Queueing Syst (2014) 78:287–311
DOI 10.1007/s11134-014-9413-y

On two-queue Markovian polling systems with exhaustive service

Jan-pieter L. Dorsman · Onno J. Boxma · Robert D. van der Mei

Received: 13 March 2014 / Revised: 16 June 2014 / Published online: 24 July 2014

Abstract We consider a class of two-queue polling systems with exhaustive service, where the order in which the server visits the queues is governed by a discrete-time Markov chain. For this model, we derive an expression for the probability generating function of the joint queue length distribution at polling epochs. Based on these results, we obtain explicit expressions for the Laplace–Stieltjes transforms of the waiting-time distributions and the probability generating function of the joint queue length distribution at an arbitrary point in time. We also study the heavy-traffic behaviour of properly scaled versions of these distributions, which results in compact and closed-form expressions for the distribution functions themselves. The heavy-traffic behaviour turns out to be similar to that of cyclic polling models, provides insights into the main effects of the model parameters when the system is heavily loaded, and can be used to derive closed-form approximations for the waiting-time distribution or the queue length distribution.

J. L. Dorsman (✉) · O. J. Boxma
EURANDOM and Department of Mathematics and Computer Science, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands
e-mail: j.l.dorsman@tue.nl

O. J. Boxma
e-mail: o.j.boxma@tue.nl

J. L. Dorsman · R. D. van der Mei
Stochastics, Centrum Wiskunde & Informatica (CWI), P.O. Box 94079, 1090 GB Amsterdam, The Netherlands
e-mail: R.D.van.der.Mei@cwi.nl

R. D. van der Mei
Department of Mathematics, VU University Amsterdam, De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands

Keywords Markovian routing · Waiting-time distribution · Queue length distribution · Descendant set approach · Heavy-traffic behaviour

Mathematics Subject Classification 60K25 · 68M20 · 90B22

1 Introduction

In this paper, we study a class of queueing systems that consist of two queues, which are attended by a single server. The server visits the queues in order to provide service to customers there, and incurs positive random switch-over times when it moves from one queue to the next. Such systems are commonly called *polling systems*, and find their origin in a wealth of real-life applications, such as manufacturing environments and computer-communication systems. For an overview of the literature on polling systems, their applications and standard results on their analysis, we refer to surveys such as [4,20,25,32].

Many studies in the literature assume that the server visits the queues in a cyclic (or in the case of two queues, an alternating) order. This might in some cases, however, not be a realistic assumption, as the queue to be visited next is determined by an external random environment. As such, we study the case where the order in which the server visits the queues is governed by a Markov chain. Note that as a consequence, after a visit to a certain queue, it is now possible for the server to resume service at the same queue, after having incurred a switch-over time. Polling models with this Markovian routing mechanism occur for instance naturally in the modelling of cellular data services that implement so-called opportunistic scheduling to profit from multi-user diversity [16,31], which aims to utilise fading and shadowing of cellular users within a single cell to optimise bandwidth efficiency [15]. The basic idea of opportunistic scheduling is that a time slot (representing the right for transmission) is assigned to the user with the highest instantaneous signal-to-noise ratio among all users in a cell. In this way, access to the medium is randomly assigned to the multitude of users in a cell. Another example can be found in the context of wireless random-access networks. So-called Carrier-Sense Multiple-Access Collision-Avoidance (CSMA-CA) algorithms provide a common mechanism for governing the use of such a shared wireless medium in a distributed fashion. As is illustrated in [13], the dynamics of networks using these algorithms are under certain assumptions probabilistically equivalent to polling systems with a Markovian routing mechanism. Apart from many other applications that can be found in the field of computer-communication systems (e.g., [18]), Markovian polling systems may also be particularly useful in the modelling of production systems with machines processing multiple product types. The type of product that a machine should prioritise for processing at a certain point (equivalently, the queue that should be visited by the server at that point), may be dependent on the levels of external demand for each product type, and is thus better modelled by a random environment than a round-robin assumption.

It is surprising that in the wide body of the literature on polling systems, hardly any studies can be found that concern themselves with these so-called Markovian polling models. The reason for this may lie in the fact that the analysis of Markovian polling systems is generally considered to be much more complex than that of cyclic polling models. In particular, it is shown in [23] that the analysis of polling systems of which the queue length vectors at appropriately chosen points in time do not constitute a multi-type branching process with immigration (cf. [3]), is far less tractable than that of systems which do satisfy this so-called branching property. The few studies that can be found include [9], in which an expression for the expected amount of work in the system at an arbitrary moment is derived for a few service disciplines. This work is extended in [33], where expressions for the moments of the (joint) queue lengths for the same service disciplines are found. More recently, these performance measures have been derived for a much more general class of service disciplines in [13]. Results for a slightly more general form of Markovian routing, where the routing probabilities may depend on the event whether a queue is empty or not, are derived in [24]. The references mentioned mostly restrict their attention to the analysis of the first few moments of performance measures such as the waiting times or the joint queue length. Unlike these papers, we consider the special setting of two-queue Markovian polling models, where the queues are served exhaustively (i.e., the server will only start a switch-over period if the current queue is completely empty). These assumptions have the advantage that the system now actually does satisfy the branching property, which enables us to analyse the *complete* distributions of the performance measures mentioned.

Initially, we will be concerned with the waiting-time and queue length distributions when the workload offered to the server is such that the queues are stable. The analysis of studies on non-trivial two-queue polling systems, such as [7], oftentimes includes a solution to a Riemann-Hilbert boundary value problem. We, however, follow an approach similar to the analysis of [34], which uses a recursive iteration of a functional equation for the probability generating function (PGF) of the joint queue length distribution at moments the server starts a visit period, and therefore, avoids such a boundary value problem.

We also study the behaviour of the system in a heavy-traffic regime, i.e., when the workload offered to the server is scaled to such a proportion that the queues are on the verge of instability. Many techniques have been proposed to obtain the heavy-traffic behaviour of polling models. Initial studies for cyclic polling models can be found in [10, 11], where the occurrence of a so-called heavy-traffic averaging principle is established. This principle implies that, although the total scaled workload in the system tends to a Bessel-type diffusion in the heavy-traffic regime, the total workload in the system may be considered as a constant during the course of a polling cycle, while the loads of the individual queues fluctuate like a fluid model. In [27], several heavy-traffic limits have been established by taking limits in known expressions for the Laplace–Stieltjes transform (LST) of the waiting-time distribution. Alternatively, [21] provides similar results, by studying the behaviour of the descendant set approach (a numerical computation method, cf. [19]) in the heavy-traffic limit. Another tool in the heavy-traffic analysis of polling models is branching theory, theorems of which led to heavy-traffic results in [28]. Other methods for obtaining heavy-traffic behaviour include perturbation techniques, which have been exploited in [5] to study a specific class of non-branching polling models, and mean-value analysis (cf. [29]). In the heavy-traffic analysis of this paper, we partly use the key ideas of [21].

The main contributions of this paper can be summarised as follows. Under the assumption of a stable system, we obtain explicit expressions for several performance measures of the two-queue Markovian polling model with exhaustive service. In particular, we derive explicit expressions for (transforms of) the waiting-time distributions and the joint queue length distribution. Although these expressions consist of infinite products and are thus not in closed form, the products converge fast so that truncation leads to accurate approximations. We also consider the behaviour of the waiting-time and queue length distributions in a heavy-traffic regime. From a theoretical perspective, these results are interesting, since unlike previous studies, the complete distributions of the waiting times and queue lengths are analysed. The results in this paper are only proved for the two-queue exhaustive case, and are not easily extendable to more general assumptions. Nevertheless, they may offer some insights into the general case. For instance, we will show that, except for some minor adjustments, the heavy-traffic behaviour of two-queue Markovian polling models with exhaustive service is similar to that of cyclic polling models as derived in the literature. It seems that this relation also exists under more general assumptions, as we will conclude in Sect. 5. From a practical perspective, the results are useful, as they not only provide closed-form approximations for several performance measures that perform well when the system is heavily loaded (as is usual in practice), but also give insights into the key effects of the model parameters on the waiting times and queue lengths.

The remainder of this paper is structured as follows. In Sect. 2, we provide a detailed model description and the necessary notation. Section 3 derives the expressions for the mentioned performance measures under the assumption of a stable system, by taking a functional equation for the PGF for the joint queue length distribution at polling epochs as a starting point. Building on these results, we obtain the heavy-traffic behaviour of the system in Sect. 4. Finally, we formulate our conclusions and provide directions for further research in Sect. 5.

2 Model description and notation

In this section, we give a description of the polling system that we consider, and we introduce the notation required. We study a queueing system that consists of two infinite-buffer queues, Q_1 and Q_2, and a single server. Customers arrive at Q_i according to a Poisson process with intensity λ_i. We also refer to these customers as type-i customers. The generic service requirement of a type-i customer is represented by the random variable B_i, of which the LST is given by $\widetilde{B}_i(s) = \mathbb{E}[e^{-sB_i}]$ and the first two moments $\mathbb{E}[B_i]$ and $\mathbb{E}[B_i^2]$ are assumed to be finite. The workload that Q_i brings to the system is denoted by $\rho_i = \lambda_i \mathbb{E}[B_i]$. The aggregate workload offered to the server is denoted by $\rho = \rho_1 + \rho_2$. Initially, we study the system in case the aggregate workload is less than one, so that the queues are stable. After that, we study the system in the so-called *heavy-traffic* regime: the case where ρ tends to one, i.e., the point at which the queues are at the verge of instability.

The single server can only serve customers of one queue at a time. Hence, after serving a given number of customers at one queue (a visit period), the server will commence a switch-over period, also called a setup period, to initiate a new visit period at any queue. Such a setup takes a random amount of time. In most studies on two-queue polling systems, it is assumed that the server visits the queues in an alternating order.

We, however, adopt a more general server routing mechanism. We assume that when the server completes a visit period at Q_1, it commences with probability $p_1 \in [0, 1)$ a switch-over period to set up for yet another visit period at Q_1. In the other case (which occurs with probability $1 - p_1$), the server sets up for a visit to Q_2. Similarly, after visiting Q_2, the server prepares for another visit period at Q_2 with probability $p_2 \in [0, 1)$, otherwise it will set up for service at Q_1. This particular routing regime captures the often-assumed alternating routing regime by taking $p_1 = p_2 = 0$.

Observe that this routing mechanism falls in the class of so-called Markovian routing mechanisms: the position of the server is governed by a two-state discrete-time Markov chain of which the transition matrix has diagonal elements p_1 and p_2. By calculating the limiting distribution of this Markov chain, one finds that a fraction $\pi_1 = \frac{1-p_2}{2-p_1-p_2}$ of the switch-overs correspond to setups to Q_1, and the remaining fraction $\pi_2 = \frac{1-p_1}{2-p_1-p_2}$ are setups to Q_2. The probability $r_{i,j}$ that, provided the server is currently visiting Q_j, the server visited Q_i during the previous visit period follows straightforwardly from these computations. It is trivial to see that $r_{1,1} + r_{2,1} = 1$ and $r_{1,2} + r_{2,2} = 1$. In particular, we have that $r_{1,1} = \frac{p_1\pi_1}{p_1\pi_1+(1-p_2)\pi_2} = p_1$, $r_{1,2} = \frac{(1-p_1)\pi_1}{(1-p_1)\pi_1+p_2\pi_2} = 1 - p_2$, $r_{2,1} = \frac{(1-p_2)\pi_2}{p_1\pi_1+(1-p_2)\pi_2} = 1 - p_1$ and $r_{2,2} = \frac{(1-p_1)\pi_1}{(1-p_1)\pi_1+p_2\pi_2} = p_2$.

Over the course of a visit period, the server serves the queues in an exhaustive manner. In other words, the server will completely empty the queue, before it commences a switch-over period. To gain more insight in the dynamics of the exhaustive service discipline, let P_i denote the duration of a busy period in an $M/G/1$ queue with the same arrival process and service-time distribution as Q_i. This busy period consists of the service of its first customer, the services of the customers arriving during the service of the first customer (i.e., the children), the services of the customers arriving during the service of the children (i.e., the grandchildren), and so forth. The LST of P_i, denoted by $\widetilde{P}_i(s) = \mathbb{E}[e^{-sP_i}]$, is well known to satisfy the functional equation

$$\widetilde{P}_i(s) = \widetilde{B}_i(s + \lambda_i(1 - \widetilde{P}_i(s))). \tag{1}$$

We denote the number of customers that arrive at Q_j over the course of a busy period at Q_i with $K_{i,j}$. Its PGF $\widetilde{K}_{i,j}(z) = \mathbb{E}[z^{K_{i,j}}]$ is given by

$$\widetilde{K}_{i,j}(z) = \sum_{k=0}^{\infty} z^k \int_{t=0}^{\infty} e^{-\lambda_j t} \frac{(\lambda_j t)^k}{k!} \mathrm{d}\mathbb{P}(P_i < t) = \widetilde{P}_i(\lambda_j(1 - z)).$$

If a server starts a visit period at Q_i when there are n customers in that queue, the duration of that visit period is the n-fold convolution of P_i. It is important to note that if the server sets up for service at the same queue afterwards, Q_i is not necessarily empty at the start of the new visit period, as customers may have arrived over the course of the intermediate switch-over period.

We assume the distribution of the durations of the switch-over periods to depend on the queue the server just visited as well as the destination queue. In particular, we assume that a switch-over from Q_i to Q_j takes a continuously distributed stochastic amount of time $S_{i,j}$, of which the LST is given by $\widetilde{S}_{i,j}(s) = \mathbb{E}[e^{-sS_{i,j}}]$, $i, j \in \{1, 2\}$. The average duration of an arbitrary switch-over period incurred by the server is given by $\sigma = \sum_{i=1}^{2} \sum_{j=1}^{2} r_{i,j}\pi_j \mathbb{E}[S_{i,j}]$. Let $M_{i,j}^{(k)}$ be the number of arriving type-k customers over a switch-over period from Q_i to Q_j. Similar to the computations above, it can then be derived that the two-dimensional PGF $\widetilde{M}_{i,j}(z_1, z_2) = \mathbb{E}[\prod_{k=1}^{2} z_k^{M_{i,j}^{(k)}}]$ is given by

$$\widetilde{M}_{i,j}(z_1, z_2) = \int_{t=0}^{\infty} \sum_{n_1=0}^{\infty} \sum_{n_2=0}^{\infty} \prod_{k=1}^{2} \left(z_k^{n_k} e^{-\lambda_k t} \frac{(\lambda_k t)^{n_k}}{n_k!} \right) \mathrm{d}\mathbb{P}(S_{i,j} < t)$$
$$= \widetilde{S}_{i,j}(\lambda_1(1 - z_1) + \lambda_2(1 - z_2)).$$

We assume all interarrival times, service times and switch-over times to be independent.

In the remainder of this article, we are interested in the waiting-time distributions and the queue length distributions (including any customer in service) at several time epochs. Let $F_{i,j}$ be the number of customers present (waiting and in service) at Q_j when the server starts a visit period at Q_i (i.e., a polling epoch at Q_i). The joint distribution of $F_{i,1}$ and $F_{i,2}$ is represented by the two-dimensional PGF $\widetilde{F}_i(z_1, z_2) = \mathbb{E}[z_1^{F_{i,1}} z_2^{F_{i,2}}]$. Similarly, G_i represents the number of type-i customers at a polling epoch of Q_i, provided that the previous visit period of the server was at Q_{3-i}, and its PGF is given by $\widetilde{G}_i(z) = \mathbb{E}[z^{G_i}]$. The random variable L_j represents the number of customers at Q_j at an arbitrary point in time and the corresponding two-dimensional PGF is given by $\widetilde{L}(z_1, z_2) = \mathbb{E}[z_1^{L_1} z_2^{L_2}]$. The waiting time of a type-i customer that arrives at an arbitrary point in time is given by W_i, and its LST is given by $\widetilde{W}_i(s) = \mathbb{E}[e^{-sW_i}]$.

We analyse the system under stability conditions ($\rho < 1$) and heavy-traffic conditions ($\rho \uparrow 1$). More specifically, in the latter regime we scale the total arrival rate $\lambda_1 + \lambda_2$, while the ratio $\frac{\lambda_2}{\lambda_1}$ remains fixed. In this way, the heavy-traffic limit is uniquely defined. It is, moreover, convenient, for any variable x that is a function of ρ, to denote its value evaluated at $\rho = 1$ as $\hat{x}$. For example, $\hat{\rho}_i = \frac{\rho_i}{\rho}$, so that $\hat{\rho} = \hat{\rho}_1 + \hat{\rho}_2 = 1$ and $\hat{\lambda}_i = \frac{\hat{\rho}_i}{\mathbb{E}[B_i]}$. The waiting times and queue lengths tend to infinity in heavy traffic, and as such their distributions are not well defined in the limiting case. Therefore, we study the distributions of the scaled waiting times $\mathcal{W}_i = (1-\rho)W_i$ and the scaled queue lengths $\mathcal{L}_i = (1 - \rho)L_i$. The LST of the scaled waiting time is given by $\widetilde{\mathcal{W}}_i(s) = \mathbb{E}[e^{-s\mathcal{W}_i}]$. Likewise, the PGF of the scaled queue length is given by $\widetilde{\mathcal{L}}_i(z) = \mathbb{E}[z^{\mathcal{L}_i}]$.

Finally, we use $\boldsymbol{\lambda}(\boldsymbol{z})$ throughout this article as short-hand notation for $\lambda_1(1 - z_1) + \lambda_2(1 - z_2)$. Furthermore, $\mathbb{1}_{\{A\}}$ represents the indicator function of the event A. Any expression for an LST $\widetilde{C}(s) = \mathbb{E}[e^{-sC}]$ that we derive in this paper corresponding to any random variable C, holds for $\Re(s) > 0$. Likewise, any one-dimensional PGF $\widetilde{C}(z_1) = \mathbb{E}[z_1^C]$ or two-dimensional PGF $\widetilde{C}(z_1, z_2) = \mathbb{E}[\prod_{k=1}^{2} z_k^{C_k}]$ derived in this paper holds for any z_1 and z_2 for which $|z_1|$ and $|z_2|$ do not exceed one. In both cases, (cross-)moments of C can be computed by differentiation. In many cases, symbolic inversion of the transforms to the original cumulative distribution functions is hard when $\rho < 1$, due to the complex nature of the cumulative distribution functions.

However, the cumulative distribution functions themselves can be obtained numerically by a wide array of methods (e.g., [1,2,12]).

3 Analysis for arbitrarily loaded systems

In this section, we derive explicit expressions for the waiting-time distributions of either queue and the joint queue length distribution. In Sect. 3.1, we first obtain expressions for $\widetilde{F}_i(z_1, z_2)$, the joint queue length PGF at a polling epoch at Q_i. These results ultimately lead in Sect. 3.2 to expressions for the quantities $\widetilde{W}_1(s)$, $\widetilde{W}_2(s)$ and $\widetilde{L}(z_1, z_2)$. Throughout this section, we assume that $\rho < 1$, i.e., the case where the queues are stable. In Sect. 4, we will study the limiting case $\rho \uparrow 1$, the case where the system becomes critically loaded.

3.1 Joint queue length at polling epochs

To obtain explicit expressions for the PGF $\widetilde{F}_i(z_1, z_2)$, we start with a functional equation for this function. Such a functional equation has already been derived in [13,33] for a setting consisting of multiple queues and a wide class of service disciplines. Applying these results to our case, we obtain

$$\widetilde{F}_1(z_1, z_2) = r_{1,1}\widetilde{F}_1(\widetilde{K}_{1,2}(z_2), z_2)\widetilde{M}_{1,1}(z_1, z_2) + r_{2,1}\widetilde{F}_2(z_1, \widetilde{K}_{2,1}(z_1))\widetilde{M}_{2,1}(z_1, z_2). \tag{2}$$

This equation can be seen to hold by the following observations. With probability $r_{i,1}$, a visit to Q_1 is preceded by a visit period at Q_i, during which each type-i customer initially present and all of its offspring is served (i.e., not only the customer himself, but also his children, grandchildren, and so on). Over the course of each service of a type-i customer, a number of type-j customers, represented by the PGF $\widetilde{K}_{i,j}(z_j)$, arrives at Q_j. During the switch-over period $S_{i,1}$ between the two visits, the population of customers in the system grows with a number of arriving customers that is represented by $\widetilde{M}_{i,1}(z_1, z_2)$. By similar observations, we have that

$$\widetilde{F}_2(z_1, z_2) = r_{1,2}\widetilde{F}_1(\widetilde{K}_{1,2}(z_2), z_2)\widetilde{M}_{1,2}(z_1, z_2) + r_{2,2}\widetilde{F}_2(z_1, \widetilde{K}_{2,1}(z_1))\widetilde{M}_{2,2}(z_1, z_2). \tag{3}$$

We now develop explicit expressions for $\widetilde{F}_1(\widetilde{K}_{1,1}(z_2), z_2)$ and $\widetilde{F}_2(z_1, \widetilde{K}_{2,1}(z_1))$, so that (2) and (3) in turn offer explicit expressions for $\widetilde{F}_1(z_1, z_2)$ and $\widetilde{F}_2(z_1, z_2)$. To this end, we note that substituting $z_1 = \widetilde{K}_{1,2}(z_2)$ in (2) leads to

$$\widetilde{F}_1(\widetilde{K}_{1,2}(z_2), z_2) = \frac{r_{2,1}\widetilde{M}_{2,1}(\widetilde{K}_{1,2}(z_2), z_2)}{1 - r_{1,1}\widetilde{M}_{1,1}(\widetilde{K}_{1,2}(z_2), z_2)}\widetilde{F}_2(\widetilde{K}_{1,2}(z_2), \widetilde{K}_{2,1}(\widetilde{K}_{1,2}(z_2))). \tag{4}$$

Similarly, a substitution of $z_2 = \widetilde{K}_{2,1}(z_1)$ in (3) leads to

$$\widetilde{F}_2(z_1, \widetilde{K}_{2,1}(z_1)) = \frac{r_{1,2}\widetilde{M}_{1,2}(z_1, \widetilde{K}_{2,1}(z_1))}{1 - r_{2,2}\widetilde{M}_{2,2}(z_1, \widetilde{K}_{2,1}(z_1))}\widetilde{F}_1(\widetilde{K}_{1,2}(\widetilde{K}_{2,1}(z_1)), \widetilde{K}_{2,1}(z_1)). \tag{5}$$

A combination of (4) and (5) gives

$$\widetilde{F}_1(\widetilde{K}_{1,2}(z_2), z_2) = a_1(z_2)\widetilde{F}_1(\widetilde{K}_{1,2}(f_1(z_2)), f_1(z_2)), \tag{6}$$

where

$$a_1(z_2) = \frac{r_{2,1}\widetilde{M}_{2,1}(\widetilde{K}_{1,2}(z_2), z_2)}{1 - r_{1,1}\widetilde{M}_{1,1}(\widetilde{K}_{1,2}(z_2), z_2)} \frac{r_{1,2}\widetilde{M}_{1,2}(\widetilde{K}_{1,2}(z_2), f_1(z_2))}{1 - r_{2,2}\widetilde{M}_{2,2}(\widetilde{K}_{1,2}(z_2), f_1(z_2))} \text{ and}$$
$$f_1(z_2) = \widetilde{K}_{2,1}(\widetilde{K}_{1,2}(z_2)). \tag{7}$$

Observe that (6) constitutes an expression for $\widetilde{F}_1(\widetilde{K}_{1,2}(z_2), \cdot)$ in terms of $\widetilde{F}_1(\widetilde{K}_{1,2}(z_2), \cdot)$ itself. Therefore, iteration of (6) leads to

$$\widetilde{F}_1(\widetilde{K}_{1,2}(z_2), z_2) = \widetilde{F}_1(\widetilde{K}_{1,2}(f_1^{(\infty)}(z_2)), f_1^{(\infty)}(z_2)) \prod_{j=0}^{\infty} a_1(f_1^{(j)}(z_2)), \tag{8}$$

where $f_1^{(0)}(z_2) = z_2$ and $f_1^{(j)}(z_2) = f_1(f_1^{(j-1)}(z_2))$. By repeating the analysis above for $\widetilde{F}_2(z_1, \widetilde{K}_{2,1}(z_1))$, we obtain that

$$\widetilde{F}_2(z_1, \widetilde{K}_{2,1}(z_1)) = \widetilde{F}_2(f_2^{(\infty)}(z_1), \widetilde{K}_{2,1}(f_2^{(\infty)}(z_1))) \prod_{j=0}^{\infty} a_2(f_2^{(j)}(z_1)), \tag{9}$$

where

$$a_2(z_1) = \frac{r_{1,2}\widetilde{M}_{1,2}(z_1, \widetilde{K}_{2,1}(z_1))}{1 - r_{2,2}\widetilde{M}_{2,2}(z_1, \widetilde{K}_{2,1}(z_1))} \frac{r_{2,1}\widetilde{M}_{2,1}(f_2(z_1), \widetilde{K}_{2,1}(z_1))}{1 - r_{1,1}\widetilde{M}_{1,1}(f_2(z_1), \widetilde{K}_{2,1}(z_1))} \text{ and}$$
$$f_2(z_1) = \widetilde{K}_{1,2}(\widetilde{K}_{2,1}(z_1)), \tag{10}$$

$f_2^{(0)}(z_1) = z_1$ and $f_2^{(j)}(z_1) = f_2(f_2^{(j-1)}(z_1))$.

Now that we have derived two explicit expressions for $\widetilde{F}_1(\widetilde{K}_{1,2}(z_2), z_2)$ and $\widetilde{F}_2(z_1, \widetilde{K}_{2,1}(z_1))$, we show in the following two lemmas that $\widetilde{F}_1(\widetilde{K}_{1,2}(f_1^{(\infty)}(z_2)), f_1^{(\infty)}(z_2))$ and $\widetilde{F}_2(f_2^{(\infty)}(z_1), \widetilde{K}_{2,1}(f_2^{(\infty)}(z_1)))$ are well-defined constants and that the infinite products actually converge.

Lemma 3.1 *For* $z_1, z_2 \in \{z \in \mathbb{C} : |z| \leq 1\}$, $\widetilde{F}_1(\widetilde{K}_{1,2}(f_1^{(\infty)}(z_2)), f_1^{(\infty)}(z_2))$ *and* $\widetilde{F}_2(f_2^{(\infty)}(z_1), \widetilde{K}_{2,1}(f_2^{(\infty)}(z_1)))$ *are well-defined constants equal to one.*

Proof See Appendix 1. □

Lemma 3.2 *For* $z_1, z_2 \in \{z \in \mathbb{C} : |z| \leq 1\}$, *the products* $\prod_{j=0}^{\infty} a_1(f_1^{(j)}(z_2))$ *and* $\prod_{j=0}^{\infty} a_2(f_2^{(j)}(z_1))$ *converge.*

Proof See Appendix 2. □

Based on the analysis of $\widetilde{F}_1(\widetilde{K}_{1,2}(z_2), z_2)$ and $\widetilde{F}_2(z_1, \widetilde{K}_{2,1}(z_1))$, we can now derive expressions for $\widetilde{F}_1(z_1, z_2)$ and $\widetilde{F}_2(z_1, z_2)$ as follows.

Theorem 3.3 *Explicit expressions for $\widetilde{F}_1(z_1, z_2)$ and $\widetilde{F}_2(z_1, z_2)$ involving converging infinite products are given by*

$$\widetilde{F}_1(z_1, z_2) = r_{1,1}\widetilde{M}_{1,1}(z_1, z_2)\prod_{j=0}^{\infty} a_1(f_1^{(j)}(z_2)) + r_{2,1}\widetilde{M}_{2,1}(z_1, z_2)\prod_{j=0}^{\infty} a_2(f_2^{(j)}(z_1)) \tag{11}$$

and

$$\widetilde{F}_2(z_1, z_2) = r_{1,2}\widetilde{M}_{1,2}(z_1, z_2)\prod_{j=0}^{\infty} a_1(f_1^{(j)}(z_2)) + r_{2,2}\widetilde{M}_{2,2}(z_1, z_2)\prod_{j=0}^{\infty} a_2(f_2^{(j)}(z_1)). \tag{12}$$

Proof The theorem follows by combining (2), (3), (8), (9), Lemmas 3.1 and 3.2. □

We use the expressions of Theorem 3.3 to obtain the (PGF of the) joint queue length distribution at an arbitrary point in time in Sect. 3.2. We conclude this section with a couple of remarks.

Remark 3.1 The infinite products that arise in (11) and (12) have a clear interpretation. To see this, observe that when substituting $z_2 = 1$ in (11), one obtains $\widetilde{F}_1(z_1, 1) = \mathbb{E}[z_1^{F_{1,1}}]$, the PGF of the type-1 customers currently present at a polling epoch of Q_1. This yields

$$\widetilde{F}_1(z_1, 1) = r_{1,1}\widetilde{M}_{1,1}(z_1, 1) + r_{2,1}\widetilde{M}_{2,1}(z_1, 1)\prod_{j=0}^{\infty} a_2(f_2^{(j)}(z_1)), \tag{13}$$

since $a_1(1) = f_1(1) = 1$. This expression can be interpreted as follows. At the end of the previous visit period at Q_1, there are no type-1 customers in the system. Thus, with probability $r_{1,1}$, the number of type-1 customers that have arrived since the previous visit period at Q_1, did so over the course of a switch-over period $S_{1,1}$. This number of customers is represented by the PGF $\widetilde{M}_{1,1}(z_1, 1)$. With probability $r_{2,1}$, the previous visit period was at Q_2, so that $\widetilde{F}_1(z_1, 1)$ equals G_1 in this case, i.e., the number of type-1 customers present at a polling epoch of Q_1 given that the server's previous visit was at Q_2. This number of type-1 customers present not only consists of type-1 customers that arrived during a switch-over period $S_{2,1}$, but also type-1 customers that arrived between the end of the previous visit period at Q_1 and the end of the latest visit period at Q_2. As the former number of customers is evidently represented by $\widetilde{M}_{2,1}(z_1, 1)$, the infinite product $\prod_{j=0}^{\infty} a_2(f_2^{(j)}(z_1))$ must be the PGF of the latter category of customers. From this it also follows that $\widetilde{G}_1(z) = \widetilde{M}_{2,1}(z, 1)\prod_{j=0}^{\infty} a_2(f_2^{(j)}(z))$.

Another way to see that the infinite product $\prod_{j=0}^{\infty} a_2(f_2^{(j)}(z_1))$ represents the number of arriving type-1 customers between the last visit period end at Q_1 and subsequently the last visit period end at Q_2 is the following. Any type-1 customer currently present (i.e., at a polling epoch of Q_1) is a customer that either arrived during a switch-over period (an ancestor) or belongs to the offspring of another type-1 or type-2 customer that arrived during a switch-over period in the past (a descendant). The currently present type-1 customers that are (descendants of) ancestors that arrived during a particular period in the past are referred to as the contribution of that period to the current polling epoch. The expression $a_2(z_1)$ (cf. (10)) now represents the complete contribution of the period that lasted until the end of the last visit to Q_2, and started at the most recent visit to Q_2 before that time that directly preceded a Q_1 visit. This period starts with a switch-over period $S_{2,1}$, of which the contribution is easily seen to be given by $\widetilde{M}_{2,1}(f_2(z_1), \widetilde{K}_{2,1}(z_1))$. After that, a geometric number of switch-over periods from Q_1 to Q_1 occur, of which the (PGF of the) contribution is given by

$$\sum_{k=0}^{\infty} r_{2,1} r_{1,1}^k \widetilde{M}_{1,1}^k(f_2(z_1), \widetilde{K}_{2,1}(z_1)) = \frac{r_{2,1}}{1 - r_{1,1}\widetilde{M}_{1,1}(f_2(z_1), \widetilde{K}_{2,1}(z_1))}.$$

Similarly, the contribution of the succeeding switch-over period $\widetilde{S}_{1,2}$ and the geometric number of switch-over periods from Q_2 to Q_2 are given by $\widetilde{M}_{1,2}(z_1, \widetilde{K}_{2,1}(z_1))$ and $\frac{r_{1,2}}{1-r_{2,2}\widetilde{M}_{2,2}(z_1,\widetilde{K}_{2,1}(z_1))}$, respectively. The product of these expressions forms $a_2(z_1) = a_2(f_2^{(0)}(z_1))$, the contribution of the latest 'inter visit-end period' of Q_2. Based on this, it is not hard to see, by the nature of $f_2(z_1)$, that $a_2(f_2^{(1)}(z_1))$ represents the contribution of the inter visit-end period preceding the latest inter visit-end period. Extending this observation, $a_2(f_2^{(j)}(z_1))$ represents the contribution of the j-th to last inter visit-end period of Q_2. As the customers currently present at Q_1 can be the contribution of any inter visit-end period of Q_2 in the past, the number sought is given by $\prod_{j=0}^{\infty} a_2(f_2^{(j)}(z_1))$, which represents the contribution of all inter visit-end periods that have past. An interpretation for $a_1(f_1^{(j)}(z_2))$ can be derived in a similar way.

Remark 3.2 Views similar to the contribution interpretation as presented in Remark 3.1 have in the past led to numerical methods for several systems, such as the descendant set approach as developed in [19] for cyclic polling systems. It is shown there that by truncating the infinite products, accurate approximations of (the PGFs of) the marginal queue length distribution arise. This supports numerical observations that the infinite-product expressions as derived in this paper give rise to efficient numerical means of computing queue length distributions.

3.2 Expressions for the waiting-time distribution and the joint queue length distribution

Based on the derived expression for the PGF $\widetilde{F}_i(z_1, z_2)$ pertaining to the queue length at a polling epoch of Q_i, we now derive $\widetilde{W}_i(s)$, the LST of the waiting-time distribution of type-i customers, and $\widetilde{L}(z_1, z_2)$, the PGF of the joint queue length at an arbitrary point in time.

3.2.1 *Analysis of* $\widetilde{W}_i(s)$

To extract an expression for $\widetilde{W}_i(s)$ from the expressions found in Sect. 3.1, we use the observation given in [33, pp. 90–91] that the analysis found in [25, Section 4.3] applied to Markovian polling systems leads to

$$\widetilde{W}_1(\lambda_1(1-z)) = \frac{\pi_1(1-\rho)(1-\widetilde{F}_1(z,1))}{\sigma\lambda_1(\widetilde{B}_1(\lambda_1(1-z))-z)} \text{ and}$$
$$\widetilde{W}_2(\lambda_2(1-z)) = \frac{\pi_2(1-\rho)(1-\widetilde{F}_2(1,z))}{\sigma\lambda_2(\widetilde{B}_2(\lambda_2(1-z))-z)}, \tag{14}$$

where σ, as defined in Sect. 2, denotes the average duration of an arbitrary switch-over period. This observation leads to expressions for $\widetilde{W}_i(s)$ as stated in the following theorem.

Theorem 3.4 *An explicit expression for* $\widetilde{W}_j(s)$ *involving converging infinite products is given by*

$$\widetilde{W}_j(s) = \frac{\pi_j(1-\rho)}{\sigma(s-\lambda_j(1-\widetilde{B}_j(s)))}\left(1-\sum_{i=1}^{2} r_{i,j}\widetilde{S}_{i,j}(s)\left(\mathbb{1}_{\{i=j\}} + \mathbb{1}_{\{i\neq j\}}\prod_{k=0}^{\infty} a_i\left(f_i^{(k)}\left(1-\frac{s}{\lambda_j}\right)\right)\right)\right). \tag{15}$$

Proof By substituting $s = \lambda_1(1-z)$ and $s = \lambda_2(1-z)$, respectively, in (14), we obtain

$$\widetilde{W}_1(s) = \frac{\pi_1(1-\rho)\left(1-\widetilde{F}_1\left(1-\frac{s}{\lambda_1},1\right)\right)}{\sigma(s-\lambda_1(1-\widetilde{B}_1(s)))} \text{ and}$$
$$\widetilde{W}_2(s) = \frac{\pi_2(1-\rho)\left(1-\widetilde{F}_2\left(1,1-\frac{s}{\lambda_2}\right)\right)}{\sigma(s-\lambda_2(1-\widetilde{B}_2(s)))}. \tag{16}$$

Combining these expressions with (13) and its equivalent for $\widetilde{F}_2(1, z_2)$ leads to the theorem. □

3.2.2 *Analysis of* $\widetilde{L}(z_1, z_2)$

To obtain $\widetilde{L}(z_1, z_2)$, we use an approach that is introduced in [8] and already applied in [13] to Markovian polling systems with an arbitrary number of queues. Before we derive a PGF of the joint queue length at an arbitrary point in time, we first regard $\widetilde{X}_i(z_1, z_2) = \mathbb{E}[z_1^{X_{i,1}} z_2^{X_{i,2}}]$, the PGF of the queue lengths $X_{i,1}$ and $X_{i,2}$ of Q_1 and Q_2 at an arbitrary point during a visit period at Q_i. By applying the results of [13, Section

3.2] to our setting, we obtain that

$$\widetilde{X}_1(z_1, z_2) = \frac{\pi_1(1-\rho)}{\rho_1\sigma} \frac{z_1(\widetilde{F}_1(z_1, z_2) - \widetilde{F}_1(\widetilde{K}_{1,2}(z_2), z_2))}{z_1 - \widetilde{B}_1(\boldsymbol{\lambda}(z))} \frac{1 - \widetilde{B}_1(\boldsymbol{\lambda}(z))}{\boldsymbol{\lambda}(z)} \tag{17}$$

and

$$\widetilde{X}_2(z_1, z_2) = \frac{\pi_2(1-\rho)}{\rho_2\sigma} \frac{z_2(\widetilde{F}_2(z_1, z_2) - \widetilde{F}_2(z_1, \widetilde{K}_{2,1}(z_1)))}{z_2 - \widetilde{B}_2(\boldsymbol{\lambda}(z))} \frac{1 - \widetilde{B}_2(\boldsymbol{\lambda}(z))}{\boldsymbol{\lambda}(z)}. \tag{18}$$

Furthermore, the results of [13, Section 3.2] reveal that $\widetilde{Y}_{i,j}(z_1, z_2) = \mathbb{E}[z_1^{Y_{i,j,1}} z_2^{Y_{i,j,2}}]$, the PGF of the queue lengths $Y_{i,j,1}$ and $Y_{i,j,2}$ of Q_1 and Q_2 at an arbitrary point during a switch-over period from Q_i to Q_j is given by

$$\widetilde{Y}_{1,j}(z_1, z_2) = \widetilde{F}_1(\widetilde{K}_{1,2}(z_2), z_2) \frac{1 - \widetilde{M}_{1,j}(z_1, z_2)}{\boldsymbol{\lambda}(z)\mathbb{E}\left[S_{1,j}\right]} \tag{19}$$

and

$$\widetilde{Y}_{2,j}(z_1, z_2) = \widetilde{F}_2(z_1, \widetilde{K}_{2,1}(z_1)) \frac{1 - \widetilde{M}_{2,j}(z_1, z_2)}{\boldsymbol{\lambda}(z)\mathbb{E}\left[S_{2,j}\right]}. \tag{20}$$

We now combine the expressions (17)–(20) into one expression for $\widetilde{L}(z_1, z_2)$, the PGF of the joint queue length at an arbitrary point in time. Observe that the server serves Q_i a fraction ρ_i of the time. In the remaining fraction $1-\rho$ of the time, the server is setting up for service at another queue. Of the time the server is in a switch-over period, he spends a fraction $\frac{r_{i,j}\pi_j\mathbb{E}[S_{i,j}]}{\sigma}$ setting up from Q_i to Q_j. Therefore, we have that

$$\widetilde{L}(z_1, z_2) = \sum_{i=1}^{2} \left(\rho_i \widetilde{X}_i(z_1, z_2) + \frac{1-\rho}{\sigma} \sum_{j=1}^{2} r_{i,j}\pi_j \mathbb{E}\left[S_{i,j}\right] \widetilde{Y}_{i,j}(z_1, z_2) \right). \tag{21}$$

This leads to the following theorem.

Theorem 3.5 *An explicit expression for $\widetilde{L}(z_1, z_2)$ involving converging infinite products is given by*

$$\widetilde{L}(z_1, z_2) = \frac{1-\rho}{\boldsymbol{\lambda}(z)\sigma} \sum_{i=1}^{2}\sum_{j=1}^{2} \pi_j \left(\frac{z_j(1 - \widetilde{B}_j(\boldsymbol{\lambda}(z)))}{z_j - \widetilde{B}_j(\boldsymbol{\lambda}(z))} (r_{i,j}\widetilde{M}_{i,j}(z_1, z_2) - \mathbb{1}_{\{i=j\}}) + r_{i,j}\left(1 - \widetilde{M}_{i,j}(z_1, z_2)\right) \right) \prod_{k=0}^{\infty} a_i(f_i^{(k)}(z_{3-i})).$$

Proof The theorem follows by combining (8), (9), Lemma 3.1 and Theorem 3.3 with (17)–(21). □

4 Heavy-traffic asymptotics

In Sect. 3, we have derived expressions for the LSTs of the waiting-time distributions and the PGF of the joint queue length distribution. These expressions are suitable for computational purposes, as theoretical and numerical evidence shows that the infinite products contained in these expressions converge very fast. However, the expressions are not in closed form, and the PGFs and the LSTs may be hard to invert symbolically. In an effort to obtain closed-form expressions for the distributions themselves, we consider the heavy-traffic asymptotics of the system, i.e., the behaviour of the system when $\rho \uparrow 1$. Recall that we study the case where the heavy-traffic limit $\rho \uparrow 1$ is taken by scaling the total arrival rate $\lambda_1 + \lambda_2$ such that the ratio $\frac{\lambda_2}{\lambda_1}$ remains fixed, so that $\frac{\hat{\lambda}_2}{\hat{\lambda}_1} = \frac{\lambda_2}{\lambda_1}$, with $\hat{\lambda}_i$ as defined in Sect. 2. In this regime, the waiting times and the queue lengths tend to infinity. Therefore, we now study the scaled waiting times $\mathcal{W}_i$ as well as the scaled queue lengths $\mathcal{L}_i$, and obtain closed-form expressions directly for their distributions. These expressions are not only easy to implement, but they also give insight into the primary effects of the model parameters on the waiting times and queue lengths, when the system operates under a heavy load. In Sect. 4.1, we derive the heavy-traffic behaviour of the waiting times and queue lengths incurred by the customers based on previous results for cyclic polling systems and some insightful observations. Subsequently, we rigorously prove these results in Sect. 4.2.

4.1 Initial study of the heavy-traffic behaviour

Before we study the heavy-traffic behaviour of the model in its full generality, we first consider the degenerate case $p_1 = p_2 = 0$ of our model. Note that for $p_1 = p_2 = 0$, the server always switches from Q_1 to Q_2 or from Q_2 to Q_1. Thus, in this particular case, the server follows a fixed alternating (or cyclic) routing mechanism. The heavy-traffic behaviour of cyclic polling models that are of a branching type and consist of an arbitrary number of queues has already been established in e.g., [21,27,28]. Translating this to our setting with two queues, exhaustive service and cyclic routing ($p_1 = p_2 = 0$), these results readily imply the following.

Proposition 4.1 *For $p_1 = p_2 = 0$, the LST of the limiting scaled waiting-time distribution is given by*

$$\lim_{\rho \uparrow 1} \widetilde{\mathcal{W}}_i(s) = \frac{1}{s(1-\hat{\rho}_i)(\mathbb{E}\left[S_{1,2}\right] + \mathbb{E}\left[S_{2,1}\right])} \left(1 - \left(\frac{{\mu_i}^{cyc}}{{\mu_i}^{cyc} + s}\right)^{\alpha^{cyc}}\right),$$

where

$$\alpha^{cyc} = \frac{2\hat{\rho}_1\hat{\rho}_2\left(\mathbb{E}\left[S_{1,2}\right] + \mathbb{E}\left[S_{2,1}\right]\right)}{\hat{\lambda}_1\mathbb{E}\left[B_1^2\right] + \hat{\lambda}_2\mathbb{E}\left[B_2^2\right]} \textit{ and } {\mu_i}^{cyc} = \frac{2\hat{\rho}_i}{\hat{\lambda}_1\mathbb{E}\left[B_1^2\right] + \hat{\lambda}_2\mathbb{E}\left[B_2^2\right]}.$$

Equivalently,

$$\lim_{\rho\uparrow 1} \mathbb{P}(\mathcal{W}_i \le t) = \mathbb{P}(UI \le t),$$

where U is a uniformly [0, 1] *distributed random variable and I is a gamma distributed random variable with shape parameter* $\alpha^{cyc} + 1$ *and scale parameter* μ_i^{cyc}, *and U and I are independent.*

The given distribution function immediately follows from inversion of the limiting LST. We observe that for the cyclic system, the complete heavy-traffic distribution of the waiting time only depends on the switch-over times through their first moments. In fact, the scaled waiting-time distribution only depends on $\mathbb{E}[S_{1,2}] + \mathbb{E}[S_{2,1}]$, the first moment of the total switch-over time incurred between two polling epochs at Q_1.

Next, we observe for the general case (i.e., $0 \le p_1, p_2 < 1$) the following. A period between two polling epochs at Q_1 can be divided in a number of subperiods:

(i) The first visit period at Q_1 after having visited Q_2;
(ii) A geometric (p_1) number of switch-over periods from Q_1 to Q_1 and subsequent 'revisit' periods at Q_1;
(iii) The switch-over period from Q_1 to Q_2;
(iv) The first visit period at Q_2 after having visited Q_1;
(v) A geometric (p_2) number of switch-over periods from Q_2 to Q_2 and subsequent 'revisit' periods at Q_2; and
(vi) The switch-over period from Q_2 to Q_1.

The support of the geometric random variables associated with subperiods (ii) and (v) is $\{0, 1, 2, \ldots\}$. In this view, we can draw a connection between the general case and the cyclic polling model as described above. In particular, we do so by reordering the subperiods as follows:

(a) All visit periods between a polling epoch at Q_1 and the first polling epoch at Q_2 to occur afterwards;
(b) A geometric (p_1) number of switch-over periods from Q_1 to Q_1;
(c) The switch-over period from Q_1 to Q_2;
(d) All visit periods between the polling epoch at Q_2 and the first polling epoch at Q_1 to occur afterwards;
(e) A geometric (p_2) number of switch-over periods from Q_2 to Q_2; and
(f) The switch-over period from Q_2 to Q_1.

Thus, the 'revisit' periods from the subperiods (ii) and (v) are shifted to the subperiods (a) and (d). In the heavy-traffic regime, the additional customers served in the subperiods (a) and (d) with respect to those in the original subperiods (i) and (iv) are negligible. This is the case since they are finite in number (they constitute arrivals during finitely long switch-over times), whereas the customers served in the original subperiods are infinite in number. As a result, the limiting waiting-time distribution of the customers served in the periods (a) and (d) coincides with that of the customers served in the reordered subperiods (i) and (iv), respectively. Note that in this reordered scheme, the polling system can be interpreted as a cyclic model, as the subperiods (b) and (c) together form a switch-over period from Q_1 to Q_2, and the subperiods (e) and (f) together form a switch-over period from Q_2 to Q_1. The switch-over period from Q_1 to Q_2 in this cyclic equivalent then consists of a geometric (p_1) number of original switch-over periods from Q_1 to Q_1 and an original switch-over period from Q_1 to Q_2 of the Markovian model. Similarly, the switch-over period from Q_2 to Q_1 in the cyclic equivalent consists of a geometric (p_2) number of switch-over periods from Q_2 to Q_2 and a subsequent switch-over period from Q_2 to Q_1.

Finally, we observe that the first moment of the total switch-over time incurred between two polling epochs at Q_1 in the Markovian polling model is given by

$$\mathbb{E}\left[S^{tot}\right] = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\left(i\mathbb{E}\left[S_{1,1}\right]+\mathbb{E}\left[S_{1,2}\right]+\mathbb{E}\left[S_{2,1}\right]+j\mathbb{E}\left[S_{2,2}\right]\right)(1-p_1)p_1^i(1-p_2)p_2^j$$
$$= \frac{p_1}{1-p_1}\mathbb{E}\left[S_{1,1}\right] + \mathbb{E}\left[S_{1,2}\right] + \mathbb{E}\left[S_{2,1}\right] + \frac{p_2}{1-p_2}\mathbb{E}\left[S_{2,2}\right]. \tag{22}$$

Combining all of the observations above, it is reasonable to assume that the heavy-traffic behaviour of the general case is similar to the heavy-traffic behaviour as derived in Proposition 4.1 for the cyclic case, except that the term $\mathbb{E}[S_{1,2}] + \mathbb{E}[S_{2,1}]$ should be replaced by $\mathbb{E}[S^{tot}]$. We formulate this result below; a rigorous proof will be given in Sect. 4.2.

Theorem 4.2 *For $0 \le p_1, p_2 < 1$, the LST of the limiting scaled waiting-time distribution is given by*

$$\lim_{\rho\uparrow 1} \widetilde{\mathcal{W}}_i(s) = \frac{1}{s(1-\hat{\rho}_i)\mathbb{E}\left[S^{tot}\right]}\left(1-\left(\frac{\mu_i}{\mu_i+s}\right)^{\alpha}\right), \tag{23}$$

where

$$\alpha = \frac{2\hat{\rho}_1\hat{\rho}_2\mathbb{E}\left[S^{tot}\right]}{\hat{\lambda}_1\mathbb{E}\left[B_1^2\right]+\hat{\lambda}_2\mathbb{E}\left[B_2^2\right]}, \mu_i = \frac{2\hat{\rho}_i}{\hat{\lambda}_1\mathbb{E}\left[B_1^2\right]+\hat{\lambda}_2\mathbb{E}\left[B_2^2\right]} \tag{24}$$

and $\mathbb{E}[S^{tot}]$ is given in (22). Equivalently,

$$\lim_{\rho\uparrow 1}\mathbb{P}(\mathcal{W}_i \le t) = \mathbb{P}(UI \le t), \tag{25}$$

where U is a uniformly $[0, 1]$ distributed random variable, I is a gamma distributed random variable with shape parameter $\alpha + 1$ and scale parameter μ_i, and U and I are independent.

Based on this theorem concerning the scaled waiting-time distribution, we can also derive the heavy-traffic behaviour of the scaled queue length distribution. From Little's law, it is immediate that $\mathbb{E}[\mathcal{L}_i] = \hat{\lambda}_i\mathbb{E}[\mathcal{W}_i]$. Furthermore, in many queueing models under heavy-traffic conditions, the scaled virtual waiting time processes and queue length processes exhibit so-called state-space collapse: the one process is in heavy traffic essentially the same as the other process multiplied by a scalar constant (cf. [22]). It is thus reasonable to assume that in heavy traffic the distribution of $\mathcal{L}_i$ equals the distribution of $\mathcal{W}_i$ scaled by a factor $\hat{\lambda}_i$. This leads to the following statement, for which, again, a rigorous proof will be given in Sect. 4.2.

Theorem 4.3 *For* $0 \leq p_1, p_2 < 1$, *the limiting scaled marginal queue length distribution is given by*

$$\lim_{\rho \uparrow 1} \mathbb{P}(\mathcal{L}_i \leq t) = \mathbb{P}(UI \leq t),$$

where U *is a uniformly* $[0, 1]$ *distributed random variable,* I *is a gamma distributed random variable with shape parameter* $\alpha + 1$ *and scale parameter* $\frac{\mu_i}{\hat{\lambda}_i}$ *(*α *and* μ_i *as defined in* (24)*). Furthermore, the random variables* U *and* I *are independent.*

Remark 4.1 Besides the distribution of a uniform times a gamma random variable, the limiting distribution of $(1-\rho)W_i$ as given in Theorem 4.2 can also be interpreted as the residual (overshoot) of a gamma distribution. To see this, observe that (23) can be rewritten as

$$\lim_{\rho \uparrow 1} \widetilde{\mathcal{W}}_i(s) = \frac{1 - \left(\frac{\mu_i}{\mu_i + s}\right)^{\alpha}}{s\frac{\alpha}{\mu_i}}.$$

As $(\frac{\mu_i}{\mu_i+s})^{\alpha}$ is the LST of a gamma (α, μ_i) distribution with first moment $\frac{\alpha}{\mu_i}$, the limiting distribution constitutes the residual of a gamma distribution with shape parameter α and scale parameter μ_i. A similar observation holds for the limiting distribution of $(1-\rho)W_i$ in the cyclic case as provided in Proposition 4.1.

Remark 4.2 Theorems 4.2 and 4.3 can immediately be used as approximations for the marginal waiting-time distributions and queue length distributions in stable systems with a load $\bar{\rho} < 1$:

$$\mathbb{P}\left(W_i < \frac{t}{1-\bar{\rho}}\right) \approx \lim_{\rho \uparrow 1} \mathbb{P}(\mathcal{W}_i < t) \text{ and } \mathbb{P}\left(L_i < \frac{x}{1-\bar{\rho}}\right) \approx \lim_{\rho \uparrow 1} \mathbb{P}(\mathcal{L}_i < x).$$

As shown in [21], approximations of this type are reasonably accurate for heavily loaded polling models (i.e., a load close to one). This is not surprising, as the approximation becomes exact by construction as $\bar{\rho}$ tends to one. Moreover, it is interesting to note that the limiting distributions of the scaled waiting times and queue lengths only depend on the first two moments of the service-time distribution as well as the first moment of the total switch-over time between two polling epochs at Q_1. They do not require higher moments, and are thus useful for practical purposes, as in reality, information about third- and higher-order moments is often hard to get.

4.2 Proofs of Theorems 4.2 and 4.3

In this section, we prove Theorems 4.2 and 4.3. For the former theorem, we rely in part on the results found in [21]. That paper provides an analysis of the heavy-traffic behaviour of periodic polling systems, of which the marginal queue length distribution at polling epochs can be (numerically) computed by the descendant set approach (cf. [19]), by analysing the mechanics of this technique in the heavy-traffic regime. The results that we particularly rely on are [21, Theorems 3 and 4], which give the limiting behaviour of a marginal queue length H of Q_1 observed at predefined epochs in time, of which the PGF $\widetilde{H}(z) = \mathbb{E}[z^H]$ can be written as

$$\widetilde{H}(z) = \prod_{c=0}^{\infty} \widetilde{R}_1\left(\lambda_1(1-\widetilde{A}_{1,c-1}(z)) + \lambda_2(1-\widetilde{A}_{2,c}(z))\right) \widetilde{R}_2\left(\lambda_1(1-\widetilde{A}_{1,c-1}(z)) + \lambda_2(1-\widetilde{A}_{2,c-1}(z))\right), \tag{26}$$

where $\widetilde{R}_1(s)$ and $\widetilde{R}_2(s)$ are LSTs of two arbitrary positive random variables R_1 and R_2,

$$\begin{aligned} \widetilde{A}_{1,c}(z) &= \widetilde{P}_1(\lambda_2(1-\widetilde{A}_{2,c}(z))) = \widetilde{K}_{1,2}(\widetilde{A}_{2,c}(z)), & \widetilde{A}_{1,-1}(z) &= z, \\ \widetilde{A}_{2,c}(z) &= \widetilde{P}_2(\lambda_1(1-\widetilde{A}_{1,c-1}(z))) = \widetilde{K}_{2,1}(\widetilde{A}_{1,c-1}(z)), & \widetilde{A}_{2,-1}(z) &= 1, \end{aligned} \tag{27}$$

and $\widetilde{P}_i(s)$ as defined in Sect. 2. The results of [21] state that under these conditions, $(1-\rho)H$ converges in distribution, as $\rho \uparrow 1$, to a gamma distribution with shape parameter $\frac{2\hat{\rho}_1\hat{\rho}_2(\mathbb{E}[R_1]+\mathbb{E}[R_2])}{\hat{\lambda}_1\mathbb{E}[B_1^2]+\hat{\lambda}_2\mathbb{E}[B_2^2]}$ and scale parameter $\frac{2\hat{\rho}_i}{\hat{\lambda}_i(\hat{\lambda}_1\mathbb{E}[B_1^2]+\hat{\lambda}_2\mathbb{E}[B_2^2])}$. Furthermore, it is stated that $\lim_{\rho\uparrow 1}\mathbb{E}[(1-\rho)^k H^k]$ coincides with the k-th moment of this distribution.

We have now only stated the results of [21] applied to two-queue polling systems with alternating and exhaustive service. A more general statement for polling systems with a general number of queues and periodic routing is shown to hold in [21] by exploiting several useful observations based on the descendant set approach.

As noted in Remark 3.2, however, the expressions that we obtained for the PGF of the queue length distribution in Sect. 3 allow for an interpretation in the spirit of the descendant set approach. As such, the results of [21] as stated above almost directly lead to the following lemma pertaining to G_i, the number of type-i customers in the system at a polling epoch of Q_i that follows a visit period at Q_{3-i}.

Lemma 4.4 *The distribution of $(1-\rho)G_i$ converges, as $\rho \uparrow 1$, in distribution to a gamma distribution with shape parameter α and scale parameter $\frac{\mu_i}{\hat{\lambda}_i}$, where α and μ_i are defined in (24). Furthermore, we have that $\lim_{\rho\uparrow 1}\mathbb{E}[(1-\rho)^k G_i^k]$ coincides with the k-th moment of this distribution.*

Proof We focus on the limiting distribution of $(1-\rho)G_1$. In Remark 3.1, we already concluded that $\widetilde{G}_1(z) = \widetilde{M}_{2,1}(z,1)\prod_{j=0}^{\infty} a_2(f_2^{(j)}(z))$. With some effort, it is straightforward to see that alternatively this can be written as

$$\widetilde{G}_1(z) = \widetilde{H}(z)\frac{1-r_{1,1}\widetilde{M}_{1,1}(z,1)}{r_{2,1}\widetilde{M}_{2,1}(z,1)}, \tag{28}$$

with $\widetilde{H}(z)$ as in (26), where

$$\widetilde{R}_j(s) = \widetilde{S}_{j,3-j}(s)\frac{r_{j,3-j}}{1-r_{3-j,3-j}\widetilde{S}_{3-j,3-j}(s)},$$

i.e., R_j is chosen to be the convolution of a switch-over time from Q_j to Q_{3-j} and a geometric $(r_{3-j,3-j})$ number of switch-over times from Q_{3-j} to Q_{3-j}. From this definition, it is easily verified that $\mathbb{E}[R_1]+\mathbb{E}[R_2]=\mathbb{E}[S^{tot}]$. As $\lim_{\rho\uparrow 1}\widetilde{M}_{1,1}(z^{1-\rho},1)=\lim_{\rho\uparrow 1}\widetilde{M}_{1,2}(z^{1-\rho},1)=1$, it is clear by (28) that the PGF of the scaled distribution $\widetilde{G}_1(z^{1-\rho})=\mathbb{E}[z^{(1-\rho)G_1}]$ satisfies $\lim_{\rho\uparrow 1}\widetilde{G}_1(z^{1-\rho})=\lim_{\rho\uparrow 1}\widetilde{H}(z^{1-\rho})$. Thus, the distributions of the scaled versions of G_1 and H coincide in the heavy-traffic limit. For $i=1$, the lemma now follows from the results of [21] as described above. For $i=2$, the lemma follows by interchanging indices. □

Now that we have established the heavy-traffic behaviour of G_i, we are able to prove Theorem 4.2 by making use of (16).

Proof of Theorem 4.2 Again, we focus on the case $i=1$ with the understanding that the proof for the case $i=2$ follows by interchanging indices. By (13) and (16), we have that

$$\begin{aligned}\lim_{\rho\uparrow 1}\widetilde{\mathcal{W}}_1((1-\rho)s)&=\lim_{\rho\uparrow 1}\frac{\pi_1(1-\rho)}{\sigma((1-\rho)s-\lambda_1(1-\widetilde{B}_1((1-\rho)s)))}\\&\quad\times\lim_{\rho\uparrow 1}\left(1-r_{1,1}\widetilde{M}_{1,1}\left(1-\frac{(1-\rho)s}{\lambda_1},1\right)\right.\\&\quad\left.-r_{2,1}\widetilde{G}_1\left(1-\frac{(1-\rho)s}{\lambda_1}\right)\right).\end{aligned}\tag{29}$$

By applying L'Hôpital's rule and observing that $\frac{\pi_1}{\sigma}=(r_{2,1}\mathbb{E}[S^{\text{tot}}])^{-1}$, we obtain that

$$\begin{aligned}\lim_{\rho\uparrow 1}\frac{\pi_1(1-\rho)}{\sigma((1-\rho)s-\lambda_1(1-\widetilde{B}_1((1-\rho)s)))}&=\lim_{\rho\uparrow 1}\frac{-\pi_1}{\sigma s(-1+\lambda_1\mathbb{E}\left[B_1e^{-(1-\rho)sB_1}\right])}\\&=\frac{1}{r_{2,1}s(1-\hat{\rho}_1)\mathbb{E}[S^{\text{tot}}]}.\end{aligned}$$

Furthermore, it is clear that $\lim_{\rho\uparrow 1}\widetilde{M}_{1,1}\left(1-\frac{(1-\rho)s}{\lambda_1},1\right)=1$. Deriving $\lim_{\rho\uparrow 1}\widetilde{G}_1(1-\frac{(1-\rho)s}{\lambda_1})$, however, takes a bit more effort. By invoking a Taylor expansion in G_1, we have that

$$\begin{aligned}\lim_{\rho\uparrow 1}\widetilde{G}_1\left(1-\frac{(1-\rho)s}{\lambda_1}\right)&=\lim_{\rho\uparrow 1}\mathbb{E}\left[\left(1-\frac{(1-\rho)s}{\lambda_1}\right)^{G_1}\right]\\&=\lim_{\rho\uparrow 1}\mathbb{E}\left[\sum_{k=0}^{\infty}\frac{\log^k\left(1-\frac{(1-\rho)s}{\lambda_1}\right)G_1^k}{k!}\right].\end{aligned}$$

To further reduce this expression, observe that a Taylor expansion around $\rho = 1$ yields $\log(1-(1-\rho)c) = -\sum_{j=1}^{\infty} \frac{(1-\rho)^j c^j}{j}$ for any $c \in \mathbb{R}$. As such,

$$\lim_{\rho\uparrow 1} \widetilde{G}_1\left(1 - \frac{(1-\rho)s}{\lambda_1}\right) = \lim_{\rho\uparrow 1} \mathbb{E}\left[\sum_{k=0}^{\infty} \frac{(-1)^k \left(\sum_{j=1}^{\infty}(1-\rho)^j s^j \lambda_1^{-j}/j\right)^k G_1^k}{k!}\right]. \tag{30}$$

Note, however, that due to Lemma 4.4, we have for any $j > k$ that $\lim_{\rho\uparrow 1} \mathbb{E}[(1-\rho)^j G_1^k] = \lim_{\rho\uparrow 1}(1-\rho)^{j-k} \lim_{\rho\uparrow 1} \mathbb{E}[(1-\rho)^k G_1^k] = 0$. Therefore, second and higher order terms of the inner sum of (30) disappear in the limit, so that the expression as a whole reduces to

$$\begin{aligned}\lim_{\rho\uparrow 1} \widetilde{G}_1\left(1 - \frac{(1-\rho)s}{\lambda_1}\right) &= \lim_{\rho\uparrow 1} \mathbb{E}\left[\sum_{k=0}^{\infty} \frac{(-1)^k (1-\rho)^k s^k \lambda_1^{-k} G_1^k}{k!}\right] \\ &= \lim_{\rho\uparrow 1} \mathbb{E}\left[e^{-(1-\rho)\frac{s}{\lambda_1} G_1}\right] = \left(\frac{\mu_1}{\mu_1 + s}\right)^{\alpha},\end{aligned}$$

where the last equality follows from Lemma 4.4. By combining the limits found above, we can reduce (29) to

$$\lim_{\rho\uparrow 1} \widetilde{\mathcal{W}}_1(s) = \frac{1}{r_{2,1} s (1-\hat{\rho}_1)\mathbb{E}\left[S^{tot}\right]}\left(1 - r_{1,1} - r_{2,1}\left(\frac{\mu_1}{\mu_1 + s}\right)^{\alpha}\right),$$

which is equivalent to (23). Equation (25) then follows by inversion of the LST. □

Now that Theorem 4.2 is proved, Theorem 4.3 follows almost immediately by the proof below.

Proof of Theorem 4.3 We make use of the distributional form of Little's law (cf. [17]), which states that

$$\widetilde{L}_i(z) = \widetilde{W}_i(\lambda_i(1-z))\widetilde{B}_i(\lambda_i(1-z)).$$

As such, we have that

$$\begin{aligned}\lim_{\rho\uparrow 1} \widetilde{\mathcal{L}}_i(z) = \lim_{\rho\uparrow 1} \widetilde{L}_i(z^{1-\rho}) &= \lim_{\rho\uparrow 1} \widetilde{W}_i(\lambda_i(1-z^{1-\rho}))\widetilde{B}_i(\lambda_i(1-z^{1-\rho})) \\ &= \lim_{\rho\uparrow 1} \widetilde{\mathcal{W}}_i\left(\frac{\lambda_i(1-z^{1-\rho})}{1-\rho}\right).\end{aligned} \tag{31}$$

As $\lim_{\rho\uparrow 1} \frac{\lambda_i(1-z^{1-\rho})}{1-\rho} = -\hat{\lambda}_i \log(z)$, a combination of Theorem 4.2 and (31) now implies that

$$\lim_{\rho\uparrow 1} \widetilde{\mathcal{L}}_i(z) = \frac{1}{-\hat{\lambda}_i \log(z)(1-\hat{\rho}_i)\mathbb{E}\left[S^{tot}\right]}\left(1 - \left(\frac{\mu_i}{\mu_i - \hat{\lambda}_i \log(z)}\right)^{\alpha}\right).$$

The latter expression is the PGF of the distribution mentioned in the theorem and this concludes the proof. □

Remark 4.3 The striking similarity between the heavy-traffic asymptotics of cyclic polling systems and those of the class of systems that we consider may in part be explained by the following. Despite the fact that Markovian polling systems generally do not satisfy the branching property as introduced in Sect. 1, the subset of two-queue exhaustive models does actually satisfy this property. More specifically, in the model that we consider in this paper, the joint queue length process observed at Q_i polling epochs constitutes a multi-type branching process with immigration (see e.g., [3]). As a consequence, this model fits in the framework considered in [28], and Lemma 4.4 follows alternatively from [28, Theorem 5] by taking the particle offspring functions $f^{(i)}(z_1, z_2)$ and the immigration function $g(z_1, z_2)$ as introduced in [28, Equations (3) and (4)] equal to $f^{(1)}(z_1, z_2) = \widetilde{K}_{1,2}(\widetilde{K}_{2,1}(z_1))$, $f^{(2)}(z_1, z_2) = \widetilde{K}_{2,1}(z_1)$ and $g(z_1, z_2) = a_2(z_1)\frac{\widetilde{M}_{2,1}(z_1,z_2)}{\widetilde{M}_{2,1}(f_2(z_1),\widetilde{K}_{2,1}(z_1))}$.

5 Conclusions and topics for further research

In this paper, we have obtained expressions for (the LSTs of) the waiting-time distributions of type-i customers and (the PGF of) the joint queue length distribution for two-queue Markovian polling systems with exhaustive service. Although these expressions are of independent interest and are suitable for implementation purposes, we have also used these expressions as a basis to obtain the heavy-traffic behaviour of the system. The established heavy-traffic asymptotics provide insights into the key effects of the model parameters when the system is heavily loaded and turn out to be very similar to the heavy-traffic asymptotics of cyclic polling models. This analysis provides closed-form heavy-traffic approximations directly for the distribution functions of the waiting times and the queue lengths.

The results obtained give rise to a variety of directions for further research. These avenues of further research include the study of the model with more than two queues. Although an equivalent of Theorem 3.3 seems hard to find for this case, functional equations similar to (2) and (3) exist for a larger number of queues. A heavy-traffic analysis may be found by carefully inspecting the behaviour of this functional equation under heavy-traffic scalings.

Another assumption that one might wish to relax is the assumption of exhaustive service at both queues. Although an analysis in the spirit of Sect. 3 also seems hard to perform when steering away from the exhaustive assumption, preliminary investigations of the authors suggest that the heavy-traffic limits of the waiting times and queue lengths still allow for compact and closed-form expressions. For instance, in the case of two-queue Markovian models with gated service (where, during a visit period, the server only serves the customers that were present at the start of it), the heavy-traffic limits seem to coincide with the heavy-traffic limits of a cyclic polling model in a similar way as established in this paper for the exhaustive case. The service discipline of this cyclic model, however, amounts to the κ-gated discipline as introduced in [30], but where κ is a geometric random variable rather than a constant. As this ‘geometric gated' service discipline defies the branching property, heavy-traffic asymptotics for the cyclic equivalent are not readily available in the literature, and thus require more study.

A final suggestion for further research is the refinement of the closed-form approximations as given in Remark 4.2. These approximations perform very well for heavily loaded models due to their exact behaviour in the heavy-traffic limit, but their performance degrades when the load offered to the system is only moderate. To this end, one may consider to construct approximations by interpolating between the found heavy-traffic asymptotics and light-traffic behaviour based on the actual offered load in the spirit of [6, 14].

Acknowledgments The authors wish to thank Marko Boon, Sem Borst and Maria Vlasiou for valuable comments on earlier drafts of the present paper. Funded in the framework of the STAR-project "Multilayered queueing systems" by the Netherlands Organization for Scientific Research (NWO). The research of Onno J. Boxma is performed in the IAP Bestcom project, funded by the Belgian government.

Appendix 1: Proof of Lemma 3.1

Proof We first focus on the value of $\left|1 - f_1^{(\infty)}(z_2)\right| = \lim_{j\to\infty}\left|1 - f_1^{(j)}(z_2)\right|$. For arbitrary $j > 0$, we have for any z_2 in the unit circle that

$$\begin{aligned}\left|1 - f_1^{(j)}(z_2)\right| &= \left|1 - f_1(f_1^{j-1}(z_2))\right| \\ &= \left|\int_{t=0}^{\infty}(1 - e^{-\lambda_1(1-\widetilde{K}_{1,2}(f_1^{(j-1)}(z_2)))t})\mathrm{d}\mathbb{P}(P_2 < t)\right| \\ &\le \int_{t=0}^{\infty}\left|1 - e^{-\lambda_1(1-\widetilde{K}_{1,2}(f_1^{(j-1)}(z_2)))t}\right|\mathrm{d}\mathbb{P}(P_2 < t),\end{aligned}$$

where the inequality constitutes the triangle inequality. Note that $\left|1 - e^{-x}\right| \le |x|$ for any $x \in \{z \in \mathbb{C} : \Re(z) > 0\}$, so that

$$\begin{aligned}\left|1 - f_1^{(j)}(z_2)\right| &\le \int_{t=0}^{\infty}\lambda_1 t\left|1 - \widetilde{K}_{1,2}(f_1^{(j-1)}(z_2))\right|\mathrm{d}\mathbb{P}(P_2 < t) \\ &= \lambda_1\mathbb{E}[P_2]\left|1 - \widetilde{K}_{1,2}(f_1^{(j-1)}(z_2))\right| \\ &\le \lambda_1\mathbb{E}[P_2]\left|\int_{t=0}^{\infty}(1 - e^{-\lambda_2(1-f_1^{(j-1)}(z_2))t})\mathrm{d}\mathbb{P}(P_1 < t)\right| \\ &\le \lambda_1\mathbb{E}[P_2]\lambda_2\mathbb{E}[P_1]\left|1 - f_1^{(j-1)}(z_2)\right|. \end{aligned} \tag{32}$$

Iteration of (32) leads to

$$\left|1 - f_1^{(j)}(z_2)\right| \leq (\lambda_1 \mathbb{E}[P_2]\, \lambda_2 \mathbb{E}[P_1])^j \, |1 - z_2| \, . \tag{33}$$

By (1) we have that $\mathbb{E}[P_i] = \mathbb{E}[B_i](1-\rho_i)^{-1}$, so that

$$\lambda_1 \mathbb{E}[P_2]\, \lambda_2 \mathbb{E}[P_1] = \frac{\rho_1}{1-\rho_2} \frac{\rho_2}{1-\rho_1} < 1. \tag{34}$$

The inequality follows since the queues are assumed to be stable, i.e., $0 \leq \rho < 1$. Therefore, $\rho_1 = \rho - \rho_2 < 1 - \rho_2$, and similarly $\rho_2 < 1 - \rho_1$. A combination of (32) and (34) now leads to

$$0 \leq \lim_{j\to\infty} \left|1 - f_1^{(j)}(z_2)\right| \leq \lim_{j\to\infty} (\lambda_1 \mathbb{E}[P_2]\, \lambda_2 \mathbb{E}[P_1])^j \, |1 - z_2| = 0.$$

Since $\lim_{j\to\infty} \left|1 - f_1^{(j)}(z_2)\right| = 0$, we must have that $f_1^{(\infty)}(z_2) = \lim_{j\to\infty} f_1^{(j)}(z_2) = 1$.

By similar arguments, it can be shown that $f_2^{(\infty)}(z_1) = 1$ for any z_1 in the unit circle. Finally, it is evident that $\widetilde{K}_{1,2}(1) = \widetilde{K}_{2,1}(1) = \widetilde{F}_1(1,1) = \widetilde{F}_2(1,1) = 1$. The lemma now follows. □

Appendix 2: Proof of Lemma 3.2

Proof We initially focus on the product $\prod_{j=0}^{\infty} a_1(f_1^{(j)}(z_2))$. By the theory of infinite products (see e.g., [26, Chapter 1]), we have that $\prod_{j=0}^{\infty} a_1(f_1^{(j)}(z_2))$ converges iff $\sum_{j=0}^{\infty}(1 - a_1(f_1^{(j)}(z_2)))$ converges. To establish the latter, it is enough to prove that $\sum_{j=0}^{\infty} \left|1 - a_1(f_1^{(j)}(z_2))\right|$ converges. We observe that

$$\begin{aligned}
&\left|1 - a_1(f_1^{(j)}(z_2))\right| \\
&= \left|1 - \frac{r_{2,1}\widetilde{M}_{2,1}(\widetilde{K}_{1,2}(f_1^{(j)}(z_2)), f_1^{(j)}(z_2))}{1 - r_{1,1}\widetilde{M}_{1,1}(\widetilde{K}_{1,2}(f_1^{(j)}(z_2)), f_1^{(j)}(z_2))} \frac{r_{1,2}\widetilde{M}_{1,2}(\widetilde{K}_{1,2}(f_1^{(j)}(z_2)), f_1^{(j)}(z_2))}{1 - r_{2,2}\widetilde{M}_{2,2}(\widetilde{K}_{1,2}(f_1^{(j)}(z_2)), f_1^{(j)}(z_2))}\right| \\
&= \left|\frac{\sum_{i=1}^{2} A_{1,i}(f_1^{(j)}(z_2))(1 - \widetilde{M}_{i,1}(\widetilde{K}_{1,2}(f_1^{(j)}(z_2)), f_1^{(j)}(z_2)))}{D(z_2)}\right. \\
&\quad \left. + \frac{\sum_{i=1}^{2} A_{2,i}(f_1^{(j)}(z_2))(1 - \widetilde{M}_{i,2}(\widetilde{K}_{1,2}(f_1^{(j)}(z_2)), f_1^{(j)}(z_2)))}{D(z_2)}\right|,
\end{aligned} \tag{35}$$

where

$$\begin{aligned}
A_{1,1}(z_2) &= r_{1,1}(1 - r_{2,2}), \\
A_{1,2}(z_2) &= (1 - r_{1,1})(1 - r_{2,2}), \\
A_{2,1}(z_2) &= (1 - r_{1,1})(1 - r_{2,2})\widetilde{M}_{1,2}(\widetilde{K}_{1,2}(z_2), z_2), \\
A_{2,2}(z_2) &= r_{2,2}(1 - r_{1,1}\widetilde{M}_{1,1}(\widetilde{K}_{1,2}(z_2), z_2)) \text{ and} \\
D(z_2) &= (1 - r_{1,1}\widetilde{M}_{1,1}(\widetilde{K}_{1,2}(f_1^{(j)}(z_2)), f_1^{(j)}(z_2))) \\
&\quad \times (1 - r_{2,2}\widetilde{M}_{2,2}(\widetilde{K}_{1,2}(f_1^{(j)}(z_2)), f_1^{(j)}(z_2))).
\end{aligned}$$

Using the triangle inequality and similar arguments as those in the proof of Lemma 3.1, we note that for $1 \leq i, k \leq 2$ and $j > 0$,

$$\begin{aligned}
&|1 - \widetilde{M}_{i,k}(\widetilde{K}_{1,2}(f_1^{(j)}(z_2)), f_1^{(j)}(z_2))| \\
&\leq \int_{t=0}^{\infty} \left|1 - e^{-(\lambda_1(1-\widetilde{K}_{1,2}(f_1^{(j)}(z_2)))+\lambda_2(1-f_1^{(j)}(z_2)))t}\right| \mathrm{d}\mathbb{P}(S_{i,k} < t) \\
&\leq \mathbb{E}\left[S_{i,k}\right]\left(\lambda_1\left|1 - \widetilde{K}_{1,2}(f_1^{(j)}(z_2))\right| + \lambda_2\left|1 - f_1^{(j)}(z_2)\right|\right) \\
&\leq \mathbb{E}\left[S_{i,k}\right]\lambda_2(\lambda_1\mathbb{E}\left[P_1\right] + 1)\left|1 - f_1^{(j)}(z_2)\right|.
\end{aligned}$$

Moreover, it is trivially seen that $\left|A_{i,k}(z_2)\right| \leq 1$ for $1 \leq i, k \leq 2$ and any z_2 in the unit circle. Furthermore, since $\left|\widetilde{M}_{i,k}(\widetilde{K}_{1,2}(z_2), z_2)\right| \leq 1$, we have that $|D(z_2)| \geq (1 - r_{1,1})(1 - r_{2,2})$. Therefore, a combination of (33) and (35) with the triangle inequality leads to

$$\begin{aligned}
\left|1 - a_1(f_1^{(j)}(z_2))\right| \leq\ & \frac{\mathbb{E}\left[S_{1,1}\right] + \mathbb{E}\left[S_{1,2}\right] + \mathbb{E}\left[S_{2,1}\right] + \mathbb{E}\left[S_{2,2}\right]}{(1 - r_{1,1})(1 - r_{2,2})} \\
& \times \lambda_2(\lambda_1\mathbb{E}\left[P_1\right] + 1)\left(\lambda_1\mathbb{E}\left[P_2\right]\lambda_2\mathbb{E}\left[P_1\right]\right)^j |1 - z_2|
\end{aligned}$$

This result obviously shows, in combination with (34), that $\sum_{j=0}^{\infty}\left|1 - a_1(f_1^{(j)}(z_2))\right|$ is bounded from above by a converging geometric sum. As such, $\sum_{j=0}^{\infty}\left|1 - a_1(f_1^{(j)}(z_2))\right|$ converges, so that $\prod_{j=0}^{\infty} a_1(f_1^{(j)}(z_2))$ converges. The convergence of the product $\prod_{j=0}^{\infty} a_2(f_2^{(j)}(z_1))$ can be established similarly. □

References

1. Abate, J., Whitt, W.: Numerical inversion of probability generating functions. Oper. Res. Lett. **12**, 245–251 (1992)
2. Abate, J., Whitt, W.: Numerical inversion of Laplace transforms of probability distributions. ORSA J. Comput. **7**, 36–43 (1995)
3. Athreya, K.B., Ney, P.E.: Branching Processes. Springer, New York (1972)

4. Boon, M.A.A., van der Mei, R.D., Winands, E.M.M.: Applications of polling systems. Surv. Oper. Res. Manag. Sci. **16**, 67–82 (2011)
5. Boon, M.A.A., Winands, E.M.M.: Heavy-traffic analysis of k-limited polling systems. Probab. Eng. Inf. Sci. (2014, to appear)
6. Boon, M.A.A., Winands, E.M.M., Adan, I.J.B.F., van Wijk, A.C.C.: Closed-form waiting time approximations for polling systems. Perform. Eval. **68**, 290–306 (2011)
7. Boxma, O.J., Groenendijk, W.P.: Two queues with alternating service and switching times. In: Boxma, O.J., Syski, R. (eds.) Queueing Theory and its Applications (Liber Amicorum for J. W. Cohen), pp. 261–282. North-Holland, Amsterdam (1988)
8. Boxma, O.J., Kella, O., Kosiński, K.M.: Queue lengths and workloads in polling systems. Oper. Res. Lett. **39**, 401–405 (2011)
9. Boxma, O.J., Weststrate, J.: Waiting times in polling systems with Markovian server routing. In: Stiege, G., Lie, J.S. (eds.) Messung, Modellierung und Bewertung von Rechensystemen und Netzen, pp. 89–104. Springer, Berlin (1989)
10. Coffman, E.G., Puhalskii, A.A., Reiman, M.I.: Polling systems with zero switch-over times: a heavy-traffic principle. Ann. Appl. Probab. **5**, 681–719 (1995)
11. Coffman, E.G., Puhalskii, A.A., Reiman, M.I.: Polling systems in heavy-traffic: a Bessel process limit. Math. Oper. Res. **23**, 257–304 (1998)
12. den Iseger, P.: Numerical transform inversion using Gaussian quadrature. Probab. Eng. Inf. Sci. **20**, 1–44 (2006)
13. Dorsman, J.L., Borst, S.C., Boxma, O.J., Vlasiou, M.: Markovian polling systems with an application to wireless random-access networks. Technical Report 2014–001, Eurandom Preprint Series, 2014. http://www.eurandom.tue.nl/reports/ (2014)
14. Dorsman, J.L., van der Mei, R.D., Winands, E.M.M.: A new method for deriving waiting-time approximations in polling systems with renewal arrivals. Stoch. Models **27**, 318–332 (2011)
15. Grossglauser, M., Tse, D.: Mobility increases the capacity of ad hoc wireless networks. IEEE/ACM Trans. Netw. **10**, 477–486 (2002)
16. Holma, H., Toskala, A.: HSDPA/HSUPA for UMTS: High Speed Radio Access for Mobile Communications. Wiley, Hoboken, NJ (2006)
17. Keilson, J., Servi, L.D.: The distributional form of Little's law and the Fuhrmann–Cooper decomposition. Oper. Res. Lett. **9**, 239–247 (1990)
18. Kleinrock, L., Levy, H.: The analysis of random polling systems. Oper. Res. **36**, 716–732 (1988)
19. 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)
20. Levy, H., Sidi, M.: Polling systems: application, modeling and optimization. IEEE Trans. Commun. **38**, 1750–1760 (1990)
21. Olsen, T.L., van der Mei, R.D.: Polling systems with periodic server routeing in heavy traffic: distribution of the delay. J. Appl. Probab. **40**, 305–326 (2003)
22. Reiman, M.I.: Some diffusion approximations with state space collapse. In: Modelling and Performance Evaluation Methodology (Paris. 1983), Lecture Notes in Control and Information Sciences, pp. 209–240. Springer, Berlin (1984)
23. Resing, J.A.C.: Polling systems and multitype branching processes. Queueing Syst. **13**, 409–426 (1993)
24. Srinivasan, M.M.: Nondeterministic polling systems. Manag. Sci. **37**, 667–681 (1991)
25. Takagi, H.: Analysis of Polling Systems. MIT Press, Cambridge, MA (1986)
26. Titchmarsh, E.C.: Theory of Functions. Oxford University Press, London (1939)
27. van der Mei, R.D.: Distribution of the delay in polling systems in heavy traffic. Perform. Eval. **38**, 133–148 (1999)
28. van der Mei, R.D.: Towards a unifying theory on branching-type polling systems in heavy traffic. Queueing Syst. **57**, 29–46 (2007)
29. van der Mei, R.D., Winands, E.M.M.: Heavy traffic analysis of polling models by mean value analysis. Perform. Eval. **65**, 400–416 (2008)
30. van Wijk, A.C.C., Adan, I.J.B.F., Boxma, O.J., Wierman, A.: Fairness and efficiency for polling models with the κ-gated service discipline. Perform. Eval. **69**, 274–288 (2012)
31. Vanghi, V., Damnjanovic, A., Vojcic, B.: The Cdma 2000 System for Mobile Communications: 3G Wireless Evolution. Prentice Hall PTR, Englewood Cliffs, NJ (2004)
32. Vishnevskii, V.M., Semenova, O.M.: Mathematical models to study the polling systems. Autom. Remote Control **67**, 173–220 (2006)

33. Weststrate, J.A.: Analysis and Optimization of Polling Models. PhD thesis, Katholieke Universiteit Brabant (1992)
34. Weststrate, J.A., van der Mei, R.D.: Waiting times in a two-queue model with exhaustive and Bernoulli service. Zeitschrift für Oper. Res. **40**, 289–303 (1994)