i)}(\mathbf{z})/\widehat{V}_e^{(i)}(\mathbf{1})$. Furthermore, it is readily seen that $\mathbb{E}\left[N_V^{(i)}\right] = \widehat{S}_s^{(i)}(\mathbf{1})/\widehat{V}_s^{(i)}(\mathbf{1})$ and that $\mathbb{E}\left[N_X^{(i)}\right] = \widehat{X}_s^{(i)}(\mathbf{1})/\widehat{V}_s^{(i)}(\mathbf{1})$. Since the number of starts and ends of a service, resp., extra time at $Q^{(i)}$ during the first k events differ at most by 1, $\widehat{S}_s^{(i)}(\mathbf{1}) = \widehat{S}_e^{(i)}(\mathbf{1})$ and $\widehat{X}_s^{(i)}(\mathbf{1}) = \widehat{X}_e^{(i)}(\mathbf{1})$ and we can write, by dividing Eq. (19) by $\widehat{V}_s^{(i)}(\mathbf{1})$,

$$\mathbb{E}\left[N_V^{(i)}\right] \left(\widetilde{S}_s^{(i)}(\mathbf{z}) - \widetilde{S}_e^{(i)}(\mathbf{z}) \right) = \alpha^{(i)}(\mathbf{z}) - \beta^{(i)}(\mathbf{z}) - \mathbb{E}\left[N_X^{(i)}\right] \left(\widetilde{X}_s^{(i)}(\mathbf{z}) - \widetilde{X}_e^{(i)}(\mathbf{z}) \right). \tag{21}$$

Theorem 5 *For a polling system operating under a self-ruling server discipline, the p.g.f. of the joint queue length distribution at a service start, given that the server starts a service at $Q^{(i)}$, is given by*

$$\tilde{S}_s^{(i)}(\mathbf{z}) = \frac{1}{\mathbb{E}[N_V^{(i)}]} \left(\frac{z_i}{z_i - H_+^{(i)}(\mathbf{z})} \right) \left(\alpha^{(i)}(\mathbf{z}) - \alpha^{(i)}(\mathbf{z}_\star^{(i)}) \frac{1 - H_{X+}^{(i)}(z_i, \mathbf{z})}{1 - H_{X+}^{(i)}(H_+^{(i)}(\mathbf{z}), \mathbf{z})} \right).$$

Proof Consider either a service or an extra time. The number of customers at the end of such a period is the sum of the customers that are present at the beginning of that same period plus the indirect replacement or the new customers that joined the system during that period, so we can write

$$\tilde{S}_e^{(i)}(\mathbf{z}) = \tilde{S}_s^{(i)}(\mathbf{z}) \frac{H^{(i)}(\mathbf{z})}{z_i}, \tag{22}$$

and

$$\tilde{X}_e^{(i)}(\mathbf{z}) = \tilde{X}_s^{(i)}(\mathbf{z}) H_X^{(i)}(z_i, \mathbf{z}).$$

We can interpret $\mathbb{E}[N_X^{(i)}] \tilde{X}_s^{(i)}(\mathbf{z})$ as the sum of all the p.g.f.’s at the start of the extra times during a visit to $Q^{(i)}$, so we can write

$$\mathbb{E}[N_X^{(i)}] \tilde{X}_s^{(i)}(\mathbf{z}) = \alpha^{(i)}(\mathbf{z}_\star^{(i)}) \sum_{j=0}^{\infty} (H_{X+}^{(i)}(H_+^{(i)}(\mathbf{z}), \mathbf{z}))^j = \frac{\alpha^{(i)}(\mathbf{z}_\star^{(i)})}{1 - H_{X+}^{(i)}(H_+^{(i)}(\mathbf{z}), \mathbf{z})}.$$

Next note that $\sum_{j=1}^M (\beta^{(j)}(\mathbf{z}) - \alpha^{(j+1)}(\mathbf{z})) = -\sum_{j=1}^M (\alpha^{(j)}(\mathbf{z}) - \beta^{(j)}(\mathbf{z}))$. Together with the above equations, Eqs. (20) and (21), and Th. 1, the theorem follows after some algebra. □

Remark 6 To derive this theorem, we could also have looked at the start of a service for an arbitrary customer during a visit. For the n th served customer with $n \leq N_s^{(i)}$, the number of customers at $Q^{(i)}$ to be served has decreased by $n - 1$, but there were also $n - 1$ replacement moments. For $n > N_s^{(i)}$, we should also include the replacement during (at least) one extra time followed by extra services. Elaborating along this line gives the same result. We chose to use the approach along the line of Eisenberg in [7] since this also applies in the more involved settings of the service-based strategy.

6.2 The queue length and workload at an arbitrary time

For a system where the service times, the arrival process and the switchovers are specified in continuous time (see Sect. 3.3), both the joint queue length and the workload at an arbitrary time are important performance measures. Other important performance measures, namely the expected time spent by a customer in the system or at a specific $Q^{(i)}$, are directly related to the expected queue length by Little’s law. For convenience, we only consider the case that at least one idle time or switchover time has a nonzero expectation. The queue length distribution at an arbitrary service start can be analyzed in the same way as in Sect. 6.1. Let $T_{SO}^{(i,i+1)}$ denote the switchover time from $Q^{(i)}$ to

$Q^{(i+1)}, T_C$ denote the cycle time of a server, $\widehat{N}_Q(\mathbf{z})$ the p.g.f. of the joint queue length distribution at an arbitrary time and $\widehat{V}(\mathbf{z})$ the p.g.f. of the joint workload distribution at an arbitrary time. Furthermore, we define the LST of $T_{S_{\pm}^{(i)}}$, the service time of an arbitrary customer at $Q^{(i)}$, by

$$\widetilde{T}_{S_{\pm}^{(i)}}(s) := \widetilde{T}_{S_{-}^{(i)}}(s) + (1 - p^{(i)})\widetilde{T}_{S_{+}^{(i)}}(s),$$

and the p.g.f. of its indirect replacement by

$$H_{*\pm}^{(i)}(\mathbf{z}) := p^{(i)}\widetilde{T}_{S_{-}^{(i)}}(\Lambda(\mathbf{z}))r^{(i)}(\mathbf{z})r_{-}^{(i)}(\mathbf{z}) + (1 - p^{(i)})\widetilde{T}_{S_{+}^{(i)}}(\Lambda(\mathbf{z}))r_{+}^{(i)}(\mathbf{z}).$$

In continuous time, it is convenient to consider the cycle time, T_C , of the server, that is, the time between two consecutive visit starts at, say, $Q^{(1)}$. To find the expectation $\mathbb{E}[T_C]$, we remark that the expected number of customers served at $Q^{(i)}$ during a cycle of the server satisfies (cf. Eq. (18))

$$\begin{aligned} \mathbb{E}[N_V^{(i)}] &= \sum_{j=1}^M \left\{ \mathbb{E}[N_V^{(j)}] \left(\lambda \mathbb{E}[T_{S_{\pm}^{(j)}}] \mathbb{E}[A^{(i)}] + r_i^{(j)} \right) \right. \\ &\quad \left. + \lambda \mathbb{E}[N_X^{(j)}] \mathbb{E}[W_I^{(j)}] \mathbb{E}[A^{(i)}] + \lambda \mathbb{E}[T_{SO}^{(j,j+1)}] \mathbb{E}[A^{(i)}] \right\}. \end{aligned}$$

The expected cycle time of a server is now readily found since

$$\mathbb{E}[T_C] = \sum_{j=1}^M \left\{ \mathbb{E}[N_V^{(j)}] \mathbb{E}[T_{S_{\pm}^{(j)}}] + \mathbb{E}[N_X^{(j)}] \mathbb{E}[W_I^{(j)}] + \mathbb{E}[T_{SO}^{(j,j+1)}] \right\}.$$

Lemma 2 *For a polling system with batch Poisson arrivals operating under a service-based SRS service discipline, the p.g.f. of the joint queue length distribution at a service start, given that the server is at $Q^{(i)}$, is given by*

$$\begin{aligned} \widetilde{S}_s^{(i)}(\mathbf{z}) &= \frac{1}{\mathbb{E}[N_V^{(i)}]} \left(\frac{z_i}{z_i - H_{*\pm}^{(i)}(\mathbf{z})} \right) \left(\frac{z_i - H^{(i)}(\mathbf{z})}{z_i - H_{+}^{(i)}(\mathbf{z})} \right) \\ &\quad \times \left(\alpha^{(i)}(\mathbf{z}) - \alpha^{(i)}(\mathbf{z}_{\star}^{(i)}) \frac{1 - H_{X+}^{(i)}(z_i, \mathbf{z})}{1 - H_{X+}^{(i)}(H_{+}^{(i)}(\mathbf{z}), \mathbf{z})} \right). \end{aligned}$$

Proof We proceed a similar way to the proof of Th. 5, where Eq. (22) replaced by $\widetilde{S}_e^{(i)}(\mathbf{z}) = \widetilde{S}_s^{(i)} H_{*\pm}^{(i)}(\mathbf{z})/z_i$. This is because here we have the service-based discipline. □

Remark 7 Th. 5 still holds when we assume that the members of a family are served consecutively, that $\mathbb{E}[N_V^{(j)}]$ represents the number of first-generation customers that

is served during a visit and that the end of a service corresponds to the end of the service of the last customer in a family being served. In other words, the total time to serve a family is then seen as one service time.

Theorem 6 *For a polling system with batch Poisson arrivals operating under a service-based SRS service discipline, the p.g.f. of the joint queue length distribution at an arbitrary time in the above continuous-time system satisfies*

$$\widehat{N_Q}(\mathbf{z}) = \frac{1}{\Lambda(\mathbf{z})\mathbb{E}[T_C]} \sum_{j=1}^M \mathbb{E}\left[N_V^{(j)}\right] \widetilde{T}_{S\pm}^{(j)}(\Lambda(\mathbf{z})) \left(r^{(j)}(\mathbf{z}) - z_j\right) \frac{\widetilde{S}_s^{(j)}(\mathbf{z})}{z_j},$$

where $\widetilde{S}_s^{(j)}(\mathbf{z})$ is given in Lemma 5.

Proof Along the same lines as the proof of Theorem 1 in [3], we can write, by the stochastic mean value theorem, that

$$\begin{aligned} \widehat{N_Q}(\mathbf{z}) = \frac{1}{\mathbb{E}[T_C]} \sum_{j=1}^M \left\{ \mathbb{E}\left[N_V^{(j)}\right] \mathbb{E}\left[T_{S\pm}^{(j)}\right] \widehat{N_{QS}^{(j)}}(\mathbf{z}) + \mathbb{E}\left[N_X^{(j)}\right] \mathbb{E}\left[W_I^{(j)}\right] \widehat{N_{QX}^{(j)}}(\mathbf{z}) \right. \\ \left. + \mathbb{E}\left[T_{SO}^{(j,j+1)}\right] \widehat{N_{QO}^{(j)}}(\mathbf{z}) \right\}, \end{aligned} \tag{23}$$

where $\widehat{N_{QS}^{(j)}}(\mathbf{z})$, $\widehat{N_{QX}^{(j)}}(\mathbf{z})$, and $\widehat{N_{QO}^{(j)}}(\mathbf{z})$ denote the p.g.f.’s of the joint queue length distribution at an arbitrary time during a service at $Q^{(j)}$, during an extra time at $Q^{(j)}$, and during a switchover period between $Q^{(j)}$ and $Q^{(j+1)}$. To find these p.g.f.’s, we have to multiply the p.g.f. of the queue length at the start of an interval by the p.g.f. of the number of arriving customers between the start and the arbitrary moment. We then find, by using Eq. (22), that

$$\begin{aligned} \widehat{N_{QS}^{(j)}}(\mathbf{z}) &= \widetilde{S}_s^{(j)}(\mathbf{z}) \frac{1 - \widetilde{T}_{S\pm}^{(j)}(\Lambda(\mathbf{z}))}{\Lambda(\mathbf{z})\mathbb{E}\left[T_{S\pm}^{(j)}\right]} = \frac{\widetilde{S}_s^{(j)}(\mathbf{z}) - \widetilde{S}_e^{(j)}(\mathbf{z})z_j/r^{(j)}(\mathbf{z})}{\Lambda(\mathbf{z})\mathbb{E}\left[T_{S\pm}^{(j)}\right]}, \\ \widehat{N_{QX}^{(j)}}(\mathbf{z}) &= \widetilde{X}_s^{(j)}(\mathbf{z}) \frac{1 - \widetilde{W}_I^{(j)}(\Lambda(\mathbf{z}))}{\Lambda(\mathbf{z})\mathbb{E}\left[W_I^{(j)}\right]} = \frac{\widetilde{X}_s^{(j)}(\mathbf{z}) - \widetilde{X}_e^{(j)}(\mathbf{z})}{\Lambda(\mathbf{z})\mathbb{E}\left[W_I^{(j)}\right]}, \end{aligned} \tag{24}$$

and

$$\widehat{N_{QO}^{(j)}}(\mathbf{z}) = \beta^{(j)}(\mathbf{z}) \frac{1 - \widehat{T}_{SO}^{(j,j+1)}(\Lambda(\mathbf{z}))}{\Lambda(\mathbf{z})\mathbb{E}\left[T_{SO}^{(j,j+1)}\right]} = \frac{\beta^{(j)}(\mathbf{z}) - \alpha^{(j+1)}(\mathbf{z})}{\Lambda(\mathbf{z})\mathbb{E}\left[T_{SO}^{(j,j+1)}\right]}.$$

When we write the numerator of the RHS of Eq. (24) as $\widetilde{S}_s^{(j)}(\mathbf{z}) - \widetilde{S}_e^{(j)}(\mathbf{z})z_j/r^{(j)}(\mathbf{z}) = \widetilde{S}_s^{(j)}(\mathbf{z}) - \widetilde{S}_e^{(j)}(\mathbf{z}) + \widetilde{S}_e^{(j)}(\mathbf{z})(1 - z_j/r^{(j)}(\mathbf{z}))$, the result then easily follows by combining the above equations for the p.g.f.’s with Eqs. (20), (22), and (23). \square

Remark 8 Consider a system where customers visit only one queue. We can then write $\mathbb{E}[N_V^{(j)}] = \lambda \mathbb{E}[T_C] \mathbb{E}[A^{(j)}]$ and

$$\widehat{N}_Q(\mathbf{z}) = \sum_{j=1}^M \frac{\mathbb{E}[A^{(j)}] (1 - z_j)}{1 - \widehat{A}(\mathbf{z})} \widetilde{S}_e^{(j)}(\mathbf{z}),$$

with $\widetilde{S}_e^{(j)}(\mathbf{z})$ as given in Eq. (22). Consider the marginal distribution of the queue length at $Q^{(i)}$ at an arbitrary epoch, that is,

$$\widehat{N}_Q(\mathbf{1}_{z_i}^{(i)}) = \frac{\mathbb{E}[A^{(i)}] (1 - z_i)}{1 - \widehat{A}(\mathbf{1}_{z_i}^{(i)})} \widetilde{S}_e^{(j)}(\mathbf{1}_{z_i}^{(i)}),$$

where $\mathbf{1}_{z_i}^{(i)}$ is a vector of size M with entries equal to one with the i th entry replaced by z_i . By multiplying this equation by the p.g.f. of the number of customers arriving to $Q^{(i)}$ in front of an arbitrary customer for $Q^{(i)}$ in the same batch, which is given by $(1 - \widehat{A}(\mathbf{1}_{z_i}^{(i)})) / \mathbb{E}[A^{(i)}] (1 - z_i)$, we get the p.g.f. of the queue length observed by an arbitrary customer arriving to $Q^{(i)}$. This is a well-known result from batch arrival queues: the distribution at the departure of a customer equals the distribution at the arrival of a customer where we assume that the customers enter one by one, albeit at the same time.

Remark 9 Using Little’s law, the expectation of W_T , the sojourn time of an arbitrary customer in the system, is given by

$$\mathbb{E}[W_T] = \frac{\frac{d}{dz} \widehat{N}_Q(z, \dots, z) \Big|_{z=1}}{\lambda \frac{d}{dz} \widehat{A}(z, \dots, z) \Big|_{z=1}} = \frac{\frac{d}{dz} \widehat{N}_Q(z, \dots, z) \Big|_{z=1}}{\lambda \sum_{j=1}^M \mathbb{E}[A^{(j)}]},$$

and the expectation of $W_T^{(i)}$, the sojourn time of an arbitrary customer per visit to $Q^{(i)}$, by

$$\mathbb{E}[W_T^{(i)}] = \frac{\mathbb{E}[T_C]}{\mathbb{E}[N_V^{(j)}]} \frac{\partial \widehat{N}_Q}{\partial z_i}(\mathbf{1}).$$

Before we give the LST of the workload in the system, we introduce some notation. Let $\widetilde{\mathbf{T}}_{S_{\pm}}^{(*)}(\mathbf{s}) = (\widetilde{T}_{S_{\pm}}^{(1)}(s_1), \dots, \widetilde{T}_{S_{\pm}}^{(M)}(s_M))$. Then, the joint LST of the workload at the start of a service at $Q^{(j)}$, without the service time of the freshly started service, is given by $\widehat{V}_S^{(j)}(\mathbf{s}) = \widetilde{S}_s^{(j)}(\widetilde{\mathbf{T}}_{S_{\pm}}^{(*)}(\mathbf{s})) / \widetilde{T}_{S_{\pm}}^{(j)}(s_j)$. Furthermore, observe that the joint LST of past (T_P) and remaining (T_R) service time at an arbitrary epoch during an arbitrary service at $Q^{(j)}$ is given by $\mathbb{E}[e^{-sT_P - tT_R}] = (\widetilde{T}_{S_{\pm}}^{(j)}(s) - \widetilde{T}_{S_{\pm}}^{(j)}(t)) / (\mathbb{E}[T_{S_{\pm}}^{(j)}](t - s))$. Using

these observations and following both the proof of Th. 6 and [3] leads to the following theorem.

Theorem 7 *For a polling system with batch Poisson arrivals operating under a service-based SRS service discipline, the LST of the joint workload distribution at an arbitrary time satisfies*

$$\widehat{V}(\mathbf{s}) = \frac{1}{\Lambda(\widetilde{\mathbf{T}}_{S\pm}^{(*)}(\mathbf{s})) \mathbb{E}[T_C]} \sum_{j=1}^M \mathbb{E}[N_V^{(j)}] \frac{\widetilde{S}_s^{(j)}(\widetilde{\mathbf{T}}_{S\pm}^{(*)}(\mathbf{s}))}{\widetilde{T}_{S\pm}^{(j)}(s_j)} \times \left(\frac{s_j \widetilde{T}_{S\pm}^{(j)}(s_j) - \Lambda(\widetilde{\mathbf{T}}_{S\pm}^{(*)}(\mathbf{s})) \widetilde{T}_{S\pm}^{(j)}(\Lambda(\widetilde{\mathbf{T}}_{S\pm}^{(*)}(\mathbf{s})))}{\Lambda(\widetilde{\mathbf{T}}_{S\pm}^{(*)}(\mathbf{s})) - s_j} + r^{(j)}(\widetilde{\mathbf{T}}_{S\pm}^{(*)}(\mathbf{s})) \right).$$

7 Numerical results for a system with the 1 + 1 SRS discipline

In this section, we apply the results from the previous section to obtain numerical results for an example system with two queues with Poisson batch arrivals; one of the queues is operating under the 1+1 SRS discipline (see Example 1) with additional idle time while the other queue is served according to the gated discipline.

In the example system, there is a single server attending two types of tasks, for example, tasks A and B. All jobs need task B before leaving the system but some need also task A, where task A has to be finished before task B starts. It takes an exponentially distributed time with mean one for the server to switch from task A to B and vice versa, so that $\mathbb{E}[T_{SO}^{(AB)}] = \mathbb{E}[T_{SO}^{(BA)}] = 1$. The server, therefore, decides to handle first a (random) number of A tasks before switching to B tasks. The served task A jobs join the task B queue. After switching, the server will handle all B tasks present at the start time of his handling according to the gated service discipline. After handling these tasks B, the server switches immediately to task A jobs. In line with the description above, we assume a 1 + 1 SRS discipline for task A with the probability that a task A job is the last to be processed before the server switches to task B equal to $p^{(A)}$. Moreover, a job will always go to the server, both for task A and for task B, so $q^{(A)} = q^{(B)} = 1$. If the server does not decide to switch before the queue for task A is empty, he waits 2 time units for another task A job to arrive. The service times of task A have an Erlang distribution with LST $\widetilde{T}_S^{(A)}(s) = 16/(s + 4)^2$ and the service times of task B have a hyper exponential distribution with LST $\widetilde{T}_S^{(B)}(s) = 3/(s + 4) + 1/(4 + 5s)$. Customers arrive in batches of size $N_{BS} = 2$ according to a Poisson process with rate $\lambda = 1/3$. A fraction of $p_{TB} = 1/3$ of the arriving customers go directly to the queue of task B, independent of other customers. This makes the joint p.g.f. of the number of jobs in an arriving batch starting with task A and, respectively, task B, $\widehat{A}(z_A, z_B) = (4z_A^2 + 4z_A z_B + z_B^2)/9$. In the following, we shall determine for which $p^{(A)}$ the expected sojourn time of an arbitrary customer is minimal.

Before we find the best value of $p^{(A)}$, we first determine its possible range. Let $p_S^{(A)}$ be the supremum of all $p^{(A)}$ for which the system is stable and suppose $p^{(A)} < p_S^{(A)}$. Then

$$\mathbb{E}[T_C] = \lambda N_{BS} \mathbb{E}[T_C] \left((1 - p_{TB}) \mathbb{E}\left[T_{S\pm}^{(A)}\right] + \mathbb{E}\left[T_{S\pm}^{(B)}\right] \right) + \mathbb{E}\left[T_{SO}^{(AB)} + T_{SO}^{(BA)}\right] + \mathbb{P}(N_X^{(A)} > 0) \mathbb{E}\left[N_X^{(A)} | N_X^{(A)} > 0\right] \mathbb{E}\left[W_I^{(A)}\right].$$

This gives

$$\mathbb{E}[T_C] = \frac{\mathbb{P}(N_X^{(A)} > 0) \mathbb{E}\left[N_X^{(A)} | N_X^{(A)} > 0\right] \mathbb{E}\left[W_I^{(A)}\right]}{1 - \rho^*}$$

with

$$\rho^* = \lambda N_{BS} \left((1 - p_{TB}) \mathbb{E}\left[T_{S\pm}^{(A)}\right] + \frac{\mathbb{E}\left[T_{SO}^{(AB)} + T_{SO}^{(BA)}\right]}{\lambda N_{BS} \mathbb{E}[T_C]} + \mathbb{E}\left[T_{S\pm}^{(B)}\right] \right). \tag{25}$$

Obviously, for the system to be stable, $\rho^* < 1$. Since the server switches after at most a geometrically distributed number of task A jobs, the average number of served task A jobs during a cycle is at most $1/p^{(A)}$ so we need $1/p^{(A)} > \lambda N_{BS} (1 - p_{TB}) \mathbb{E}[T_C]$. Combine this with Eq. (25) to find that

$$p^{(A)} < p_U^{(A)} := \frac{1 - \lambda N_{BS} \left((1 - p_{TB}) \mathbb{E}\left[T_{S\pm}^{(A)}\right] + \mathbb{E}\left[T_{S\pm}^{(B)}\right] \right)}{\lambda N_{BS} (1 - p_{TB}) \left(\mathbb{E}\left[T_{SO}^{(AB)} + T_{SO}^{(BA)}\right] \right)} \tag{26}$$

is a necessary condition for stability. To investigate whether this condition is also sufficient, we can argue as follows: Intuitively, when $p^{(A)}$ is getting closer to $p_S^{(A)}$, the expected number of task A jobs at the beginning of a cycle will become large, which implies that $\mathbb{P}(N_X^{(A)} > 0)$ tends to zero and so

$$\mathbb{E}[T_C] \approx \lambda N_{BS} \mathbb{E}[T_C] \left((1 - p_{TB}) \mathbb{E}\left[T_{S\pm}^{(A)}\right] + \mathbb{E}\left[T_{S\pm}^{(B)}\right] \right) + \mathbb{E}\left[T_{SO}^{(AB)} + T_{SO}^{(BA)}\right],$$

which gives that

$$\mathbb{E}[T_C] \approx \frac{\mathbb{E}\left[T_{SO}^{(AB)} + T_{SO}^{(BA)}\right]}{1 - \lambda N_{BS} \left((1 - p_{TB}) \mathbb{E}\left[T_{S\pm}^{(A)}\right] + \mathbb{E}\left[T_{S\pm}^{(B)}\right] \right)}.$$

The expected number of task A jobs that can be handled during a cycle equals $1/p^{(A)} > 1/p_U^{(A)} \approx \lambda N_{BS} \mathbb{E}[T_C] (1 - p_{TB})$, the expected number of task A jobs that arrive during

Table 1 The expected queue length at the beginning and end of a server visit to a queue

$p^{(A)}$	Server visit to Task A				Server visit to Task B			
	Beginning		End		Beginning		End	
	$\mathbb{E}[N_A]$	$\mathbb{E}[N_B]$	$\mathbb{E}[N_A]$	$\mathbb{E}[N_B]$	$\mathbb{E}[N_A]$	$\mathbb{E}[N_B]$	$\mathbb{E}[N_A]$	$\mathbb{E}[N_B]$
0.05	3.194	0.916	0.918	6.020	1.363	6.242	2.750	0.694
0.10	3.503	0.841	1.376	5.351	1.820	5.573	3.059	0.619
0.15	3.918	0.781	1.910	4.811	2.355	5.033	3.473	0.559
$p_*^{(A)}$	3.974	0.775	1.979	4.756	2.423	4.978	3.530	0.553
0.20	4.493	0.732	2.584	4.367	3.027	4.589	4.047	0.510
0.25	5.319	0.691	3.493	3.995	3.938	4.217	4.875	0.469
0.30	6.590	0.656	4.835	3.678	5.279	3.900	6.146	0.433
0.35	8.745	0.625	7.050	3.406	7.494	3.628	8.300	0.403
0.40	13.069	0.599	11.426	3.169	11.871	3.391	12.624	0.377
0.45	26.184	0.576	24.587	2.961	25.032	3.184	25.739	0.354

Table 2 The expected queue length at the beginning of a service and at an arbitrary time and the expected sojourn time of an arbitrary customer

$p^{(A)}$	Service beg. A		Service beg. B		Arbitrary time		Arbitrary customer		
	$\mathbb{E}[N_A]$	$\mathbb{E}[N_B]$	$\mathbb{E}[N_A]$	$\mathbb{E}[N_B]$	$\mathbb{E}[N_A]$	$\mathbb{E}[N_B]$	$\mathbb{E}[W_A]$	$\mathbb{E}[W_B]$	$\mathbb{E}[W_T]$
0.05	3.604	5.162	2.828	6.274	2.493	5.442	5.609	8.163	11.902
0.10	3.886	4.489	3.108	5.600	2.775	4.766	6.244	7.149	11.311
0.15	4.282	3.938	3.505	5.049	3.170	4.218	7.133	6.326	11.082
$p_*^{(A)}$	4.336	3.88	3.56	4.9930	3.225	4.161	7.256	6.242	11.080
0.20	4.842	3.473	4.073	4.589	3.731	3.764	8.394	5.647	11.243
0.25	5.661	3.073	4.907	4.196	4.550	3.385	10.235	5.078	11.901
0.30	6.927	2.724	6.192	3.856	5.816	3.065	10.238	5.078	11.903
0.35	9.080	2.415	8.369	3.559	7.969	2.793	13.087	4.597	13.322
0.40	13.407	2.146	12.726	3.304	12.296	2.565	27.662	3.849	22.290
0.45	26.525	1.901	25.878	3.075	25.414	2.370	57.181	3.5550	41.676

a cycle for $p^{(A)}$ close to $p_S^{(A)}$. Hence, Eq. (26) also seems to be sufficient. For the parameters in the example system, this implies that $p_S^{(A)} = 0.5$.

In Table 1, we present the expected queue length for task A (N_A) and task B (N_B) at the beginning and end of a visit. Table 2 contains the queue length both at the beginning of a service and at an arbitrary time. Furthermore, we present the sojourn time at an arbitrary time, where we consider the sojourn time of an arbitrary customer visiting task A (W_A), visiting task B (W_B), and in the whole system (W_T). Using a numerical search, we find that the optimal value of $p^{(A)}$ that minimizes the expected sojourn time is $p_*^{(A)} = 0.15573$.

8 Conclusion

We consider stable polling systems with a self-ruling server and find for these a relation between the joint p.g.f. of the queue lengths at the end of a server visit and $Q^{(i)}$, and the joint p.g.f. of the queue lengths at the start of the visit to this queue for a class of polling systems with the branching-type discipline. In [1, Sec. 6], the method implementation and the test of its performance from a computational point of view is analyzed. Along the lines of [7], once we have found the p.g.f.'s at the start and end of a visit, we derive the joint p.g.f. of queue lengths at the start of a customer service. For the general pure time-limited systems with Poisson arrivals, the departures of a server can be seen as a Poisson process, which implies that the p.g.f. at the end of a visit also represents the p.g.f. of the queue lengths of an arbitrary arrival. In this case, we can derive the joint LST of the workload at arbitrary time in the queues and the expected sojourn of an arbitrary customer in the queues.

A key assumption in this paper is the customer limit, which is geometrically distributed ($p^{(i)}$ is fixed per queue). As future work, it would be interesting to relax this assumption to cover customer limits with a discrete phase-type distribution. Another research direction is the determination of the stability condition of the SRS disciplines. To give an indication of why it is a study on its own, consider the following counter intuitive example, where extra arrivals make a system stable: Consider a system with two queues with the following characteristics:

- independent single Poisson arrival streams at the queues with rates λ_1 (specified later) and $\lambda_2 = 2$;
- at $Q^{(1)}$ the service time is 0, $p^{(1)} = 1$, that is, the server always decides that the first customer is the final one and the time the server waits idle for a customer to arrive, if any, has an exponential time length with rate $\xi = 1$;
- the service times at $Q^{(2)}$ have an exponential distribution with rate $\mu_2 = 4$, $p^{(2)} = 1$, and the server leaves immediately after becoming idle, so during a visit to $Q^{(2)}$, the expected number of arrivals to $Q^{(1)}$ is $\lambda_1/4$ and to $Q^{(2)}$ $1/2$.

When $\lambda_1 = 0$, $Q^{(1)}$ will be empty and the expected number of extra customers to $Q^{(2)}$ after an idle period at $Q^{(1)}$ is $\lambda_2/\xi = 2$, so $Q^{(2)}$ is not stable since the server serves only one customer per visit to $Q^{(2)}$. On the other hand, when $\lambda_1 = 3$, both queues are stable. We can see this as follows: $Q^{(2)}$ will always become empty when $Q^{(1)}$ is full since the expected number of new arrivals to $Q^{(2)}$ between two visit starts at this queue is $\lambda_2/\mu_2 = 1/2$; once $Q^{(2)}$ is empty, $Q^{(1)}$ will become empty. Note that, in this case, the expected number of extra customers in $Q^{(2)}$ after an idle period at $Q^{(1)}$ will be $\lambda_2/(\lambda_1 + \xi) = 1/2$. Since not every visit to $Q^{(1)}$ ends with an idle time, the expected number of new customers arriving at $Q^{(2)}$ in between to visit starts at $Q^{(2)}$ is less than one, which makes $Q^{(2)}$ stable.

For some special cases of the exponentially time-limited systems, stability proofs can be found in [4, Sec. 3.3 and 3.4], which are strongly related to the stability proof of Fricker and Jaibi [10] for a class of polling systems with non-preemptive and work-conserving service disciplines. This approach is also promising for time-limited systems with general underlying branching-type service disciplines.

Acknowledgements We thank the referees for carefully reading our manuscript. Following their comments resulted in a more general and more applicable paper.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

1. Al Hanbali, A., de Haan, R., Boucherie, R.J., van Ommeren, J.-K.: Time-limited polling systems with batch arrivals and phase-type service times. *Ann. Oper. Res.* **198**(1), 57–82 (2012)
2. Borst, S., Boxma, O.: Polling: past, present, and perspective. *TOP* **26**(3), 335–369 (2018)
3. Boxma, O., Kella, O., Kosinski, K.: Queue lengths and workloads in polling systems. *Oper. Res. Lett.* **39**(6), 401–405 (2011)
4. de Haan, R.: Queueing models for mobile Ad hoc networks. Ph.D. thesis, Enschede (June 2009). <http://doc.utwente.nl/61385/>
5. de Haan, R., Boucherie, R.J., Hanbali, A.A., van Ommeren, J.-K.: Transient analysis for exponential time-limited polling models under the preemptive repeat random policy. *Adv. Appl. Probab.* (accepted) (2019)
6. de Haan, R., Boucherie, R.J., van Ommeren, J.-K.: A polling model with an autonomous server. *Queueing Syst.* **62**(3), 279–308 (2009)
7. Eisenberg, M.: Queues with periodic service and changeover times. *Oper. Res.* **20**(2), 440–451 (1972)
8. Eliazar, I., Yechiali, U.: Polling under the randomly-timed gated regime. *Stoch. Mod.* **14**(1), 79–93 (1998)
9. Eliazar, I., Yechiali, U.: Randomly timed gated queueing systems. *SIAM J. Appl. Math.* **59**(2), 423–441 (1998)
10. Fricker, C., Jaibi, M.: Monotonicity and stability of periodic polling models. *Queueing Syst.* **15**(1–4), 211–238 (1994)
11. Fuhrmann, S.W.: A decomposition result for a class of polling models. *Queueing Syst.* **11**(1–2), 109–120 (1992)
12. Resing, J.: Polling systems and multitype branching processes. *Queueing Syst.* **13**(4), 409–426 (1993)
13. Sidi, M., Levy, H., Fuhrmann, S.: A queueing network with a single cyclically roving server. *Queueing Syst.* **11**(1–2), 121–144 (1992)
14. Takagi, H.: Queueing analysis of polling models: progress in 1990–1994. *Front. Queueing Mod. Appl. Sci. Eng.* **7**, 119 (1997)
15. Takagi, H.: Analysis and application of polling models. In: *Performance Evaluation: Origins and Directions*, LNCS 1769. Springer, Berlin, pp. 423–442 (2000)
16. Vishnevskii, V.M., Semenova, O.V.: Mathematical methods to study the polling systems. *Autom. Rem. Control* **67**(2), 173–220 (2006)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.