Queueing Syst (2007) 55: 161–178
DOI 10.1007/s11134-007-9010-4

Iterative approximation of k-limited polling systems

M. van Vuuren · E.M.M. Winands

Received: 1 May 2006 / Revised: 15 December 2006 / Published online: 21 March 2007

Abstract The present paper deals with the problem of calculating queue length distributions in a polling model with (exhaustive) k-limited service under the assumption of general arrival, service and setup distributions. The interest for this model is fueled by an application in the field of logistics. Knowledge of the queue length distributions is needed to operate the system properly. The multi-queue polling system is decomposed into single-queue vacation systems with k-limited service and state-dependent vacations, for which the vacation distributions are computed in an iterative approximate manner. These vacation models are analyzed via matrix-analytic techniques. The accuracy of the approximation scheme is verified by means of an extensive simulation study. The developed approximation turns out to be accurate, robust and computationally efficient.

Keywords Polling systems · k-limited service · Approximation · Decomposition

Mathematics Subject Classification (2000) 60K25 · 68M20 · 90B30

This research is supported by the Technology Foundation STW, applied science division of NWO and the technology programme of the Dutch Ministry of Economic Affairs.

M. van Vuuren
Department of Mathematics and Computer Science, Technische Universiteit Eindhoven, P.O. Box 513, 5600 MB Eindhoven, The Netherlands
e-mail: m.v.vuuren@tue.nl

E.M.M. Winands (✉)
Department of Mathematics and Computer Science, Department of Technology Management, Technische Universiteit Eindhoven, P.O. Box 513, 5600 MB Eindhoven, The Netherlands
e-mail: e.m.m.winands@tue.nl

1 Introduction

A typical polling system consists of a number of queues, attended by a single server in a fixed order. There is a huge body of literature on polling systems that has developed since the late 1950s, when the papers [29, 30] concerning a patrolling repairman model for the British cotton industry were published. Polling systems have a wide range of applications in communication, production, transportation and maintenance systems. Excellent surveys on polling systems and their applications may be found in [39, 41, 42] and in [28].

The vast majority of the literature is concerned with the two traditional service disciplines, the *exhaustive* and *gated* policies. Exhaustive means that a queue must be empty before the server moves on, whereas in case of gated service only those customers in the queue at the polling start are served. Suggested references for readers who would like to pursue their study of the exhaustive and gated policies are [39, 41, 42]. The main drawback of these traditional policies is the inability to prioritize among the different queues for improving total system performance. A more sophisticated service strategy offering this possibility is the *k-limited* service strategy. Under this k-limited strategy the server continues working at a queue until either a predefined number of k customers is served or until the queue becomes empty, whichever occurs first. Note that the case $k \to \infty$ is equivalent to the exhaustive service strategy. In many applications of polling systems, the objective function typically depends not only on the *mean* queue lengths, but on the *complete* marginal queue length distributions (an illustrative application is described at the end of the present section). The present paper, therefore, aims to study the marginal queue length distributions in continuous-time polling systems with

k-limited service under the assumption of general arrival, service and setup distributions.

To this very day, not only hardly any *exact* results for polling systems with the k-limited service policy have been obtained [25, 36, 37, 47], but also their derivations give little hope for extensions to more realistic systems. This deficiency of exact results is due to the fact that the k-limited service discipline does not satisfy a well-known branching property independently ascertained by [17] and [38]. This branching property causes a striking dichotomy in complexity across the analysis of various polling systems, where the k-limited service policy is on the wrong side of the borderline implying that even mean queue lengths are in general not known. In the absence of exact results for the marginal queue length distributions, people have resorted to *numerical* approaches, such as the power series algorithm [2] and techniques based on discrete Fourier transforms [26]. The main disadvantage of both methods is that time and memory requirements are exponential functions of the number of queues.

A feasible *approximate* approach for the queue length distribution in a k-limited polling system is the decomposition method, in which the polling system is decomposed in vacation systems, for which the vacation distributions are computed in an iterative approximate manner. At each step in the iteration the mathematical analysis focusses on one single queue, whereas the other queues in the system determine the length of the vacation period. This decomposition method is adopted by the present research as well. We have to remark that these decomposition methods seem to be applicable to a wide variety of queueing systems (see, e.g., [9, 18, 44, 45]). In the past, some systems related to the one of the present paper have been studied by the decomposition approach, i.e., a k-limited polling system with finite buffers under the assumption of Poisson arrival processes [23] or a k-limited polling system in combination with a reservation mechanism [24]. The qualitative observations of these studies seem to carry over to the system of the present paper.

The key observation, which is at the same time the mathematical motivation of the present study, is the fact that it is extremely important to capture the correlations among the different queues, since these correlations have a significant impact on the performance measures. Whereas [23] does not take these dependencies into account, [24] proposes to take a weighted sum of a completely uncorrelated and a perfectly correlated system in each step of the iteration by using a pre-defined mixing probability. Although the method of [24] clearly outperforms the procedure that ignores the correlations, this procedure is unable to compensate for correlations in systems with only two queues and is also difficult to apply for systems with more than two queues. That is, since the quality of the procedure strongly depends on the mixing probability, it is rather complicated to find an expression of this probability providing accurate results over the entire range of parameters. Further, the procedure of [24] is based on generating functions, the numerical determination of zeros and the numerical inversion of characteristic functions, considerably increasing the computational time of the algorithm. Finally, due to special features of the protocol studied in [24] the correlations between the queue lengths are relatively small compared to our system (e.g., in case all queues have a service limit of 1 the correlations vanish), which makes the approach of [24] well suited for that particular protocol.

Therefore, the goal of the present study is the development of a computationally efficient iterative approximation method for the marginal queue length distributions in the k-limited polling model. The main challenge can be found in the estimation of the correlations between the queue lengths in each step of the iterative algorithm. The vast majority of the literature on polling systems is devoted to delay figures, while almost no attention has been given to the analysis of such correlations. By using the recently developed *mean value analysis* for polling systems of [48] as the starting point, [32] derives heavy-traffic asymptotics for the covariances between successive visit times in polling systems with mixtures of *gated* and *exhaustive* service under the assumption of Poisson arrivals. Subsequently, [32] proposes simple closed-form approximations of these covariances for stable systems, i.e., with load less than one. However, to the best of our knowledge no results are known for the correlations among queues in polling systems with *k-limited* service.

The key ideas of the approach undertaken in the present paper for polling systems with k-limited service are as follows:

1. The dependence between the queue under consideration and the other queues is taken into account by the introduction of conditional vacations (also called intervisit periods), i.e., the length of the intervisit period is positively correlated to the length of the preceding visit period.
2. The mutual dependencies of the other queues are approximated via standard probabilistic arguments and the conditional intervisit periods.

The main contribution of the present paper is the development of a novel iterative approximation scheme for k-limited polling systems with general arrival, service and setup distributions. The algorithm developed in the present paper only needs information on the first two moments of all distributions. The accuracy of the approximation scheme is verified by means of an extensive simulation study. The approximation scheme turns out to be robust and computationally efficient, while the differences between the exact and approximate values are small within a reasonable margin. In particular, the time complexity is only polynomial in the number of queues and the service limits. The main

building block of this algorithm is a k-limited service vacation model with state-dependent vacations, which has not been studied before in the open literature. In this vacation model, the vacation length depends on the length of the preceding visit period to the queue. As a spin-off, we present an exact analysis for this vacation model with the help of matrix-analytic techniques. A final word on the applicability of the algorithm is that it can also be used as approximation for the exhaustive discipline by taking a "large" value of the service limits. Therefore, our algorithm can also be seen as extension of [13] for the exhaustive polling system with Poisson arrivals.

The remainder of the present section is devoted to the application that led us to this model. Although in the past the k-limited strategy proved its merit in communication systems (see, e.g., [3, 7]), the specific application that raised our attention is in the field of logistics. In many stochastic multi-product single-capacity make-to-stock production systems considerable setup times are incurred, i.e., the so-called *stochastic economic lot scheduling problem* (SELSP) [46]. The presence of these setup times in combination with the stochastic environment are the key complicating factors of the SELSP. On the one hand, one aims for short cycle lengths, and thus frequent production opportunities for the various products, in order to be able to react to the stochasticity in the system. On the other hand, short cycle lengths will increase the setup frequency, which has a negative influence on the amount of capacity available for production. Consequently, this effect will hinder the timely fulfillment of demand.

In the context of the SELSP, the exhaustive service discipline has been studied under the assumption of Poisson demand processes by [14, 15]. A major drawback of this exhaustive policy is that one single product, for which a high demand arrives in a certain period of time, may occupy the machine for quite a while. The impacts of this phenomenon on the other products are stock outs, highly variable cycle lengths and high costs. The k-limited policy circumvents this drawback and offers the possibility to the manager to control both the setup frequencies and the cycle lengths.

The optimal base-stock levels in this system can be obtained by solving standard newsboy problems for which the *complete* queue length distributions in (k-limited) polling systems are required. For more information on newsboy problems, see, e.g., [50]. Moreover, in many telecommunication systems the single most important performance measure is often not an aggregate measure like the mean waiting time, rather the probability that the delay exceeds a predefined threshold. In view of both the described production setting and the dimensioning of a telecommunication network, the importance of an accurate approximation of the complete queue length distribution, as obtained in the present paper, is evident.

The rest of the present paper is organized as follows. Section 2 gives, besides the introduction of the model and further notation, a high-level view of the approximation scheme. In Sect. 3 the approximations for the mean and the variance of the conditional intervisit period are presented. Building on these results, Sect. 4 analyses a k-limited vacation model with state-dependent vacations. Section 5 contains an overview of the iterative procedure to calculate the performance measures of interest. An extensive numerical study to test the accuracy of the approximation algorithm is presented in the penultimate section. Finally, the last section describes the main conclusions of the present research and indicates some possible directions for further research.

2 Model description and notation

We consider a system with one single server for $N \geq 2$ queues, in which there is infinite buffer capacity for each queue. The server visits and serves the queues in a fixed cyclic order. We index the queues by i, $i = 1, 2, \ldots, N$, in the order of the server movement. When visiting queue i, $i = 1, 2, \ldots, N$, the server continues working at this queue until either a predefined number of k_i customers is served or until the queue becomes empty, whichever occurs first. Notice that $k_i = \infty$ amounts to the standard exhaustive service policy.

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

The mean total setup time $\mathbb{E}[S]$ in a cycle is given by

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

The occupancy rate ρ_i at queue i is defined by

$$\rho_i = \frac{\mathbb{E}[B_i]}{\mathbb{E}[A_i]},$$

and the total occupancy rate ρ is given by $\rho = \sum_{i=1}^{N} \rho_i$. Note that the occupation rates do not include the setup times. Hence, especially for small values of the service limits k_i the effective load on the system is considerably higher.

The cycle length C_i of queue i, $i = 1, 2, \ldots, N$, is defined as the time between two successive arrivals of the server at this queue. It is well-known that the mean cycle length is independent of the queue involved and is given by

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

This identity can be proved by observing that the amount of work *arriving* during a cycle should on average equal the amount of work *departing* during a cycle, i.e.,

$$\rho\mathbb{E}[C] = \mathbb{E}[C] - \mathbb{E}[S]. \tag{2}$$

Unfortunately, higher moments of the cycle length are analytically intractable and, certainly, depend on the queue involved.

The visit period V_i of queue i, $i = 1, 2, \ldots, N$, is the time the server spends servicing customers at queue i excluding setup time. Since the server is working a fraction ρ_i of the time on queue i, the mean of a visit period of queue i reads

$$\mathbb{E}[V_i] = \rho_i \mathbb{E}[C], \quad i = 1, 2, \ldots, N. \tag{3}$$

Subsequently, the intervisit period I_i of queue i, the time between a departure epoch of the server from queue i and its subsequent arrival to this queue, is defined as

$$I_i := C_i - V_i, \quad i = 1, 2, \ldots, N.$$

A necessary and sufficient stability condition reads here (see [16], for a rigorous proof in the special case of Poisson arrivals)

$$\rho + \mathbb{E}[S] \max_{1,2,\ldots,N} \frac{1}{\mathbb{E}[A_i]k_i} < 1. \tag{4}$$

If the system is stable, (4) may be rewritten by using (1) as follows

$$\frac{\mathbb{E}[C]}{\mathbb{E}[A_i]} < k_i, \quad i = 1, 2, \ldots, N.$$

In words, this means that for a stable system the average number of type-i customers arriving in a cycle is smaller than the service limit k_i, i.e., the maximum number of type-i customers served in a cycle. Throughout the present paper, the assumption is made that stability condition (4) is fulfilled.

Our main interest is in L_i, the queue length at queue i at an arbitrary point in time, $i = 1, 2, \ldots, N$. The main result of the present paper is the development of an iterative scheme to approximate the *complete* distribution of L_i. For the special case of Poisson arrivals, our results for the queue length distribution can be readily translated into results for the distribution of the customer delay via the distributional form of Little's law [20].

We continue the present section with a high-level description of our approximation method. The key approximation idea is that we decompose the original k-limited polling system with N queues into a set of N separate k-limited *single*-queue models with vacations. At each step in the iteration the mathematical analysis focusses on one single queue i, whereas the other queues in the system determine the length of the vacation period (intervisit period) of queue i, $i = 1, 2, \ldots, N$. The bottleneck in this approximation is the derivation of the distribution of the intervisit period, which will be done in an iterative way. If we assume that the distribution of the intervisit period is known in step n of the iteration, the distribution of the visit period in step $n+1$ is derived by means of a queueing analysis for the k-limited single-queue model with vacations (see Sect. 4). On its turn, the latter distribution can be used to compute the distribution of the length of the intervisit period in step $n+1$ (see Sect. 3).

Since it is more likely that a long (short) visit period is followed by a long (short) intervisit period, conditional intervisit periods are introduced. That is, the length of an intervisit period is assumed to be positively correlated to the number of customers served in the preceding visit period. The subsequent two sections aim to answer the following questions:

1. What are the first two moments of an intervisit period for queue i given that $l = 0, 1, \ldots, k_i$ customers are served in queue i in the preceding visit period (see Sect. 3).
2. What is the distribution of a visit period for queue i given the first two moments of the conditional intervisit periods (see Sect. 4).

3 Intervisit period

The present section computes the first two moments of an intervisit period for queue i given that $l = 0, 1, \ldots, k_i$ customers are served in queue i in the preceding visit period. The input of the present section are the stationary probabilities $\pi_i(l)$ that l customers are served during this visit period of queue i. These probabilities follow from the analysis of the vacation model in the previous iteration step as expounded in Sect. 4. For presentation reasons, we omit throughout this section the superscript n in all random variables denoting the corresponding iteration step n.

3.1 First moments

The intervisit period of a queue i is obviously positively correlated to the preceding visit period of queue i, $i = 1, 2, \ldots, N$. Therefore, we introduce so-called *conditional* visit periods $V_i(l)$, intervisit periods $I_i(l)$ and cycles $C_i(l)$

conditioned on the number of customers $D_i = l$ served in the visit period of queue i, $l = 0, 1, \ldots, k_i$.

The mean conditional cycle lengths may be approximated by using approximate balance equations for $C_i(l)$ as proposed by [22],

$$(\rho - \rho_i)\mathbb{E}[C_i(l)] + l\mathbb{E}[B_i] \approx \mathbb{E}[C_i(l)] - \mathbb{E}[S],$$
$$i = 1, 2, \ldots, N,\ l = 0, 1, \ldots, k_i, \tag{5}$$

which equate the amount of work arriving (left hand side) and the amount of work departing during conditional cycles (right hand side). The balance equation (5) is obviously an approximation, since it assumes balance within each conditional cycle which may not hold. Notice the similarity with the *exact* balance equation for the *unconditional* cycle length, for which work-in is equal to work-out. Solving (5) results in

$$\mathbb{E}[C_i(l)] \approx \frac{l \cdot \mathbb{E}[B_i] + \mathbb{E}[S]}{1 - \rho + \rho_i},$$
$$i = 1, 2, \ldots, N,\ l = 0, 1, \ldots, k_i.$$

We extend the approximation of [22] by multiplying the individual values $\mathbb{E}[C_i(l)]$ with a scaling factor $c_i \in \mathbb{R}$ in such a way that the correct unconditional cycle length as given by (1) is maintained, i.e.,

$$c_i = \frac{\mathbb{E}[C]}{\sum_{l=0}^{k_i} \pi_i(l)\mathbb{E}[C_i(l)]}, \quad i = 1, 2, \ldots, N,$$

where $\pi_i(l)$ are obtained via the analysis of the vacation model in the previous iteration step (see Sect. 4). This scaling obviously facilitates the convergence and stability of the algorithm.

Then, the mean conditional intervisit periods $I_i(\cdot)$ can be approximated in the following way,

$$\mathbb{E}[I_i(l)] \approx \mathbb{E}[C_i(l)] - l \cdot \mathbb{E}[B_i],$$
$$i = 1, 2, \ldots, N,\ l = 0, 1, \ldots, k_i. \tag{6}$$

Finally, we define a conditional visit period $V_i^j(l)$ as the length of the visit period of queue j given that in the preceding visit to queue i precisely l customers are served, $l = 0, 1, \ldots, k_i$. The mean of this random variable reads

$$\mathbb{E}[V_i^j(l)] \approx \rho_j \mathbb{E}[C_i(l)],$$
$$i = 1, 2, \ldots, N,\ l = 0, 1, \ldots, k_i,$$
$$j = i + 1, \ldots, N, 1, \ldots, i - 1, \tag{7}$$

which completes the analysis of the conditional first moments.

We have to remark that the approximations of the present subsection only compensate for the correlations between the visit period and the *immediately following* intervisit period. Although it is not inconceivable that one may come up with more sophisticated approximations, the numerical evaluation of Sect. 6 shows that our approximations are still very effective in capturing the correlations among the queues.

3.2 Second moments

The goal of the present subsection is the development of an approximation for the variance of the *conditional* intervisit periods $I_i(\cdot)$. The starting point of our analysis are the *unconditional* intervisit periods I_i. Since the setup times are assumed to be uncorrelated (see Sect. 2), the variance of such an unconditional intervisit period I_i is given by

$$\mathbb{V}\mathrm{ar}[I_i] = \sum_{j \neq i} \mathbb{V}\mathrm{ar}[V_j] + \sum_j \mathbb{V}\mathrm{ar}[S_j] + 2\sum_{j \neq i} \sum_{\substack{k > j \\ k \neq i}} \mathbb{C}\mathrm{ov}[V_j, V_k] + \sum_j \sum_{k \neq i} \mathbb{C}\mathrm{ov}[S_j, V_k], \tag{8}$$

where the latter two summations include all the covariances among the various visit periods and among the setup times, respectively, within an intervisit period of queue i. Therefore, the $>$ sign in this summation means that queue k is visited after queue j in this intervisit period.

The terms $\mathbb{V}\mathrm{ar}[V_j]$ in the right-hand side of (8) represent the variance of an unconditional visit periods V_j of queue j. The second moment of such a visit period can be approximated as follows. Conditioning on the number of customers served during the visit period of this queue and ignoring the correlations between the length of the service times and the number of customers served during the visit period yields

$$\mathbb{E}[V_i^2] = \sum_{l=0}^{k_i} \pi_i(l)\mathbb{E}[V_i^2(l)] \approx \sum_{l=0}^{k_i} \pi_i(l)(l\mathbb{E}[B_i^2] + l(l-1)\mathbb{E}[B_i]^2), \quad i = 1, 2, \ldots, N,$$

with the remark that the probabilities $\pi_i(\cdot)$ are still unknown at this stage. These probabilities are obtained from the analysis of the vacation model in the previous iteration step, see Sect. 4. Now, the variance of V_i can be obtained via standard probabilistic arguments.

Since the terms $\mathbb{V}\mathrm{ar}[S_j]$ are assumed to be input of the system (see Sect. 2), one does not need to approximate them. By definition, the covariance terms $\mathbb{C}\mathrm{ov}[V_j, V_k]$ appearing in (8) can be rewritten as

$$\mathbb{C}\mathrm{ov}[V_j, V_k] = \mathbb{E}[V_j V_k] - \mathbb{E}[V_j]\mathbb{E}[V_k],$$

where the terms $\mathbb{E}[V_j]$ and $\mathbb{E}[V_k]$ follow from (3). To compute the unknown quantity $\mathbb{E}[V_j V_k]$, we condition on the

number D_j of customers served in queue j during the last visit period as follows

$$\mathbb{E}[V_j V_k] = \sum_{l=0}^{k_j} \mathbb{E}[V_j V_k | D_j = l]\pi_j(l)$$
$$\approx \sum_{l=0}^{k_j} l\mathbb{E}[B_j]\mathbb{E}[V_j^k(l)]\pi_j(l),$$

where $\pi_j(l)$ follow from the analysis of Sect. 4 and $\mathbb{E}[V_j^k(l)]$ can be approximated by (7).

Finally, in case a queue k is visited before queue j in the intervisit period of queue i, V_k and S_j are obviously uncorrelated. In case queue j is visited first, we assume independence between setup times and visit periods as well, i.e.,

$$\mathbb{C}\text{ov}[S_j, V_k] \approx 0,$$

and, thus, all terms in (8) have been specified. The numerical results in Sect. 6 show that this assumption is valid as long as the setup times are not too variable.

By definition, the coefficient of variation c_{I_i} of an unconditional intervisit period is, subsequently, given by

$$c_{I_i} = \frac{\sqrt{\mathbb{V}\text{ar}[I_i]}}{\mathbb{E}[I_i]}, \quad i = 1, 2, \ldots, N.$$

We approximate the variance of the conditional intervisit periods $I_i(\cdot)$ by assuming equality of the coefficients of variation of all periods, i.e.,

$$\mathbb{V}\text{ar}[I_i(l)] \approx c_{I_i}^2 \cdot \mathbb{E}[I_i(l)]^2,$$
$$l = 1, 2, \ldots, k_i, \; i = 1, 2, \ldots, N, \tag{9}$$

where an approximation of $\mathbb{E}[I_i(\cdot)]$ is given by (6). We add that we have also experimented with other approximations for the variance of conditional visit period such as assuming equality of the coefficients of variation of all conditional cycle lengths. Approximation (9), however, turned out to be the most accurate one. Finally, notice that (9) is increasing in l.

4 Visit period

The present section aims to compute the distribution of a visit period for queue i given the first two moments of the conditional intervisit periods as computed via (6) and (9) in the preceding section. By means of matrix-analytic techniques, we analyse a single-station vacation model with k-limited service, in which the vacation length depends on the length of the preceding visit period. The authors are aware of only one other study in which this specific dependency is studied under the restrictive assumption of Poisson input [27]. Comprehensive surveys on vacation models can be found in [10, 11, 40].

Since the present section is focussing on one single queue i in a specific iteration step n, the subscript i and superscript n are dropped from all random variables. Throughout the present section, the distribution functions of the arrival and the service times are needed. However, the only information available for these random variables are the first two moments. A common way to obtain an *approximate* distribution is to fit a phase-type distribution on the first two moments as elucidated in Appendix 1 (cf., e.g., [43]). In the remainder of the present section, we assume that the fitted distributions are used as substitute for the arrival and service distributions and that the number of phases needed equal n_A and n_B, respectively.

In the preceding subsection, we have computed the first two moments of the conditional intervisit periods $I(\cdot)$ conditioned on the exact number of customers served in the preceding visit period. To keep the size of the state space for the k-limited vacation model manageable, some of these intervisit periods are aggregated. That is, we draw a distinction between intervisit periods $I(0)$, $I(k)$ and $I(*)$ in which there have been zero, the maximum number or any other number of customers served in the preceding visit period, respectively. In case the service limit at a queue equals one, only $I(0)$ and $I(1)$ have to be distinguished. The period $I(*)$ is, thus, defined as,

$$I(*) := \sum_{l=1}^{k-1} \pi(l) I(l),$$

with first two moments,

$$\mathbb{E}[I(*)] := \sum_{l=1}^{k-1} \pi(l)\mathbb{E}[I(l)], \quad \text{and}$$

$$\mathbb{E}[I(*)^2] := \sum_{l=1}^{k-1} \pi(l)\mathbb{E}[I(l)^2],$$

where $\pi(l)$ follow from the previous iteration step. We have to remark that we have tested this aggregation of intervisit periods for a wide variety of cases, from which we concluded that it has only negligible (negative) impact on the results, which is outweighted by the gain in efficiency.

In sum, the system under consideration is a single-server k-limited vacation model with three different kinds of intervisit periods dependent on the number of customers served in the preceding visit period. In order to construct these intervisit periods in an efficient way, we introduce the auxiliary mutually independent random variables $\tilde{I}(*)$ and $\tilde{I}(k)$, which are independent of $I(0)$ as well. These random variables satisfy

$$I(*) = \tilde{I}(*) + I(0), \quad \text{and} \quad I(k) = \tilde{I}(k) + I(*),$$

which is always possible since the variances of the conditional intervisit periods are increasing in l as shown in (9). Thereupon, phase-type distributions are fitted on $I(0)$, $\tilde{I}(*)$ and $\tilde{I}(k)$ (see Appendix 1 for further details) in such a way that the first two moments of $I(*)$ and $I(k)$ are correct. If we assume that the number of phases needed for the description of $I(0)$, $\tilde{I}(*)$ and $\tilde{I}(k)$ equal $n_{I(0)}$, $n_{\tilde{I}(*)}$ and $n_{\tilde{I}(k)}$, respectively, the total number n_I of phases for the intervisit process is given by $n_I = n_{I(0)} + n_{\tilde{I}(*)} + n_{\tilde{I}(k)}$.

The k-limited vacation model can be described by a continuous-time Markov process with states (i, j, m). The state variable $i = 0, 1, \ldots$ denotes the total number of customers in the specific queue under consideration, whereas the state variable $j = 1, 2, \ldots, n_A$ indicates the phase of the arrival process A. Finally, $m = 1, 2, \ldots, n_D$ indicates the phase of the departure process D, which is the combination of the service process and vacation processes $I(0)$, $\tilde{I}(*)$ and $\tilde{I}(k)$. These latter two processes can be modeled by one single variable, since the server is either serving customers or is on vacation. When the server is serving customers, one has to keep track of the phase of the service process and of the number of customers already served in the corresponding visit period. On the other hand, when the server is on vacation the phase of the corresponding vacation period is needed. Consequently, the total number of states for the departure process is $n_D = k \times n_B + n_I$. The phases of this departure process are grouped as follows: first, we group all phases related to the k service processes and, then, the phases of $\tilde{I}(k)$, $\tilde{I}(*)$ and $I(0)$.

Refer by level i to the set of states with i customers in the system and group the states by these levels, so that (i, j, m) precedes (i', j', m') if $i < i'$. Within each level, the states are grouped according to the arrival phase, so that (i, j, m) precedes (i, j', m') if $j < j'$. Lastly, the states are ordered by the departure phase, so that (i, j, m) precedes (i, j, m') if $m < m'$. Now, one may verify that the introduced Markov process is a *quasi-birth-and-death* (QBD) process where the infinitesimal generator $\mathbf{Q}$ has the following block-tridiagonal structure,

$$\mathbf{Q} = \begin{pmatrix} B_{00} & B_{01} & 0 & 0 & 0 & \cdots \\ B_{10} & A_1 & A_0 & 0 & 0 & \cdots \\ 0 & A_2 & A_1 & A_0 & 0 & \cdots \\ 0 & 0 & A_2 & A_1 & A_0 & \\ \vdots & \vdots & & \ddots & \ddots & \ddots \end{pmatrix}.$$

Below we specify the submatrices in $\mathbf{Q}$, where we use the concept of *Markovian Arrival Process* (MAP) (see, e.g., [1]) to describe the arrival and departure processes. In general, a MAP is defined in terms of a continuous-time Markov process with finite state space $\{0, \ldots, m-1\}$ and generator $G_0 + G_1$. The element $G_1(i, j)$ denotes the intensity of transitions from i to j accompanied by an arrival. For $i \neq j$ element $G_0(i, j)$ denotes the intensity of the remaining transitions from i to j, while the diagonal elements $G_0(i, i)$ are strictly negative and chosen such that the row sums of $G_0 + G_1$ are zero.

The arrival process can be straightforwardly represented by such a MAP, the states of which correspond to the phases of this process. Its generator can be expressed as $G_0^A + G_1^A$, where the transition rates in G_1^A are the ones that correspond to an arrival of a customer to the system. The transition rates of the G_0^A and G_1^A matrices are listed in Appendix 2.

The MAP for the departure process with generator $G_0^D + G_1^D$ is a little more involved. All transitions related to the vacation periods do not cause departures and are, thus, within G_0^D. Completion of a service process, obviously, leads to a departure implying that the corresponding rates are in G_1^D. Transitions within a service process not causing departures are, of course, part of G_0^D. Further, we have to distinguish between the situation when there are more than two customers in the system or not. In the first situation, if a departure is not the kth departure the next service process is started and if it is the kth departure a new vacation period is begun. To deal with the situations in which there are only zero or one customers present, we have to introduce matrices $\tilde{G}_0^D$ and $\tilde{G}_1^D$, representing the transition within level 0 and the transitions from level 1 to level 0, respectively. We can recognize two differences between these matrices and $G_0^D + G_1^D$. First, when a service process is completed which is not the kth service, a vacation period is commenced instead of the next service. Second, when a vacation period is finished, we jump to process $I(0)$ instead of to the service process of the first customer in the visit period. The transition rates for G_0^D, G_1^D, $\tilde{G}_0^D$ and $\tilde{G}_1^D$ are summarized in Appendix 2.

Now, we are in the position to describe all the submatrices in $\mathbf{Q}$, i.e.,

$$\begin{aligned} B_{01} &= G_1^A \otimes I_{n_D}, \\ B_{00} &= G_0^A \otimes I_{n_D} + I_{n_A} \otimes \tilde{G}_0^D, \\ B_{10} &= I_{n_A} \otimes \tilde{G}_1^D, \\ A_0 &= G_1^A \otimes I_{n_D}, \\ A_1 &= G_0^A \otimes I_{n_D} + I_{n_A} \otimes G_0^D, \\ A_2 &= I_{n_A} \otimes G_1^D, \end{aligned}$$

where I_n is the identity matrix of size n and if A is an $n_1 \times n_2$ matrix and B an $n_3 \times n_4$ matrix the Kronecker product $A \otimes B$ is an $n_1 n_3 \times n_2 n_4$ matrix defined by

$$A \otimes B = \begin{pmatrix} A(1,1)B & \cdots & A(1,n_2)B \\ \vdots & & \vdots \\ A(n_1,1)B & \cdots & A(n_1,n_2)B \end{pmatrix}.$$

Fig. 1 Algorithm of [34] for finding the rate matrix R, where $\|.\|$ denotes a matrix-norm and ϵ some positive number

```
N := A1
L := A0
M := A2
W := A1
dif := 1

while dif > ϵ
{
   X := −N^{−1}L
   Y := −N^{−1}M
   Z := LY
   dif := ||Z||
   W := W + Z
   N := N + Z + MX
   Z := LX
   L := Z
   Z := MY
   M := Z
}
R := −A0W^{−1}
```

This completes the description of the QBD. If we let q_i denote the equilibrium probability vector of level i, the corresponding balance equations are given by

$$q_{n-1}A_0 + q_nA_1 + q_{n+1}A_2 = 0, \quad n \geq 2,$$

and

$$q_0B_{00} + q_1B_{10} = 0, \tag{10}$$

$$q_0B_{01} + q_1A_1 + qA_2 = 0. \tag{11}$$

Introducing the rate matrix R as the minimal nonnegative solution of the nonlinear matrix equation

$$A_0 + RA_1 + R^2A_2 = 0,$$

it can be proved that the equilibrium probabilities satisfy (see, e.g., [35])

$$q_{n+1} = q_nR, \quad n \geq 1.$$

To determine this matrix R we use the algorithm developed by [34] as listed in Fig. 1. The vectors q_0 and q_1 follow from the boundary conditions (10), (11), and the normalization condition. This queue length distribution q_i yields the following expression for the distribution of the length of a visit period,

$$\pi(l) = \frac{h(l)}{\sum_{i=0}^{k} h(i)}, \quad l = 0, 1, \ldots, k, \tag{12}$$

where $h(l)$ is the total rate of jumps to a vacation period after serving l customers. To calculate $h(l)$ we have to sum all transition rates from a state where $l-1$, $l = 1, 2, \ldots, k$, customers are served (or 0 customers when $l = 0$) to a vacation, multiplied by the probability of being in that specific state. Further, we recall that the indices of $q_\cdot(\cdot)$ within the brackets correspond to lexicographically ordered states of the arrival and departure processes. So,

$$\begin{aligned} h(0) = {} & \sum_{i=1}^{n_A} \sum_{j=1}^{n_{I(0)}} (q_1((i-1)n_D + kn_B + n_{\tilde{I}(k)} + n_{\tilde{I}(*)} + i) \\ & \times B_{00}((i-1)n_D + kn_B + n_{\tilde{I}(k)} + n_{\tilde{I}(*)} + i, \\ & \qquad (i-1)n_D + kn_B + n_{\tilde{I}(k)} + n_{\tilde{I}(*)} + 1)), \end{aligned}$$

$$\begin{aligned} h(l) = {} & \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} q_1((i-1)n_D + (l-1)n_B + j)) \\ & \times B_{10}((i-1)n_D + (l-1)n_B + j, \\ & \qquad (i-1)n_D + kn_B + n_{\tilde{I}(k)} + 1), \\ & l = 1, \ldots, k-1, \end{aligned}$$

$$\begin{aligned} h(k) = {} & \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} r((i-1)n_D + (k-1)n_B + j) \\ & \times A_2((i-1)n_D + (k-1)n_B + j, \\ & \qquad (i-1)n_D + kn_B + 1), \end{aligned}$$

where

$$r = \sum_{i=1}^{\infty} q_i = \sum_{i=1}^{\infty} q_1R^{i-1} = q_1(I_{n_A \times n_D} - R)^{-1},$$

which completes the analysis of the k-limited vacation model.

5 Iterative algorithm

As described at the end of Sect. 2, the performance characteristics of the k-limited polling system are approximated by an iterative scheme. The algorithm is as follows.

Outline of the algorithm

- *Step 0*: Choose initial characteristics for all queues.
- *Step 1*: For $i = 1$ to N, determine the first two moments of the conditional intervisit period $I_i(\cdot)$ for queue i from (6) and (9), respectively.
- *Step 2*: For $i = 1$ to N, determine the distribution of the visit period V_i from (12).
- *Step 3*: Repeat Step 1 and 2 until the characteristics for all queues have converged.
- *Step 4*: For $i = 1$ to N, compute the performance measures of interest for queue i.

Initialization In Step 0 of the algorithm, we have to choose initial values for $\pi_i(l)$, $l = 0, 1, \ldots, k_i$ and $i = 1, 2, \ldots, N$. The assumption is made that all of these probabilities are zero except for $\pi_i(k_i)$, $i = 1, 2, \ldots, N$. Notice that, via the approach developed in Sect. 3, the correct mean cycle lengths are obtained as computed by (1). We note that we have experimented with a large number of initial values, from which we concluded that the starting values of the algorithm have no, or at least negligible, impact on the results.

Convergence criterion After Step 1 and 2 we check whether the iterative algorithm has converged by comparing the probabilities $\pi_i(\cdot)$, $i = 1, 2, \ldots, N$, in the $(n-1)$th and nth step. We decide to stop when the maximum of the absolute values of the differences is less than ε; otherwise we repeat Step 1 and 2. Hence, the convergence criterion is

$$\max_{l=0,1,\ldots,k_i} \left|\pi_i^{(n)}(l) - \pi_i^{(n-1)}(l)\right| < \varepsilon, \quad \forall i = 1, 2, \ldots, N,$$

where ε is chosen to be 10^{-4}. Of course, we may use other stop-criteria as well, e.g., mean queue lengths or mean intervisit periods.

Complexity analysis The complexity of this method is as follows. Within the iterative algorithm, solving a subsystem consumes most of the time. In one single iteration step N subsystems are solved. The number of iterations needed is difficult to predict, but in practice this number is about 10 to 15 iterations. The time consuming part of solving a subsystem is the calculation of the R-matrix. This can be done in $O(n_i^3)$ time, where n_i is the size of the R matrix of subsystem i. Then, the time complexity of one iteration becomes $O(N \max_i(n_i^3))$. This means that the time complexity is polynomial in the number of queues, the service limits and the number of phases for each process.

6 Numerical evaluation

The present section reports on an extensive numerical study designed to assess the accuracy of the approximation method developed. We compare the first two moments and tail probabilities of the queue length distribution with the ones produced by discrete event simulation. Each simulation run is sufficiently long such that the widths of the 95% confidence intervals of the performance measures of interest are smaller than 1%. A first important remark is that the computation time of our algorithm is considerably less than the simulation time, which can mount up to fifteen minutes or more. This inefficiency of simulation techniques for (k-limited) polling systems has been observed before by, e.g., [2].

6.1 Parameter setting

We use a broad set of parameters for the tests. The number of queues in the system is varied between 2, 5 and 10, whereas the service limits are either 1, 5 or 10. The total load on the system varies between 0.45, 0.60 and 0.75; as mentioned in Sect. 2 this load does not include the setup times. Hence, especially for small values of the service limits k_i the effective load on the system is considerably higher. For this reason, some cases are unstable, meaning that (4) does not hold, and are thus removed from the test bed.

The squared coefficients of variation of the interarrival, service and setup times for each queue are identical and are varied between 0.25 and 2 and between 0.25 and 1, respectively. We have to remark that we envision production systems as the main application for the present paper (see also Sect. 1). Since the variations in the setup and service times tend to be small in such systems—in contrast to telecommunication systems where heavy-tailed random variables are common—we only consider cases in which these variations are indeed relatively small. Furthermore, we test cases for which the setup times are 10 times smaller than the service times and cases for which setup and service times are equal.

Furthermore, both balanced and imbalanced polling systems are considered. In the balanced cases we set the arrival rates of all queues equal to 1. We test imbalance in the average interarrival times by making the load of the most heavily loaded queue 10 times higher then that of the least heavily loaded queue, and by letting the arrival rates of the other queues change linearly such that the overall mean arrival rate is maintained at 1. For example, in case of 5 queues we get arrival rates (0.182, 0.591, 1.000, 1.409, 1.818). Testing imbalance in the service times proceeds along the same lines. This leads to a total of $3^4 2^5 = 2592$ test cases, which are summarized in Table 1. After removing the unstable

Table 1 Test bed

Test bed				
Parameter	Notation	Value		
		Low	Medium	High
Number of queues	N	2	5	10
Load	ρ	0.45	0.60	0.75
Service limit	k_i	1	5	10
SCV interarrival times	A_i	0.25	1	2
SCV service times	B_i	0.25	–	1
SCV setup time	S_i	0.25	–	1
Imbalance interarrival times	I_{A_i}	1:1	–	1:10
Imbalance service time	I_{B_i}	1:1	–	1:10
Ratio service and setup times	I_{B_i/S_i}	1:1	–	10:1
	Number of instances		2592	

cases, we end up with a total of 2088 cases. For further reference, we have classified the values for each parameter in the categories low, medium and high.

The performance measures under consideration in the present numerical study are the mean, standard deviation, 0.90-quantile and 0.95-quantile of the marginal queue length distributions, where the α-quantile of the distribution of a random variable X can be defined as the smallest value x such that

$$\mathbb{P}[X \leq x] \geq \alpha.$$

The importance of the quantiles of the queue length distributions lies in the fact that the optimal base-stock levels in the production application described in Sect. 1 precisely equal these quantiles.

6.2 Results

Table 2 summarizes the performance of the approach developed in the present paper showing the average errors and for four error-ranges the percentage of the cases which fall in that range. Overall, we can say that for all performance measures the average error is around 7%, while the errors are for the majority of the cases less than 10%. We believe that these errors are in general satisfactory in view of the complexity of the system under consideration: we study a k-limited service discipline—containing the exhaustive policy as special case—under the assumption of general arrival processes, whilst the fact that our interest is in the complete queue length distribution constitutes an additional complicating factor.

To give this statement a more scientific basis, we compare the performance of our approach to the standard decomposition approach. In such a standard decomposition approach the dependencies among the individual queues are completely ignored. That is, the intervisit periods are assumed to be independent of the length of the preceding visit period, thus the need for conditional cycles and conditional (inter)visit periods cancels, and the correlations among the individual visit periods are set equal to zero. Remark that the application of this standard approach to k-limited polling systems has not been published in the open literature.

The results for the latter approach are listed in Table 3. Comparing this table to Table 2, we can conclude that our approach not only halves the mean errors for all performance measures, but also that the standard approach, in contrast to our approach, quite often results in more than 30% error. This observation clearly underpins the statement made in the introduction that it is extremely important to capture the correlations among the different queues, since these correlations have a significant impact on the performance measures. In particular, the performance of the standard approach significantly degrades as the total load increases as shown in Table 5, which is in agreement with the result of [32] that the correlation between successive visit times converges to one as the total load tends to one for the cases of exhaustive and gated polling systems with Poisson arrivals. Table 4 shows that the accuracy of our approach decreases in heavy traffic as well; the decrease in accuracy is, however, not so severe as for the standard decomposition approach (see, also, Sect. 6.3).

It would also be interesting to compare the performance of our approach to the one of the alternative approach developed in [24]. In this study, it is proposed to take a weighted

Table 2 Overall results approach of present paper

Errors approach of present paper					
	Aver. (%)	0–10%	10–20%	20–30%	>30%
Mean queue lengths	7.26	76.25	17.77	5.12	0.86
SD queue lengths	8.34	71.02	20.16	5.51	3.30
0.90-quantile	6.58	75.62	14.80	5.75	3.83
0.95-quantile	7.33	73.37	15.95	6.80	3.88

Table 3 Overall results standard approach

Errors standard approach					
	Aver. (%)	0–10%	10–20%	20–30%	>30%
Mean queue lengths	15.40	40.95	30.94	15.61	12.50
SD queue lengths	15.26	40.37	29.98	16.91	12.74
0.90-quantile	13.45	57.95	16.52	11.69	13.84
0.95-quantile	13.26	54.02	17.10	14.08	14.80

Table 4 Relative errors for approach of present paper as function of total utilization ρ

Errors approach of present paper as function of ρ (%)			
	Low	Medium	High
Mean queue lengths	4.43	6.72	11.64
SD queue lengths	5.20	6.23	14.95
0.90-quantile	4.11	5.87	10.67
0.95-quantile	4.63	6.50	11.85

Table 5 Relative errors for standard approach as function of total utilization ρ

Errors standard approach as function of ρ (%)			
	Low	Medium	High
Mean queue lengths	8.22	14.54	25.88
SD queue lengths	8.32	14.09	25.78
0.90-quantile	10.11	9.22	22.75
0.95-quantile	6.06	11.71	25.55

sum of a completely uncorrelated and a perfectly correlated system in order to capture the correlations among the queues. A good choice of the desired mixing probability is an interesting problem in itself and the probability used in [24] has not been developed for the k-limited polling system covered in the present paper, rather for a modification of this system, i.e., inclusion of a reservation mechanism. Directly applying the same mixing probability to our setting would certainly wrong the approach of [24] leading to an unfair comparison. Essentially, this observation reveals a weakness of the procedure of [24]: the quality of this procedure strongly depends on the choice of the mixing probability. Taking the above into account, we confine ourselves to a more qualitative comparison between the two approaches. That is, when comparing the errors reported in [24] to the ones listed in Table 2, one can conclude that they are of the same order of magnitude. The approximation method of [24] has, however, only been tested in a system with smaller inherent dependencies for the special case of Poisson arrivals. We have to remark that Tables 6, 7, 8 and 9 show that the interarrival distribution has no or at least negligible effect on the accuracy of our approach.

More specifically, Tables 6 through 9 show the detailed results for our approach, when fixing one parameter at a certain level. When a row is partially empty, it means that this parameter is only tested on two levels. Our approximation method seems to be fairly insensitive to different parameter settings. In this respect, the parameter having the largest impact on the performance is the total utilization ρ as earlier illustrated in Table 5. Moreover, we observe that imbalance in the service times and an increase in the setup times have negative impact on the accuracy, whereas the accuracy of our approach increases as the service limits become larger. This latter observation tempts one to use the approach of the present paper as approximation for the exhaustive policy as well as touched upon in Sect. 7. In the next subsection, we present results for various asymptotic regimes in order to study the effect of the individual parameters even further.

Remark 6.1 In the past, so-called *pseudo-conservation laws*, intensity-weighted sums of mean delays, have been applied quite often to develop accurate and elegant approximations for mean delays in polling systems (and, thus, mean queue lengths as well). Throughout the present paper, we have deliberately left this approach aside, because

Table 6 Relative errors for mean queue lengths

Errors mean queue lengths (%)			
Parameter	Low	Medium	High
N	8.96	7.17	5.74
ρ	4.43	6.72	11.64
k_i	9.35	6.91	6.39
A_i	6.70	6.96	8.14
B_i	6.79	–	7.74
S_i	6.92	–	7.61
I_{A_i}	7.32	–	7.19
I_{B_i}	5.17	–	9.51
I_{B_i/S_i}	5.07	–	8.67

Table 7 Relative errors for SD queue lengths

Errors SD queue lengths (%)			
Parameter	Low	Medium	High
N	8.77	10.21	6.16
ρ	5.20	6.23	14.95
k_i	9.39	7.87	8.18
A_i	6.56	8.25	10.22
B_i	7.72	–	8.97
S_i	8.18	–	8.51
I_{A_i}	8.21	–	8.51
I_{B_i}	6.07	–	10.78
I_{B_i/S_i}	5.65	–	10.07

Table 8 Relative errors for 0.90-quantile

Errors 0.90-quantile (%)			
Parameter	Low	Medium	High
N	9.50	5.87	4.49
ρ	4.11	5.87	10.67
k_i	8.55	6.43	5.60
A_i	6.65	5.79	7.31
B_i	6.15	–	7.02
S_i	6.47	–	6.69
I_{A_i}	6.84	–	6.26
I_{B_i}	4.63	–	8.67
I_{B_i/S_i}	5.23	–	7.45

Table 9 Relative errors for 0.95-quantile

Errors 0.95-quantile (%)			
Parameter	Low	Medium	High
N	7.61	9.23	5.25
ρ	4.63	6.50	11.85
k_i	9.29	6.90	6.60
A_i	6.59	7.51	7.87
B_i	7.02	–	7.64
S_i	7.08	–	7.57
I_{A_i}	7.67	–	6.90
I_{B_i}	5.04	–	9.78
I_{B_i/S_i}	5.85	–	8.27

our approach does not use this technique and because this technique only gives approximations for mean performance measures for the special case of Poisson arrivals (for more information see, e.g., [4] and the references therein). An additional complexity that shows up when applying pseudo-conservations laws to polling systems with k-limited service is that in such systems these laws still contain some unknown terms that have to be approximated as independently shown by [5] and [12]. Note that the most accurate algorithm [6] based on such a pseudo-conservation law can still give up to 20% errors for the mean delays in k-limited polling systems.

6.3 Asymptotic regimes

The foregoing subsection showed the accuracy of the developed approximation for a wide range of cases. The test bed is, undoubtedly, not only representative for practical instances of the production application motivating the present research but also for most applications in communication systems. In the present subsection we, however, want to test the applicability of the approximation beyond all limits and test the accuracy of the approximation in the following *asymptotic* regimes:

1. Highly variable setup and/or service times
2. Heavy traffic, i.e., $\rho \uparrow 1$
3. Large setup times, i.e., $\mathbb{E}[S] \to \infty$
4. Large number of queues, i.e., $N \to \infty$

Before we discuss these regimes in detail, it is important to stress that for none of these regimes any, qualitative or quantitative, results are known for the k-limited policy. However, there are (partial) asymptotic results known for the less intricate exhaustive and gated policies (and, sometimes, for branching-type policies). We mention these results in the present subsection. First of all, we want to give the reader a feeling for what might happen for the k-limited discipline

Table 10 Test bed for highly variable setup and/or service times

Test bed	
Parameter	Value(s)
N	5
ρ	0.6
k_i	5
$c^2_{A_i}$	1
$c^2_{B_i}$	1 4 16
$c^2_{S_i}$	1 4 16
I_{A_i}	1:1
I_{B_i}	1:1
I_{B_i/S_i}	1:1
Number of instances	9

Table 11 Relative errors for highly variable setup and/or service times ($c^2_{S_i} = 1$)

Errors for highly variable setup and/or service times ($c^2_{S_i} = 1$)						
	$c^2_{B_i} = 1$		$c^2_{B_i} = 4$		$c^2_{B_i} = 16$	
	P	S	P	S	P	S
Mean queue lengths	3.8	13.3	14.6	23.6	32.8	40.8
SD queue lengths	3.2	14.0	17.6	28.3	39.4	50.2
0.90-quantile	0.0	33.3	25.0	25.0	38.6	38.6
0.95-quantile	0.0	25.0	20.0	20.0	41.7	50.0

Table 12 Relative errors for highly variable setup and/or service times ($c^2_{S_i} = 4$)

Errors for highly variable setup and/or service times ($c^2_{S_i} = 4$)						
	$c^2_{B_i} = 1$		$c^2_{B_i} = 4$		$c^2_{B_i} = 16$	
	P	S	P	S	P	S
Mean queue lengths	11.1	20.3	19.1	27.9	33.6	41.5
SD queue lengths	9.4	21.1	18.4	30.3	37.1	48.6
0.90-quantile	0.0	0.0	25.0	25.0	44.4	44.4
0.95-quantile	0.0	25.0	16.7	33.3	34.4	50.8

Table 13 Relative errors for highly variable setup and/or service times ($c^2_{S_i} = 16$)

Errors for highly variable setup and/or service times ($c^2_{S_i} = 16$)						
	$c^2_{B_i} = 1$		$c^2_{B_i} = 4$		$c^2_{B_i} = 16$	
	P	S	P	S	P	S
Mean queue lengths	27.4	35.8	30.3	38.5	36.7	44.4
SD queue lengths	28.9	42.4	27.5	41.9	31.8	45.6
0.90-quantile	33.3	33.3	28.6	42.9	36.4	45.5
0.95-quantile	25.0	37.5	33.3	44.4	33.3	46.7

Table 14 Test bed for heavy traffic

Test bed	
Parameter	Value(s)
N	5
ρ	0.75 0.8 0.85 0.9 0.95
k_i	3 5
$c^2_{A_i}$	1
$c^2_{B_i}$	1
$c^2_{S_i}$	1
I_{A_i}	1:1
I_{B_i}	1:1
I_{B_i/S_i}	10:1
Number of instances	10

Table 15 Relative errors for heavy traffic ($k = 3$)

Errors for heavy traffic ($k = 3$)										
	$\rho = 0.75$		$\rho = 0.8$		$\rho = 0.85$		$\rho = 0.90$		$\rho = 0.95$	
	P	S	P	S	P	S	P	S	P	S
Mean queue lengths	0.8	41.7	0.3	49.4	0.1	58.2	1.8	68.5	15.5	68.5
SD queue lengths	−16.5	38.0	−20.2	46.7	−19.0	57.1	−7.8	69.5	26.0	85.3
0.90-quantile	0.0	50.0	0.0	33.3	0.0	50.0	−14.3	50.0	16.2	71.4
0.95-quantile	0.0	33.3	0.0	50.0	−16.7	66.7	−10.0	70.0	22.9	85.1

Table 16 Relative errors for heavy traffic ($k = 5$)

Errors for heavy traffic ($k = 5$)										
	$\rho = 0.75$		$\rho = 0.8$		$\rho = 0.85$		$\rho = 0.90$		$\rho = 0.95$	
	P	S	P	S	P	S	P	S	P	S
Mean queue lengths	−4.0	39.4	−6.8	46.7	−10.4	55.2	−14.6	65.4	−11.5	78.9
SD queue lengths	−29.0	35.7	−42.3	43.7	−52.4	53.7	−51.6	65.9	−19.0	82.1
0.90-quantile	0.0	50.0	0.0	33.3	0.0	50.0	−33.3	66.7	−20.2	82.0
0.95-quantile	0.0	33.3	0.0	50.0	−40.0	60.0	−33.3	66.7	−20.1	82.8

in the corresponding regimes. Secondly, these results for the exhaustive and gated policies clearly show that polling systems display aberrant behavior in these asymptotic cases implying that one cannot expect to be able to develop one single algorithm which is accurate both in standard traffic settings and for all possible asymptotic regimes.

6.3.1 Highly variable setup and/or service times

The first case, as summarized in Table 10, investigates the impact of the squared coefficient of variation of both the setup and service times. Thereto, these quantities are varied between 1, 4 and 16. Tables 11, 12 and 13 summarizes the results for this case. In these tables, and all other tables throughout this subsection, the values in the column P refer to the relative errors of the approach of the present paper, whereas column S shows the relative errors of the standard approach. As observed from these tables, the accuracy of our approach is somewhat disappointing. The reason for this observation is perdu in the way of conditioning introduced in Sect. 3. That is, we condition on the *number* of customers served without taking the *length* of each service period into account. In case of highly variable service times this can cause difficulties. Conditioning on the length significantly complicates our analysis, since the length of a visit period is continuous, has an infinite support and is more difficult to be monitored. For the impact of the variance of the setup times similar observations apply, where the assumption of independence between setup and (subsequent) visit periods is the main reason for the decrease in accuracy. Finally, we should stress that the standard approach again clearly tastes defeat. Also, the alternative approach of [24] brings no relief, since this approach has specifically been developed for deterministic distributions.

Table 17 Test bed for large setup times

Test bed	
Parameter	Value(s)
N	5
ρ	0.6
k_i	100
$c^2_{A_i}$	1
$c^2_{B_i}$	1
$c^2_{S_i}$	1
I_{A_i}	1:1
I_{B_i}	1:1
I_{B_i/S_i}	1:1 1:2 1:4 1:8 1:16 1:32
Number of instances	6

6.3.2 Heavy traffic

Next, we analyze the case as shown in Table 14, where we increase the total load as follows: $\rho = 0.75$, 0.8, 0.85, 0.9, 0.95 to study the effect of heavy traffic. We have to stress that also in the extensive test bed examined in the previous subsection we studied (many) heavy-traffic cases. However, in the present paragraph the system reaches saturation due to an increase in the traffic intensity, whereas in the previous subsection the system got saturated mainly due to the

Table 18 Relative errors for large setup times

Errors for large setup times												
	$I_{B_i/S_i} = 1:1$		$I_{B_i/S_i} = 1:2$		$I_{B_i/S_i} = 1:4$		$I_{B_i/S_i} = 1:8$		$I_{B_i/S_i} = 1:16$		$I_{B_i/S_i} = 1:32$	
	P	S	P	S	P	S	P	S	P	S	P	S
Mean queue lengths	2.6	11.0	3.0	8.2	3.3	6.4	3.6	5.5	3.6	4.8	3.9	4.7
SD queue lengths	2.5	11.7	3.4	10.2	4.7	9.7	6.3	9.9	7.5	10.1	8.5	10.5
0.90-quantile	0.0	33.3	0.0	0.0	0.0	14.3	7.7	7.7	8.3	8.3	6.5	6.5
0.95-quantile	0.0	0.0	0.0	0.0	11.1	11.1	6.7	6.7	7.1	10.7	7.4	9.3

magnitude of the setup times. The difference between these two regimes is enormous, which can be observed by comparing the rigorously proven results in [33] and [49] for the complete class of branching-type policies under Poisson arrival processes. That is, [33] studies the system under an increase of the traffic intensity, which shows that a *diffusion* limit applies and that the *gamma* distribution is prevalent, for example, in the scaled cycle lengths and the marginal queue lengths at polling instants. In contrast, [49] analyzes the effect of an increase of the setup times obtaining a *fluid* limit with a central role for the *deterministic* distribution revealing itself again, e.g., in the scaled cycle lengths and the marginal queue lengths at polling instants.

Let us now return to the k-limited policy of the present paper, for which none of these (asymptotic) results have been obtained. The results are depicted in Tables 15 and 16, which show that the accuracy of our approach significantly decreases as the total load increases. Moreover, it is not a difficult task to construct cases for which this decrease in accuracy is even more severe. The smaller accuracy in heavy traffic is due to the fact that the correlation between visit periods matters a lot in this regime; a small error in the approximation causes a snowball effect in the course of the iterations having an enormous effect on the final outcome. Intuitively, one would expect that the duration of the visit periods becomes degenerate for the k-limited policy in heavy traffic which may be exploited in an improvement of our algorithm in such a regime. Finally, our approach, in comparison with the standard approach, wins by a mile.

6.3.3 Large setup times

The present case studies the effect of large setup times on the accuracy of the approximation. Therefore, we perturb the ratio between service and setup times as follows: $I_{B_i/S_i} = 1:1, 1:2, 1:4, 1:8, 1:16, 1:32$, while the values of the other parameters are shown in Table 17. The results are summarized in Table 18, from which we can conclude that our approach remains accurate in the limit of increasing setup times. Moreover, we see that also the standard approach produces acceptable results indicating that the correlations between queues remain small as the setup times increase. In fact, [49] proves for branching-type policies that the system behaves as a deterministic system (with no correlations at all) in the limit of increasing deterministic setup times.

Table 19 Test bed for large number of queues

Test bed	
Parameter	Value(s)
N	2 4 8 16 32 64
ρ	0.6
k_i	5
$c^2_{A_i}$	1
$c^2_{B_i}$	1
$c^2_{S_i}$	1
I_{A_i}	1:1
I_{B_i}	1:1
I_{B_i/S_i}	1:1
Number of instances	6

6.3.4 Large number of queues

In this paragraph we examine whether the number of queues significantly affects the quality of the approximation. The detailed input parameters are provided in Table 19, but the most important feature is that we vary the number of queues: $N = 2, 4, 8, 16, 32, 64$. The results given in Table 20 show that the accuracy of both our approach and the standard approach increases as the number of queues in the system becomes larger. The reason for this is that an increase of the number of queues has a stabilizing effect on the cycle lengths, and thus also on the delays, which facilitates the approximation. For analytical results on polling systems in the limit of $N \to \infty$, where the total load and total setup times are held at fixed values, we refer to, e.g., [8, 21]. Note that in the limit the system has no distinction of different service disciplines since the load at each queue is infinitesimally small.

Table 20 Relative errors for large number of queues

Errors for large number of queues												
	$N=2$		$N=4$		$N=8$		$N=16$		$N=32$		$N=64$	
	P	S	P	S	P	S	P	S	P	S	P	S
Mean queue lengths	8.7	15.9	4.7	14.8	2.5	9.8	1.4	5.6	−0.8	2.3	−1.9	2.2
SD queue lengths	5.9	18.2	3.6	15.8	2.3	10.0	1.4	5.6	−0.5	2.6	−2.1	1.9
0.90-quantile	0.0	25.0	0.0	0.0	0.0	33.3	0.0	0.0	0.0	0.0	0.0	0.0
0.95-quantile	0.0	20.0	0.0	25.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

7 Conclusions

In the present paper, we have created a novel iterative approximation scheme for k-limited polling systems with general arrival, service and setup distributions to compute the complete queue length distributions. The multi-queue polling system has been decomposed into single-queue vacation systems with state-dependent vacations and k-limited service. We have analyzed this vacation model by means of matrix-analytic techniques under the assumption of general arrival, service and vacation processes. The main challenge was found in the computation of the correlations among the queues in each step of the iterative scheme. The accuracy of the approximation scheme has been validated via an extensive simulation study. The developed approximation turned out be accurate, robust and computationally efficient. As shown in Sect. 6.3, possible improvement of the algorithm may be obtained in heavy traffic and in cases with highly variable input. The numerical evaluation has shown that the algorithm converged relatively fast; a rigorous proof of convergence is, however, left as subject of further research.

With minor adjustments, the algorithm developed can be carried over to variants of the considered polling systems, e.g., systems with batch arrivals, discrete-time polling systems or systems with finite buffers. Application of our algorithm to polling systems with so-called gated-type k-limited service, i.e., the servers serves only k customers in a queue who arrived before the server's visit, is also not inconceivable. A related remark is that for deterministic service times the k-limited coincides with the time-limited strategy with fixed time limits, i.e., each queue has a time limit after which it relinquishes the server. By choosing service times with a negligible coefficient of variation as input, the algorithm of the present paper can also be used for the evaluation of this time-limited policy. Moreover, due to the efficiency of the algorithm, it could be used directly as approximation for the standard exhaustive and gated policy as well by choosing a 'large' value for the service limits. In that sense, our algorithm may be considered as extension of the procedure of [13] for exhaustive and gated polling systems, which relies on a Poisson assumption.

Finally, the algorithm of the present paper may be extended to the computation of *derivatives* of performance measures with respect to the service limits. Such an extensions would allow application of gradient methods to optimize systems performance and sensitivity analysis with respect to these control variables. Due to the low computational complexity of the developed procedure, it can be used as subroutine in such an optimization procedure.

Acknowledgement The authors would like to thank Onno Boxma for several helpful discussions.

Appendix 1

To obtain an approximating distribution of a positive random variable X, one may fit a phase-type distribution on the mean $\mathbb{E}[X]$ and the coefficient of variation c_X by using the following approach [43]. First of all, a random variable X is defined to have to a Coxian distribution of order k if it has to go through up to at most k exponential phases, where phase n has rate μ_n, $n=1,2,\ldots,k$. It starts in phase 1 and after phase n, $n=1,2,\ldots,k-1$, it ends with probability $1-p_n$, whereas it enters phase $n+1$ with probability p_n. Finally, p_k is defined to equal zero.

Now, the distribution of X is approximated as follows. If $c_X^2 > 1$, then the rate and coefficient of variation of the Coxian$_2$ distribution matches with $\mathbb{E}[X]$ and c_X, provided the parameters are chosen as (cf. [31]):

$$\mu_1 = 2/\mathbb{E}[X], \qquad p_1 = \frac{1}{2c_X^2}, \quad \text{and} \quad \mu_2 = p_1\mu_1.$$

If $1/k \le c_X^2 \le 1/(k-1)$ for some $k \ge 2$, then the rate and coefficient of variation of the Erlang$_{k-1,k}$ distribution, which is a special case of a Coxian distribution of order k, matches with $\mathbb{E}[X]$ and c_X, provided the parameters are chosen as (cf. [43]):

$$p_n = 1, \quad n = 1,2,\ldots,k-2,$$

$$p_{k-1} = 1 - \frac{kc_X^2 - \sqrt{k(1+c_X^2) - k^2c_X^2}}{1+c_X^2},$$

$$\mu_1 = \mu_2 = \cdots = \mu_k = (k-p)\mathbb{E}[X].$$

Of course, also other phase-type distributions may be fitted on the mean and the coefficient of variation, but numerical experiments suggest that choosing other distributions only has a minor effect on the results, as shown in [19].

Appendix 2

The transition rates of the G_0^A and G_1^A matrices as defined in Sect. 4 are given by

$$-\mu_i^A = G_0^A(i,i), \quad i = 1,2,\ldots,n_A,$$
$$p_i^A \mu_i^A = G_0^A(i,i+1), \quad i = 1,2,\ldots,n_A - 1,$$
$$(1-p_i^A)\mu_i^A = G_1^A(i,1), \quad i = 1,2,\ldots,n_A,$$

with p_i^A and μ_i^A the parameters of the fitted phase-type distributions for the arrival processes.

Subsequently, the transition rates for G_0^D and G_1^D as introduced in Sect. 4 are

$$-\mu_i^B = G_0^D(jn_B + i, jn_B + i),$$
$$j = 0,\ldots,k-1,\ i = 1,\ldots,n_B,$$
$$p_i^B \mu_i^B = G_0^D(jn_B + i, jn_B + i + 1),$$
$$j = 0,\ldots,k-1,\ i = 1,\ldots,n_B - 1,$$
$$(1-p_i^B)\mu_i^B = G_1^D(jn_B + i, (j+1)n_B + 1),$$
$$j = 0,\ldots,k-1,\ i = 1,\ldots,n_B,$$
$$-\mu_i^{\tilde{I}(k)} = G_0^D(kn_B + i, kn_B + i),$$
$$i = 1,\ldots,n_{\tilde{I}(k)},$$
$$p_i^{\tilde{I}(k)} \mu_i^{\tilde{I}(k)} = G_0^D(kn_B + i, kn_B + i + 1),$$
$$i = 1,\ldots,n_{\tilde{I}(k)} - 1,$$
$$(1-p_i^{\tilde{I}(k)})\mu_i^{\tilde{I}(k)} = G_0^D(kn_B + i, kn_B + n_{\tilde{I}(k)} + 1),$$
$$i = 1,\ldots,n_{\tilde{I}(k)},$$
$$-\mu_i^{\tilde{I}(*)} = G_0^D(kn_B + n_{\tilde{I}(k)} + i, kn_B + n_{\tilde{I}(k)} + i),$$
$$i = 1,\ldots,n_{\tilde{I}(*)},$$
$$p_i^{\tilde{I}(*)} \mu_i^{\tilde{I}(*)} = G_0^D(kn_B + n_{\tilde{I}(k)} + i, kn_B + n_{\tilde{I}(k)} + i + 1),$$
$$i = 1,\ldots,n_{\tilde{I}(*)} - 1,$$
$$(1-p_i^{\tilde{I}(*)})\mu_i^{\tilde{I}(*)} = G_0^D(kn_B + n_{\tilde{I}(k)} + i, kn_B + n_{\tilde{I}(k)} + n_{\tilde{I}(*)} + 1),$$
$$i = 1,\ldots,n_{\tilde{I}(*)},$$
$$-\mu_i^{I(0)} = G_0^D(kn_B + n_{\tilde{I}(k)} + n_{\tilde{I}(*)} + i, kn_B + n_{\tilde{I}(k)} + n_{\tilde{I}(*)} + i),$$
$$i = 1,\ldots,n_{I(0)},$$
$$p_i^{I(0)} \mu_i^{I(0)} = G_0^D(kn_B + n_{\tilde{I}(k)} + n_{\tilde{I}(*)} + i, kn_B + n_{\tilde{I}(k)} + n_{\tilde{I}(*)} + i + 1),$$
$$i = 1,\ldots,n_{I(0)} - 1,$$
$$(1-p_i^{I(0)})\mu_i^{I(0)} = G_0^D(kn_B + n_{\tilde{I}(k)} + n_{\tilde{I}(*)} + i, 1),$$
$$i = 1,\ldots,n_{I(0)},$$

and for $\tilde{G}_0^D$ and $\tilde{G}_1^D$ (see again Sect. 4) we have

$$(1-p_i^B)\mu_i^B = \tilde{G}_1^D(jn_B + i, kn_B + n_{\tilde{I}(k)} + 1),$$
$$j = 0,\ldots,k-2,\ i = 1,\ldots,n_B,$$
$$(1-p_i^B)\mu_i^B = \tilde{G}_1^D((k-1)n_B + i, kn_B + 1),$$
$$i = 1,\ldots,n_B,$$
$$-\mu_i^{\tilde{I}(k)} = \tilde{G}_0^D(kn_B + i, kn_B + i),$$
$$i = 1,\ldots,n_{\tilde{I}(k)},$$
$$p_i^{\tilde{I}(k)} \mu_i^{\tilde{I}(k)} = \tilde{G}_0^D(kn_B + i, kn_B + i + 1),$$
$$i = 1,\ldots,n_{\tilde{I}(k)} - 1,$$
$$(1-p_i^{\tilde{I}(k)})\mu_i^{\tilde{I}(k)} = \tilde{G}_0^D(kn_B + i, kn_B + n_{\tilde{I}(k)} + 1),$$
$$i = 1,\ldots,n_{\tilde{I}(k)},$$
$$-\mu_i^{\tilde{I}(*)} = \tilde{G}_0^D(kn_B + n_{\tilde{I}(k)} + i, kn_B + n_{\tilde{I}(k)} + i),$$
$$i = 1,\ldots,n_{\tilde{I}(*)},$$
$$p_i^{\tilde{I}(*)} \mu_i^{\tilde{I}(*)} = \tilde{G}_0^D(kn_B + n_{\tilde{I}(k)} + i, kn_B + n_{\tilde{I}(k)} + i + 1),$$
$$i = 1,\ldots,n_{\tilde{I}(*)} - 1,$$
$$(1-p_i^{\tilde{I}(*)})\mu_i^{\tilde{I}(*)} = \tilde{G}_0^D(kn_B + n_{\tilde{I}(k)} + i, kn_B + n_{\tilde{I}(k)} + n_{\tilde{I}(*)} + 1),$$
$$i = 1,\ldots,n_{\tilde{I}(*)},$$

$$-\mu_i^{I(0)} = \tilde{G}_0^D(kn_B + n_{\tilde{I}(k)} + n_{\tilde{I}(*)} + i, kn_B + n_{\tilde{I}(k)} + n_{\tilde{I}(*)} + i),$$
$$i = 1,\ldots,n_{I(0)},$$
$$p_i^{I(0)} \mu_i^{I(0)} = \tilde{G}_0^D(kn_B + n_{\tilde{I}(k)} + n_{\tilde{I}(*)} + i, kn_B + n_{\tilde{I}(k)} + n_{\tilde{I}(*)} + i + 1),$$
$$i = 1,\ldots,n_{I(0)} - 1,$$
$$(1-p_i^{I(0)})\mu_i^{I(0)} = \tilde{G}_0^D(kn_B + n_{\tilde{I}(k)} + n_{\tilde{I}(*)} + i, kn_B + n_{\tilde{I}(k)} + n_{\tilde{I}(*)} + 1),$$
$$i = 1,\ldots,n_{I(0)},$$

where p_i^B, $p_i^{\tilde{I}(k)}$, $p_i^{\tilde{I}(*)}$, $p_i^{I(0)}$, μ_i^B, $\mu_i^{\tilde{I}(k)}$, $\mu_i^{\tilde{I}(*)}$ and $\mu_i^{I(0)}$ are the parameters of the fitted phase-type distributions for the service and intervisit processes.

References

1. Asmussen S, Koole G. Marked point processes as limits of Markovian arrival streams. J Appl Probab 1993;30:365–72.
2. Blanc JPC. An algorithmic solution of polling models with limited service disciplines. IEEE Trans Commun 1992;40(7):1152–5.
3. Borst SC, Boxma OJ, Levy H. The use of service limits for efficient operation of multistation single-medium communication systems. IEEE/ACM Trans Netw 1995;3(5):602–12.
4. Boxma OJ. Workloads and waiting times in single-server systems with multiple customer classes. Queueing Syst 1989;5:185–214.
5. Chang KC, Sandhu D. Pseudo-conservation laws in cyclic-service systems with a class of limited service policies. Ann Oper Res 1992;35:209–29.
6. Chang KC, Sandhu D. Mean waiting time approximations in cyclic-service systems with exhaustive limited service policy. Perform Eval 1992;15(1):21–40.
7. Charzinski J, Renger T, Tangemann M. Simulative comparison of the waiting time distributions in cyclic polling system with different service strategies. In: Proceedings of the 14th international teletraffic congress. Antibes Juan-les-Pins; 1994. p. 719–728.
8. Coffman EG Jr, Gilbert EN. A continuous polling system with constant service times. IEEE Trans Inf Theory 1986;32(4):584–91.
9. Dallery Y, David R, Xie X. Approximate analysis of transfer lines with unreliable machines and finite buffers. IEEE Trans Autom Control 1989;34(9):943–53.
10. Doshi BT. Queueing systems with vacations—a survey. Queueing Syst 1986;1(1):29–66.
11. Doshi BT. Single server queues with vacations. In: Takagi H, editors. Stochastic analysis of computer and communication systems. Amsterdam: North-Holland; 1990. p. 217–65.
12. Everitt D. A note on the pseudoconservation laws for cyclic service systems with limited service disciplines. IEEE Trans Commun 1989;37(7):781–3.
13. Federgruen A, Katalan Z. Approximating queue size and waiting time distributions in general polling systems. Queueing Syst 1994;18:353–86.
14. Federgruen A, Katalan Z. The stochastic economic lot scheduling problem: cyclical base-stock policies with idle times. Manag Sci 1996;42(6):783–96.
15. Federgruen A, Katalan Z. Determining production schedules under base-stock policies in single facility multi-item production systems. Oper Res 1998;46(6):883–98.
16. Fricker C, Jaibi R. Monotonicity and stability of periodic polling models. Queueing Syst 1994;15:211–38.
17. Fuhrmann SW. Performance analysis of a class of cyclic schedules. Bell laboratories technical memorandum 81-59531-1; 1981.
18. Gershwin SB, Burman MH. A decomposition method for analyzing inhomogeneous assembly/disassembly systems. Ann Oper Res 2000;93:91–115.
19. Johnson MA. An empirical study of queueing approximations based on phase-type distributions. Stoch Models 1993;9(4):531–61.
20. Keilson J, Servi LD. The distributional form of Little's law and the Fuhrmann-Cooper decomposition. Oper Res Lett 1990;9(4):239–47.
21. Kroese DP, Schmidt V. A continuous polling system with general service times. Ann Appl Probab 1992;2:906–27.
22. Kühn PJ. Multiqueue systems with nonexhaustive cyclic service. Bell Syst Tech J 1979;58(3):671–98.
23. Lang M, Bosch M. Performance analysis of finite capacity polling systems with limited-M service. In: Proceedings of the 13th international teletraffic congress. Copenhagen; 1991. p. 731–735.
24. Lee D-S, Sengupta B. An approximate analysis of a cyclic server queue with limited service and reservations. Queueing Syst 1992;11:153–78.
25. Lee D-S. A two-queue model with exhaustive and limited service disciplines. Stoch Models 1996;12(2):285–305.
26. Leung KK. Cyclic-service systems with probabilistically-limited service. IEEE J Sel Areas Commun 1991;9(2):185–93.
27. Levy Y. A class of scheduling policies for real-time processors with switching system applications. In: Proceedings of the 11th international teletraffic congress. Yokohama; 1985. p. 760–766.
28. Levy H, Sidi M. Polling systems: applications, modeling and optimization. IEEE Trans Commun 1990;COM-38(10):1750–60.
29. Mack C, Murphy T, Webb NL. The efficiency of N machines unidirectionally patrolled by one operative when walking time and repair times are constants. J Roy Stat Soc Ser B 1957;19(1):166–72.
30. Mack C. The efficiency of N machines uni-directionally patrolled by one operative when walking time is constant and repair times are variable. J Roy Stat Soc Ser B 1957;19(1):173–8.
31. Marie RA. Calculating equilibrium probabilities for $\lambda(n)/C_k/1/N$ queue. In: Proceedings Performance '80. Toronto; 1980. p. 117–125.
32. van der Mei RD, Winands EMM. Mean value analysis for polling systems in heavy traffic. In: Proceedings Valuetools. Pisa: ACM Press; 2006.
33. van der Mei RD. Towards a unifying theory on branching-type polling models in heavy traffic. Report. Free University; 2006.
34. Naoumov VA, Krieger UR, Wagner D. Analysis of a multiserver delay-loss system with a general markovian arrival process. In: Alfa AS, Chakravarthy SR, editors. Matrix-analytic methods in stochastic models. New York: Dekker; 1997. p. 43–66.
35. Neuts MF. Matrix-geometric solutions in stochastic models, an algorithmic approach. Baltimore: Johns Hopkins Press; 1981.
36. Ozawa T. Alternating service queues with mixed exhaustive and K-limited services. Perform Eval 1990;11:165–75.
37. Ozawa T. Waiting time distribution in a two-queue model with mixed exhaustive and gated-type K-limit service. In: Proceedings of international conference on the performance and management of complex communication networks. Tsukuba; 1997. p. 231–250.
38. Resing JAC. Polling systems and multitype branching processes. Queueing Syst 1993;13:409–26.
39. Takagi H. Queueing analysis of polling models: an update. In: Takagi H, editor. Stochastic analysis of computer and communication systems. Amsterdam: North-Holland; 1990. p. 267–318.
40. Takagi H. Queueing analysis: a foundation of performance evaluation, vacation and priority systems, part 1. Amsterdam: North-Holland; 1991.
41. Takagi H. Queueing analysis of polling models: progress in 1990–1994. In: Dshalalow JH, editor. Frontiers in queueing: models, methods and problems. Boca Raton: CRC Press; 1997. p. 119–46.
42. Takagi H. Analysis and application of polling models. In: Haring G, Lindemann C, Reiser M, editors. Performance evaluation: origins and directions. Lecture notes in computer science, vol 1769. Berlin: Springer; 2000. p. 423–42.
43. Tijms HC. Stochastic models: an algorithmic approach. Chichester: Wiley; 1994.
44. van Vuuren M, Adan IJBF, Resing-Sassen SA. Performance analysis of multi-server tandem queues with finite buffers and blocking. OR Spectr 2005;27(2–3):315–38.

45. van Vuuren M, Adan IJBF. Performance analysis of assembly systems. In: Proceedings of the Markov anniversary meeting. Charleston; 2006. p. 89–100.
46. Winands EMM, Adan IJBF, van Houtum GJ. The stochastic economic lot scheduling problem: a survey. Eindhoven: BETA WP-133, Beta Research School for Operations Management and Logistics; 2005.
47. Winands EMM, Adan IJBF, van Houtum GJ. A two-queue model with alternating limited service and state-dependent setups. In: Proceedings of analysis of manufacturing systems–production management. Zakynthos; 2005. p. 200–208.
48. Winands EMM, Adan IJBF, van Houtum GJ. Mean value analysis for polling systems. Queueing Syst 2006;54(1):45–54.
49. Winands EMM. Branching-type polling systems with large setups. Report. Eindhoven University of Technology; 2006.
50. Zipkin PH. Foundations of inventory management. London: McGraw–Hill; 2000.