Heavy traffic analysis of polling models by mean value analysis

Heavy traffic analysis of polling models by Mean Value Analysis

R.D. van der Mei[1,2] and E.M.M. Winands[3,4]

[1]Department of Mathematics, Vrije Universiteit
1081 HV Amsterdam, The Netherlands

[2]Centre for Mathematics and Computer Science (CWI)
1098 SJ Amsterdam, The Netherlands
mei@cwi.nl

[3]Department of Mathematics and Computer Science
[4]Department of Technology Management
Technische Universiteit Eindhoven
P.O. Box 513, 5600 MB Eindhoven, The Netherlands
e.m.m.winands@tue.nl

December 4, 2006

Abstract

In this paper we present a new approach to derive heavy-traffic asymptotics for polling models. We consider the classical cyclic polling model with exhaustive or gated service at each queue, and with general service-time and switch-over time distributions, and study its behavior when the load tends to one. For this model, we explore the recently proposed mean value analysis (MVA), which takes a new view on the dynamics of the system, and use this view to provide an alternative way to derive closed-from expressions for the expected asymptotic delay; the expressions were derived earlier in [31], but in a different way. Moreover, the MVA-based approach enables us to derive closed-form expressions for the heavy-traffic limits of the covariances between the successive visit periods, which are key performance metrics in many application areas. These results, which have not been obtained before, reveal a number of insensitivity properties of the covariances with respect to the system parameters under heavy-traffic assumptions, and moreover, lead to simple approximations for the covariances between the successive visit times for stable systems. Numerical examples demonstrate that the approximations are accurate when the load is close enough to one.

Keywords: polling systems, mean value analysis, heavy traffic, delay, visit time.

1 Introduction

A polling system is a multi-queue single-server system in which the server visits the queues in some order to process requests pending at the queues. Polling systems occur naturally in the modeling of systems in which service capacity (e.g., CPU, bandwidth, processing power, labor) is shared by different types of users, each type having specific traffic characteristics and performance requirements. Polling systems find many applications in the areas of computer-communication networks, production, manufacturing and maintenance, see [16] for an overview. Since the late 1960s polling systems have received much attention in the literature, see [27, 28, 29] for overviews of the available results. Exact analysis of the delay in polling systems is only possible in some cases, and even in those cases numerical techniques are usually required to obtain the expected delay at each of the queues. However, the use of numerical techniques for the analysis of polling systems has several drawbacks. Firstly, numerical techniques do not reveal explicitly how the system performance depends on the system parameters and can therefore contribute to the understanding of the system behavior only to a limited extent. Exact closed-form expressions provide much more insight into the dependence of the performance measures on the system parameters, which leads to significant insights in the behavior of the system (e.g., insensitivity and monotonicity properties). Secondly, the efficiency of the numerical algorithms tends to degrade significantly for heavily loaded, highly asymmetric systems with a large number of queues, while the proper operation of the system is particularly critical when the system is heavily loaded. These observations raise the importance of an exact asymptotic analysis of the delay in polling systems in heavy traffic.

The literature reveals a remarkable difference in the complexity of different polling systems. Resing [22] shows that for a class of polling systems the joint queue-length process embedded at polling instants at a fixed queue constitutes a multi-type branching process (MTBP) with immigration. The theory of MTBPs leads to expressions for the generating function of the joint queue-length distribution at polling instants. For polling systems with an MTBP-structure several numerical algorithms have been proposed to determine the (first few) moments of the delay at the queues by solving sets of linear equations (cf., e.g., [9, 23, 25]). Alternatively, Konheim *et al.* [12] propose the so-called descendant set approach (DSA), an iterative technique that exploits the MTBP-structure of the model by making use of the concept of descendant sets, to obtain the moments of the delay. Choudhury and Whitt [4] use numerical transform-inversion to calculate tail probabilities and transient performance measures of the delay. A common aspect of each of these approaches is that the derivation of delay distributions (or moments) is basically decomposed into two phases: (1) derivation of distribution of the number of customers at a polling instant of the server at a fixed queue, and (2) derivation of the relation between this distribution and the distribution of the delay. As an interesting alternative to these classical approaches, Winands *et al.* [40] suggest a completely different view at the system, and propose the so-called *mean value analysis* (MVA) approach for obtaining the mean delay at each of the queues via direct mean value arguments, avoiding the need to determine the queue-length distributions at polling instants. The basic idea of the MVA approach is to identify a set of linear equations for the variables $\mathbb{E}[L_{i,j}]$, where $L_{i,j}$ is the length of queue i at an arbitrary epoch within a visit time at queue j (see Section 4 for details). Polling systems that do not have an MTBP-structure generally require much more computational effort (cf., e.g., [2, 15]).

There are several approaches for deriving heavy traffic asymptotics in polling systems. Coffman *et al.* [5, 6] use a heavy-traffic averaging principle to study a two-queue model with exhaustive service at both queues and show that, under heavy-traffic assumptions and scalings, the total amount of unfinished work converges to a known process. These observations lead to explicit expressions for the moments of the delay at both queues. They also suggest that, based on a partial conjecture, the analysis can be extended to systems with more than two queues. Exploring the averaging principle, Reiman and Wein [21] and Markowitz *et al.* [17, 18] study the problem of determining optimal dynamic schedules by approximating the dynamic scheduling problems by diffusion control problems. Kudoh *et al.* [14] use the classical buffer-occupancy technique,

which is based on an expression for the probability generating function of the joint queue-length distribution at successive polling instants, to derive explicit expressions for the second moment of the delay in fully symmetric systems with gated or exhaustive service at each queue for models with two, three and four queues. They also give conjectures about the heavy-traffic limits of the first two moments of the delay for systems with an arbitrary number of queues. Kroese [13] uses the theory of age-dependent branching processes to study the heavy-traffic behavior of continuous polling systems and shows that the steady-state number of waiting customers has approximately a gamma-distribution. Van der Mei and co-authors [19, 20, 31, 32, 33, 34] explore the recursive relations of the DSA to derive closed-form expressions for the asymptotic delay distribution in heavy traffic for polling systems with an MTBP-structure in a general parameter setting, both for cyclic and periodic server routing. As a by-product, the asymptotic results presented in these papers also suggest simple approximations for the delay distributions for stable systems. Numerical results demonstrate that the accuracy of the approximations is remarkably good, particularly for the mean delay figures.

In the present paper we consider the classical asymmetric cyclic polling system with generally distributed service times and switch-over times (with finite first two moments), and with exhaustive or gated service at each of the queues (see also Remark 5.5). For this model, we propose a new technique to derive heavy-traffic limits, both for the expected delay at each of the queues and for the covariances between successive visits of the server to the queues. To this end, we use the recently proposed MVA-technique [40] as the starting point, which leads to a set of linear equations for the unknowns $\mathbb{E}[L_{i,j}]$ $(i, j = 1, \ldots, N)$, where N is the number of queues; recall that $L_{i,j}$ is the length of queue i at an arbitrary epoch within a visit time at queue j. Taking the proper heavy-traffic limits of this set, we obtain a highly simplified but dependent set of linear equations that determines the heavy-traffic limits of $\mathbb{E}[L_{i,j}]$ up to a scaling constant. Finally, the scaling constant is obtained by adding a linear equation that follows from the pseudo-conservation law of the system [3]. These results not only provide a new means to obtain heavy-traffic asymptotics for the expected delay (which were shown earlier in [31] based on the use of the DSA), but also lead to the observation that the correlations between successive visit times converge to one as the load tends to one. The latter observation gives rise to asymptotic expressions for the covariance between successive visit times. The expressions naturally lead to approximate closed-form expressions for the covariances between the visit times for stable systems (i.e., with load less than one), allowing for back-of-the-envelope calculations. Numerical results show that the approximations are accurate when the total system load is close to one.

We believe that the importance of the mean delay as a performance metric requires no further explanation, the importance of the correlation terms possibly does. The motivation for the analysis of these terms stems from the fact that polling systems can be employed to study widely used, cyclic base-stock policies in multi-product production-inventory systems (details are given in Section 6). These systems are frequently encountered in process industries, where very high levels of utilization are prevalent as the production installations are typically very expensive (see, e.g., Fransoo [10]). The asymptotic results derived for the correlations of successive visit times indicate that base-stock policies display undesirable behavior if the utilization rate is high, which reveals itself, for example, in difficulties in the coordination between stages within the production process.

The contribution of the present paper is three-fold. First, we present a new approach to derive heavy-traffic asymptotics for a class of polling models, based on the recently introduced MVA technique [40]. Second, we derive heavy-traffic asymptotics for the correlation and covariances between successive visit times, leading to new insights in the impact of system parameters on the correlations between the visits times in heavy traffic. Third, we use these results to propose simple closed-form approximations of these covariances for stable systems, and show that these approximations, which allow for back-of-the-envelope calculations, are accurate when the load is high enough. These observations make the contribution of the present paper evident, both from a queueing-theoretical and application point-of-view.

The remainder of this paper is organized as follows. Section 2 presents the model, while Section 3 discusses the performance measures of interest. Sections 4 and 5 analyze the MVA equations in heavy traffic and obtain closed-form expressions for the mean asymptotic delay and the covariances between successive visit times for exhaustive and gated service, respectively. Next, in Section 6, we specifically consider the multi-item production-inventory system, which motivated our research. Some topics for further research are addressed in Section 7.

2 Model description and notation

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

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

Throughout the present paper, it is assumed that the service discipline within a queue is *First Come First Served* (FCFS). Obviously, the mean waiting times are the same under any work-conserving non-preemptive service discipline that does not account for the actual service requests of the customers. Customers arrive at all queues according to independent Poisson processes with rates λ_i, $i = 1, 2, \ldots, N$, where the total arrival rate is denoted by $\Lambda = \sum_{i=1}^{N} \lambda_i$. The service times at queue i are independent, identically distributed random variables with mean $\mathbb{E}[B_i]$ and finite second moment $\mathbb{E}[B_i^2]$, $i = 1, 2, \ldots, N$. The first two moments of the service time of an arbitrary customer are given by, respectively,

$$\mathbb{E}[B] = \sum_{i=1}^{N} \frac{\lambda_i \mathbb{E}[B_i]}{\Lambda}, \qquad \mathbb{E}[B^2] = \sum_{i=1}^{N} \frac{\lambda_i \mathbb{E}[B_i^2]}{\Lambda}.$$

When the server starts service at queue i, a setup time is incurred of which the first and second moment are denoted by $\mathbb{E}[S_i]$ and $\mathbb{E}[S_i^2]$, $i = 1, 2, \ldots, N$, respectively. The mean and the variance of the total setup time in a cycle are given by, respectively,

$$\mathbb{E}[S] = \sum_{i=1}^{N} \mathbb{E}[S_i], \qquad \mathbb{V}\mathrm{ar}[S] = \sum_{i=1}^{N} (\mathbb{E}[S_i^2] - \mathbb{E}[S_i]^2).$$

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

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

The occupation rate ρ_i (excluding setups) at queue i is defined by $\rho_i = \lambda_i \mathbb{E}[B_i]$ and the total occupation rate ρ is given by $\rho = \sum_{i=1}^{N} \rho_i$. A necessary and sufficient condition for the stability of this polling system is obviously $\rho < 1$ (see, e.g., [11] for a rigorous proof)). In the remainder of the present paper, this stability condition is assumed to hold as we restrict ourselves to steady-state behavior.

The cycle length 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 $\mathbb{E}[C] = \frac{\mathbb{E}[S]}{1-\rho}$ (see, e.g., [27]). The present paper focusses mainly on the case $\mathbb{E}[S] > 0$, but our analysis fully applies in case the total setup time is equal to zero, *mutatis mutandis* (see, also, [40]). The visit time θ_i of queue i, $i = 1, 2, \ldots, N$, is composed of the service period of queue i, the time the server spends servicing customers at queue i, plus the *preceding* setup time in case of exhaustive service or plus the *succeeding* setup time in case of gated service. By virtue of these two different definitions, a queue is empty exactly at the end of its visit time in case of exhaustive service, while the queue before the gate is empty at the beginning of a visit time in case of gated service (all customers waiting for service are then placed behind the gate). Since the server is working a fraction ρ_i of the time on queue i, the mean of a visit period of queue i reads, for exhaustive service,

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

and, for gated service,

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

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

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

Notice that in case $j = 1$ and $j = N$, $\mathbb{E}[\theta_{i,j}]$ is equal to the mean visit period $\mathbb{E}[\theta_i]$ of queue i and the mean cycle length $\mathbb{E}[C]$, respectively. The fraction of the time $q_{i,j}$ the system is in an (i, j)-period equals $q_{i,j} = \frac{\mathbb{E}[\theta_{i,j}]}{\mathbb{E}[C]}$. Moreover, the mean of a residual (i, j)-period is given by

$$\mathbb{E}[R_{\theta_{i,j}}] = \frac{\mathbb{E}[\theta_{i,j}^2]}{2\mathbb{E}[\theta_{i,j}]}, \qquad i = 1, 2, \ldots, N, \qquad j = 1, 2, \ldots, N,$$

with the remark that the second moments $\mathbb{E}[\theta_{i,j}^2]$ are still unknown at this stage.

3 Performance measures

Delay: Let us consider the mean delay $\mathbb{E}[W_i]$ of a type-i customer, $i = 1, 2, \ldots, N$, which is defined as the time in steady state from a customer's arrival at queue i until the start of his service. By Little's Law, these mean delays are obviously related to the mean queue lengths (excluding the customer possibly in service) $\mathbb{E}[L_i]$, $i = 1, 2, \ldots, N$. To derive these quantities, we use the MVA approach [40], which renders a set of equations for the $\mathbb{E}[L_{i,n}]$, the mean queue length at queue i at an arbitrary epoch within a visit time of queue n, $i, n = 1, 2, \ldots, N$. The corresponding mean delay $\mathbb{E}[W_i]$ can be expressed in terms of $\mathbb{E}[L_{i,n}]$ as follows

$$\mathbb{E}[W_i] = \frac{1}{\lambda_i}\mathbb{E}[L_i] = \frac{1}{\lambda_i}\sum_{n=1}^{N} q_{n,1}\mathbb{E}[L_{i,n}], \qquad i = 1, 2, \ldots, N. \tag{2}$$

For further reference, we state the pseudo-conservation law for systems with exhaustive or gated service at each of the queue (see, e.g., [3])

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

where G stands for the index set of queues with gated service. Throughout the present paper $\mathbb{E}[W_i]$ is considered as function of ρ, where the arrival rates are variable, while the service time distributions and the ratios of the arrival rates are fixed. In case $\rho \uparrow 1$, all queues become unstable and, thus, $\mathbb{E}[W_i]$ tends to infinity for all i. To be precise, $\mathbb{E}[W_i]$ has a first-order pole at $\rho = 1$ (cf. [31]),

$$\mathbb{E}[W_i] = \frac{\mathbb{E}[W_i^*]}{1-\rho} + o((1-\rho)^{-1}), \qquad \rho \uparrow 1, \quad i = 1, 2, \ldots, N,$$

where $g(x) = o(f(x))$ means that $g(x)/f(x) \to 0$ as $x \uparrow 1$. The analysis of the present paper is oriented towards the determination of a closed-form expression for

$$\mathbb{E}[W_i^*] = \lim_{\rho \uparrow 1} \ (1-\rho)\mathbb{E}[W_i], \qquad i = 1, 2, \ldots, N,$$

which is referred to as the *mean asymptotic scaled delay* at queue i. Finally, the fact that $\mathbb{E}[W_i]$, $i = 1, 2, \ldots, N$, has a first-order pole at $\rho = 1$ implies that $\mathbb{E}[L_{i,n}]$, $i, n = 1, 2, \ldots, N$, has a first-order pole at $\rho = 1$ as well. Therefore, the following limits are well-defined,

$$\mathbb{E}[L_{i,n}^*] = \lim_{\rho \uparrow 1} \ (1-\rho)\mathbb{E}[L_{i,n}], \qquad i = 1, 2, \ldots, N, \quad n = 1, 2, \ldots, N.$$

For the validity of the statement that $\mathbb{E}[W_i]$ has a first-order pole at $\rho = 1$, we refer to [32, 34]. In fact, the results of these papers show that the k^{th} moment of the delay has a k^{th} order pole at $\rho = 1$ for $k = 1, 2, \ldots$.

Correlations: The set of MVA equations can also be applied for the computation of the mean of residual (i, j)-periods $\mathbb{E}[R_{\theta_{i,j}}]$. Thereupon, the variance of an (i, j)-period can be obtained via

$$\mathbb{V}\mathrm{ar}[\theta_{i,j}] = 2\mathbb{E}[R_{\theta_{i,j}}]\mathbb{E}[\theta_{i,j}] - \mathbb{E}[\theta_{i,j}]^2, \qquad i = 1, 2, \ldots, N, \quad j = 1, 2, \ldots, N, \tag{4}$$

where $\mathbb{E}[\theta_{i,j}]$ is given by (1). From these variances it also possible to compute the covariance $\mathbb{C}\mathrm{ov}[\theta_i, \theta_{i+n}]$ and correlation $\mathbb{C}\mathrm{or}[\theta_i, \theta_{i+n}]$ of the visit periods θ_i and θ_{i+n} via the following lemma.

Lemma 3.1. *Given random variables X, Y and Z, the covariance of X and Z can be obtained as follows*

$$\mathbb{C}ov[X, Z] = \frac{1}{2}(\mathbb{V}ar[X + Y + Z] - \mathbb{V}ar[X + Y] - \mathbb{V}ar[Y + Z] + \mathbb{V}ar[Y]).$$

Proof By definition,

$$\mathbb{V}\mathrm{ar}[X + Y + Z] = \mathbb{V}\mathrm{ar}[X] + \mathbb{V}\mathrm{ar}[Y] + \mathbb{V}\mathrm{ar}[Z] + 2(\mathbb{C}\mathrm{ov}[X, Y] + \mathbb{C}\mathrm{ov}[Y, Z] + \mathbb{C}\mathrm{ov}[X, Z]),$$

and

$$\begin{aligned} \mathbb{V}\mathrm{ar}[X + Y] &= \mathbb{V}\mathrm{ar}[X] + \mathbb{V}\mathrm{ar}[Y] + 2\mathbb{C}\mathrm{ov}[X, Y], \\ \mathbb{V}\mathrm{ar}[Y + Z] &= \mathbb{V}\mathrm{ar}[Y] + \mathbb{V}\mathrm{ar}[Z] + 2\mathbb{C}\mathrm{ov}[Y, Z]. \end{aligned}$$

which, after some rewriting, completes the proof. □

Using the result of Lemma 3.1, one may verify that for $i = 1, 2, \ldots, N$ and $n = 1, 2, \ldots, N-1$,

$$\mathbb{C}\mathrm{ov}[\theta_i, \theta_{i+n}] = \frac{1}{2}(\mathbb{V}\mathrm{ar}[\theta_{i,n+1}] - \mathbb{V}\mathrm{ar}[\theta_{i,n}] - \mathbb{V}\mathrm{ar}[\theta_{i+1,n}] + \mathbb{V}\mathrm{ar}[\theta_{i+1,n-1}]), \tag{5}$$

where $\mathbb{V}\mathrm{ar}[\theta_{i+1,0}] = 0$. Thus, we have for the correlation

$$\mathbb{C}\mathrm{or}[\theta_i, \theta_{i+n}] = \frac{\mathbb{C}\mathrm{ov}[\theta_i, \theta_{i+n}]}{\sqrt{\mathbb{V}\mathrm{ar}[\theta_i]\mathbb{V}\mathrm{ar}[\theta_{i+n}]}}, \qquad i = 1, 2, \ldots, N, \quad n = 1, 2, \ldots, N-1. \tag{6}$$

Similar to the mean delay $\mathbb{E}[W_i]$, the above performance metrics are considered as a function of ρ, where the arrival rates are variable. Finally, throughout the present paper, for each variable x that is a function of ρ, its value evaluated at $\rho = 1$ is denoted by $\hat{x}$.

4 Exhaustive service

In this section we explore the use of the MVA to derive heavy-traffic asymptotics for the model with exhaustive service.

4.1 Mean value analysis

Starting proof of our analysis is the following set of equations for the unknown mean queue lengths $\mathbb{E}[L_{i,n}]$, $i = 1, 2, \ldots, N$, $j = 1, \ldots, N-1$, (see [40]),

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

$$\sum_{n=i+1}^{i+j} \frac{q_{n,1}}{q_{i+1,j}}\mathbb{E}[L_{i,n}] = \lambda_i\mathbb{E}[R_{\theta_{i+1,j}}], \quad (8)$$

where the unknown $\mathbb{E}[R_{\theta_{i,j}}]$ are expressed recursively in terms of $\mathbb{E}[L_{i,n}]$ as follows, $i = 1, 2, \ldots, N$, $j = 2, 3, \ldots, N-1$,

$$\mathbb{E}[R_{\theta_{i,1}}] = \frac{1}{1-\rho_i}\left(\mathbb{E}[L_{i,i}]\mathbb{E}[B_i] + \frac{\rho_i\mathbb{E}[C]}{\mathbb{E}[\theta_{i,1}]}\mathbb{E}[R_{B_i}] + \frac{\mathbb{E}[S_i]}{\mathbb{E}[\theta_{i,1}]}\mathbb{E}[R_{S_i}]\right), \quad (9)$$

$$\mathbb{E}[R_{\theta_{i,j}}] = \frac{q_{i,1}}{q_{i,j}}\left(\frac{\mathbb{E}[R_{\theta_{i,1}}]}{\prod_{n=1}^{j-1}(1-\rho_{i+n})} + \sum_{n=1}^{j-1}\frac{\mathbb{E}[S_{i+n}] + \mathbb{E}[L_{i+n,i}]\mathbb{E}[B_{i+n}]}{\prod_{m=n}^{j-1}(1-\rho_{i+m})}\right) + (1 - \frac{q_{i,1}}{q_{i,j}})\mathbb{E}[R_{\theta_{i+1,j-1}}]. \quad (10)$$

In [40] it is remarked that the residual cycle lengths $\mathbb{E}[R_{\theta_{i,N}}]$, $i = 1, 2, \ldots, N$, which are not required for the computation of the mean delays, satisfy (10) as well. The set (7) - (10) can in general not be solved in closed-form, but in the remainder of the present section an explicit solution is derived in the limit of $\rho \uparrow 1$. Multiplying both sides of (7) - (10) by $(1-\rho)$ and letting $\rho \uparrow 1$ renders the corresponding set of equations in heavy traffic for $i = 1, 2, \ldots, N$ and $j = 1, 2, \ldots, N-1$,

$$\sum_{n=1}^{N} \hat{\rho}_n\mathbb{E}[L^*_{i,n}] = \hat{\lambda}_i\mathbb{E}[R_{\theta^*_{i+1,N-1}}], \quad (11)$$

$$\frac{\sum_{n=i+1}^{i+j}\hat{\rho}_n\mathbb{E}[L^*_{i,n}]}{\sum_{m=i+1}^{i+j}\hat{\rho}_m} = \hat{\lambda}_i\mathbb{E}[R_{\theta^*_{i+1,j}}], \quad (12)$$

with for $i = 1, 2, \ldots, N$ and $j = 2, 3, \ldots, N$,

$$\mathbb{E}[R_{\theta^*_{i,1}}] = \frac{1}{1-\hat{\rho}_i}\mathbb{E}[L^*_{i,i}]\mathbb{E}[B_i], \quad (13)$$

$$\mathbb{E}[R_{\theta^*_{i,j}}] = \frac{\hat{\rho}_i}{\sum_{n=i}^{i+j-1}\hat{\rho}_n}\left(\frac{\mathbb{E}[R_{\theta^*_{i,1}}]}{\prod_{n=1}^{j-1}(1-\hat{\rho}_{i+n})} + \sum_{n=1}^{j-1}\frac{\mathbb{E}[L^*_{i+n,i}]\mathbb{E}[B_{i+n}]}{\prod_{m=n}^{j-1}(1-\hat{\rho}_{i+m})}\right) + \frac{\sum_{n=i+1}^{i+j-1}\hat{\rho}_n}{\sum_{n=i}^{i+j-1}\hat{\rho}_n}\mathbb{E}[R_{\theta^*_{i+1,j-1}}]. \quad (14)$$

The variables $\mathbb{E}[R_{\theta^*_{i,j}}]$ are defined by

$$\mathbb{E}[R_{\theta^*_{i,j}}] = \lim_{\rho\uparrow 1}\ (1-\rho)\mathbb{E}[R_{\theta_{i,j}}], \qquad i = 1, 2, \ldots, N, \quad j = 1, 2, \ldots, N. \quad (15)$$

The fact that $\mathbb{E}[W_i]$, $i = 1, 2, \ldots, N$, has a first-order pole at $\rho = 1$ implies that $\mathbb{E}[R_{\theta_{i,j}}]$, $j = 1, 2, \ldots, N$, also has a first-order pole at $\rho = 1$ and thus the limits in (15) are well-defined.

The recursion represented by (13) - (14) has a closed-form solution in terms of $\mathbb{E}[L^*_{i,n}]$, which reads for $i = 1, 2, \ldots, N$ and $j = 1, 2, \ldots, N$,

$$\mathbb{E}[R_{\theta^*_{i,j}}] = \frac{1}{\sum_{n=i}^{i+j-1}\hat{\rho}_n}\left(\sum_{n=0}^{j-1}\mathbb{E}[B_{i+n}]\frac{\sum_{m=0}^{n}\hat{\rho}_{i+m}\mathbb{E}[L^*_{i+n,i+m}]}{\prod_{m=n}^{j-1}(1-\hat{\rho}_{i+m})}\right). \quad (16)$$

By substitution of (16) into (11) - (12), the following set of N^2 equations for equally many unknowns $\mathbb{E}[L^*_{i,n}]$ is obtained: For $i = 1, 2, \ldots, N$,

$$\sum_{n=1}^{N} \hat{\rho}_n \mathbb{E}[L^*_{i,n}] = \mathbb{E}[L^*_{i,i}], \tag{17}$$

and for $i = 1, 2, \ldots, N,\ j = 1, 2, \ldots, N-1$,

$$\sum_{n=i+1}^{i+j} \hat{\rho}_n \mathbb{E}[L^*_{i,n}] = \hat{\lambda}_i \left(\sum_{n=0}^{j-1} \mathbb{E}[B_{i+1+n}] \frac{\sum_{m=0}^{n} \hat{\rho}_{i+1+m} \mathbb{E}[L^*_{i+1+n,i+1+m}]}{\prod_{m=n}^{j-1}(1-\hat{\rho}_{i+1+m})} \right). \tag{18}$$

The set (17) - (18) can be solved up to some unknown scaling factor $c \in \mathbb{R}$ as shown in the following theorem.

Theorem 4.1. *The solution of the set* (17) - (18) *is given by*

$$\mathbb{E}[L^*_{i,i}] = c\hat{\lambda}_i(1-\hat{\rho}_i), \qquad i = 1, 2, \ldots, N, \tag{19}$$

$$\mathbb{E}[L^*_{i,i+n}] = c\hat{\lambda}_i(2\sum_{m=1}^{n-1} \hat{\rho}_{i+m} + \hat{\rho}_{i+n}), \qquad i = 1, 2, \ldots, N, \quad n = 1, 2, \ldots, N-1, \tag{20}$$

with $c \in \mathbb{R}$.

Proof: By Cramer's rule we know that the *homogeneous* set (17) - (18) has (an infinite number of) non-degenerate solutions if and only if the determinant of the coefficient matrix vanishes. This can be ascertained by elementary, but tedious, row and column operations, which we leave to the reader. Finally, the final row reduced form of the coefficient matrix shows that the rank of the matrix equals $N^2 - 1$, which completes the proof. □

Since the dimension of the null space of the coefficient matrix of (17) - (18) equals one, adding a single *non-homogeneous* equation would render a unique solution the unknown scaling factor c. This additional equation can be readily obtained from a scaled version of the pseudo-conservation law (3) as done in the theorem below.

Theorem 4.2. *The quantity* c *is given by*

$$c = \frac{1 + \beta\delta\mathbb{E}[S]}{2\beta\delta},$$

where $\beta = \frac{\mathbb{E}[B]}{\mathbb{E}[B^2]}$ *and* $\delta = 1 - \sum_{i=1}^{N} \hat{\rho}_i^2$.

Proof: Via Theorem 4.1 and (2), one obtains the unconditional mean asymptotic scaled delays,

$$\mathbb{E}[W^*_i] = c(1-\hat{\rho}_i), \qquad i = 1, 2, \ldots, N, \tag{21}$$

which satisfies a scaled version of pseudo-conservation law (3). That is, multiplying both sides of (3) by $(1-\rho)$ and letting $\rho \uparrow 1$ yields

$$\sum_{i=1}^{N} \hat{\rho}_i \mathbb{E}[W^*_i] = \frac{1 + \beta\delta\mathbb{E}[S]}{2\beta}, \tag{22}$$

where $\beta = \frac{\mathbb{E}[B]}{\mathbb{E}[B^2]}$ and $\delta = 1 - \sum_{i=1}^{N} \hat{\rho}_i^2$. Combining (21) and (22) completes the proof. □

4.2 Performance measures

Delay: The results of the previous subsection bring us in the position to obtain a closed-form expression for the mean asymptotic scaled delay $\mathbb{E}[W_i^*]$ at each of the queues as exposed in the following corollary.

Corollary 4.3. *For $i = 1, 2, \ldots, N$,*

$$\mathbb{E}[W_i^*] = (1 - \hat{\rho}_i)\frac{1 + \beta\delta\mathbb{E}[S]}{2\beta\delta}.$$

Corollary 4.3 is in agreement with results of [6] and [30] and explicitly reveals the impact of the system parameters on the mean asymptotic scaled delay as stated in the following property (see also [31]).

Property 4.4. *For $i = 1, 2, \ldots, N$,*

1. *$\mathbb{E}[W_i^*]$ is independent of the visit order;*
2. *$\mathbb{E}[W_i^*]$ depends on the service time distributions only through the first two moments of the service time of an arbitrary customer;*
3. *$\mathbb{E}[W_i^*]$ depends on the setup time distributions only through the first moment of the total setup time in a cycle.*

It is important to note that the properties discussed above are in general not valid for stable systems (i.e. with $\rho < 1$).

Correlations: Observe that (16) yields the following expression for the mean asymptotic scaled residual (i, j)-period $\mathbb{E}[R_{\theta_{i,j}^*}]$,

$$\mathbb{E}[R_{\theta_{i,j}^*}] = \frac{1 + \beta\delta\mathbb{E}[S]}{2\beta\delta}\sum_{m=i}^{i+j-1}\hat{\rho}_m, \qquad i = 1, 2, \ldots, N, \quad j = 1, 2, \ldots, N. \tag{23}$$

Combining (4) and (23) together with the following obvious observation for the mean of a scaled asymptotic (i, j)-period $\mathbb{E}[\theta_{i,j}^*]$,

$$\mathbb{E}[\theta_{i,j}^*] = \lim_{\rho\uparrow 1}\ (1 - \rho)\mathbb{E}[\theta_{i,j}] = \mathbb{E}[S]\left(\sum_{m=i}^{i+j-1}\hat{\rho}_m\right), \qquad i = 1, 2, \ldots, N, \quad j = 1, 2, \ldots, N,$$

yields for the corresponding scaled asymptotic variance $\mathbb{V}\mathrm{ar}[\theta_{i,j}^*]$,

$$\mathbb{V}\mathrm{ar}[\theta_{i,j}^*] = \lim_{\rho\uparrow 1}\ (1 - \rho)^2\mathbb{V}\mathrm{ar}[\theta_{i,j}] = \frac{\mathbb{E}[S]}{\beta\delta}\left(\sum_{m=i}^{i+j-1}\hat{\rho}_m\right)^2, \qquad i = 1, 2, \ldots, N, \quad j = 1, 2, \ldots, N.$$

The above expression in conjunction with (5) and (6) gives rise to the following result for the scaled asymptotic covariance $\mathbb{C}\mathrm{ov}[\theta_i^*, \theta_{i+n}^*]$ and the asymptotic correlation $\mathbb{C}\mathrm{or}[\theta_i^*, \theta_{i+n}^*]$ of the visit periods θ_i and θ_{i+n} under heavy traffic defined by, respectively,

$$\begin{aligned}
\mathbb{C}\mathrm{ov}[\theta_i^*, \theta_{i+n}^*] &= \lim_{\rho\uparrow 1}\ (1 - \rho)^2\mathbb{C}\mathrm{ov}[\theta_i, \theta_{i+n}], \qquad i = 1, 2, \ldots, N, \quad n = 1, 2, \ldots, N - 1,\\
\mathbb{C}\mathrm{or}[\theta_i^*, \theta_{i+n}^*] &= \lim_{\rho\uparrow 1}\ \mathbb{C}\mathrm{or}[\theta_i, \theta_{i+n}], \qquad i = 1, 2, \ldots, N, \quad n = 1, 2, \ldots, N - 1,
\end{aligned}$$

with the remark that these limits are again well-defined due to the fact that $\mathbb{E}[W_i]$, $i = 1, 2, \ldots, N$, has a first-order pole at $\rho = 1$.

Corollary 4.5. *For $i = 1, 2, \ldots, N$ and $n = 1, 2, \ldots, N-1$, we have*

$$\begin{aligned} \mathbb{C}ov[\theta_i^*, \theta_{i+n}^*] &= \frac{\hat{\rho}_i \hat{\rho}_{i+n}}{\beta\delta} \mathbb{E}[S], \\ \mathbb{C}or[\theta_i^*, \theta_{i+n}^*] &= 1. \end{aligned}$$

From Corollary 4.5 the following properties about the dependence of the scaled asymptotic covariance and the asymptotic correlation with respect to the system parameters can be perceived.

Property 4.6. *For $i = 1, 2, \ldots, N$ and $n = 1, 2, \ldots, N-1$,*

1. *$\mathbb{C}ov[\theta_i^*, \theta_{i+n}^*]$ and $\mathbb{C}or[\theta_i^*, \theta_{i+n}^*]$ are independent of the visit order;*
2. *the visit time of θ_i of queue i is perfectly correlated with the visit time θ_{i+n} of queue $i+n$;*
3. *$\mathbb{C}ov[\theta_i^*, \theta_{i+n}^*]$ depends on the service time distributions only through the first two moments of the service time of an arbitrary customer;*
4. *$\mathbb{C}ov[\theta_i^*, \theta_{i+n}^*]$ depends on the setup time distributions only through the first moment of the total setup time in a cycle.*

Remark 4.7. Corollary 4.5 shows that the scaled asymptotic covariance of successive visit times equals zero in systems with zero setup times. This observation actually holds for the covariances of visit times in stable systems as well. That is, in systems without setup times the number of visits with zero length tends to infinity and, consequently, the mean and variance of (i, j)-periods both tend to zero implying that the covariance of successive visit periods converges to zero.

Remark 4.8. Corollary 4.5 can be intuitively explained from the results of Coffman *et al.* [5, 6]. They prove a heavy traffic averaging principle (HTAP) for a two-queue polling system with exhaustive service at both queues, from which they conjecture that the same result applies for systems with more than two queues. This HTAP says that, in heavy traffic, the total workload in the system converges to a known process, while on the time scale of this process, the individual workloads change at an infinite rate. This means that the work is shifting between the queues in a rather deterministic way for a period of time, in which the total workload stays relatively constant. This deterministic behavior in the shifting of the workload manifests itself in the perfect correlations between the successive visit times as proved in the present paper. As such, the results rigorously proven in the present paper support the validity of the partially-conjectured results in [5, 6].

Remark 4.9. Corollary 4.5 comprehends the correlation of successive visit times within one single cycle. This result can, however, be extended to correlations of successive times not belonging to the same cycle by modifying (10) in an obvious way. Hence, we obtain that the visit time θ_i of queue i is perfectly correlated with the visit time θ_{i+n} of queue $i+n$, i.e.,

$$\mathbb{C}\text{or}[\theta_i^*, \theta_{i+n}^*] = 1, \qquad i = 1, 2, \ldots, N, \quad n = 1, 2, \ldots.$$

Remark 4.10. Although the main motivation of our interest in correlation of visit times is application oriented (see Section 6 for an extensive discussion), these correlations are of theoretical interest as well. For example, they give an indication of the lengths of simulation runs needed to obtain sufficient narrow confidence intervals of performance measures. The higher the correlations, the longer the simulation should be. As such, the results of the present paper provide a theoretical basis for the inefficiency of simulation techniques for (K-limited) polling systems as observed by, e.g., Blanc [1]. Further, Van Vuuren and Winands [36] recently developed an efficient decomposition method to approximate K-limited polling systems with general arrival processes, in which the covariances between successive visit times play a crucial role. The results in the present paper open possibilities for the derivation of similar algorithms for exhaustive or gated polling systems.

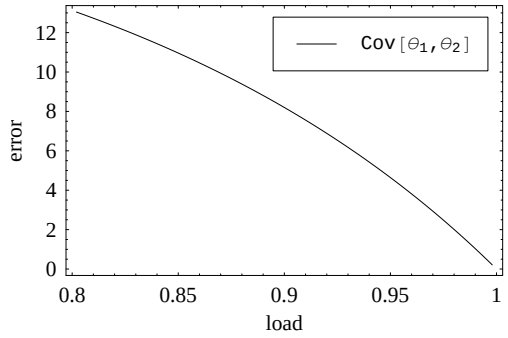

Figure 1: Quality of approximation as function of load (Example 4.11).

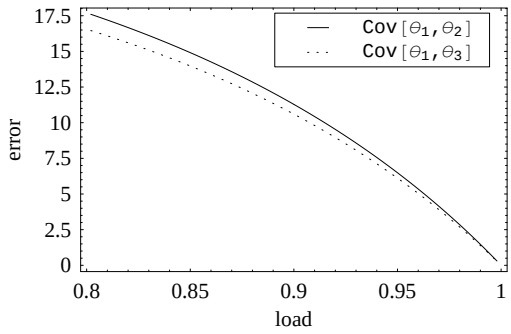

Figure 2: Quality of approximation as function of load (Example 4.12).

4.3 Approximations

Delay: Corollary 4.3 suggests that the mean delays in stable systems can be approximated as follows: For $i = 1, 2, \ldots, N$, $\rho < 1$,

$$\mathbb{E}[W_i] \approx \frac{1-\rho_i}{1-\rho} \frac{1+\beta\delta\mathbb{E}[S]}{2\beta\delta}. \tag{24}$$

We refer to [31] for an extensive discussion of the accuracy of the approximation.

Correlations: Based on Corollary 4.5 we propose the following approximation for the covariance of successive visit periods for stable systems: For $i = 1, 2, \ldots, N$ and $n = 1, 2, \ldots, N-1$, $\rho < 1$,

$$\mathbb{C}\text{ov}[\theta_i, \theta_{i+n}] \approx \frac{1}{(1-\rho)^2} \frac{\rho_i \rho_{i+n}}{\beta\delta} \mathbb{E}[S]. \tag{25}$$

Note that Corollary 4.5 implies that (25) is asymptotically exact in the limiting case $\rho \uparrow 1$. To assess the accuracy of the approximation (25) for stable systems, we have performed numerical experiments, comparing the exact results obtained via the MVA with the approximations according to (25). We define the relative error Δ of the approximation (25) as follows:

$$\Delta = \left| \frac{\mathbb{C}\text{ov}[\theta_i, \theta_{i+n}] - \frac{1}{(1-\rho)^2} \frac{\rho_i \rho_{i+n}}{\beta\delta} \mathbb{E}[S]}{\mathbb{C}\text{ov}[\theta_i, \theta_{i+n}]} \right| \times 100\%,$$

where we use the MVA results of [40] to numerically compute the exact covariances $\mathbb{C}\text{ov}[\theta_i, \theta_{i+n}]$. In this numerical evaluation, (25) is considered to be accurate as Δ is less than 10%.

Example 4.11. Consider a fully symmetric two-queue exhaustive polling system, where the service times follow exponential distributions with means equal to 1 for both customer types. Moreover, the switch-over times have mean and squared coefficient of variation equal to 0.25 and 0.5, respectively. Figure 1 depicts the relative error Δ for $\mathbb{C}\text{ov}[\theta_1, \theta_2]$, which - due to symmetry - equals $\mathbb{C}\text{ov}[\theta_2, \theta_1]$, as function of the total load ρ, in which we see that (25) is accurate when the total load exceeds 87%.

Example 4.12. Consider an asymmetric three-queue exhaustive polling system, where the ratios between the arrival rates are $4 : 16 : 1$ and where the service times follow exponential distributions with mean equal to 1 for all queues. The average switch-over time from queue 1 to queue 2 is equal to 10, while all other switch-over times have means equal to 1. Further, for all switch-over times the squared coefficient of variation equals 0.5. In Figure 2 the relative errors for $\mathbb{C}\text{ov}[\theta_1, \theta_2]$ and $\mathbb{C}\text{ov}[\theta_1, \theta_3]$ are shown, from which we can conclude that (25) is accurate when the total load is 92% or more.

5 Gated service

In the present section we present heavy-traffic results for the gated service discipline. For compactness of the presentation the details of the proofs, which proceed along the lines similar to the case of exhaustive service (see Section 4), are omitted.

5.1 Mean value analysis

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

$$L_{i,i} = \bar{L}_{i,i} + \widetilde{L}_{i,i}, \qquad i = 1, 2, \ldots, N,$$

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

$$L_{i,n} = \widetilde{L}_{i,n}, \qquad i \neq n = 1, 2, \ldots, N.$$

We again start our analysis with the MVA equations derived in [40] for the unknown conditional mean queue lengths $\mathbb{E}[\bar{L}_{i,i}]$ and $\mathbb{E}[\widetilde{L}_{i,n}]$. That is, for $i = 1, 2, \ldots, N$ and $j = 1, 2, \ldots, N$,

$$\sum_{n=i}^{i+j-1} \frac{q_{n,1}}{q_{i,j}} \mathbb{E}[\widetilde{L}_{i,n}] = \lambda_i \mathbb{E}[R_{\theta_{i,j}}], \tag{26}$$

$$(1-\rho_i) \sum_{n=1}^{N} q_{n,1} \mathbb{E}[\widetilde{L}_{i,n}] + q_{i,1} \mathbb{E}[\bar{L}_{i,i}] = \lambda_i \mathbb{E}[R_{\theta_{i,N}}], \tag{27}$$

where the unknown $\mathbb{E}[R_{\theta_{i,j}}]$ are again expressed recursively in terms of $\mathbb{E}[\bar{L}_{i,i}]$ and $\mathbb{E}[\widetilde{L}_{i,n}]$ in the following way, $i = 1, 2, \ldots, N$ and $j = 2, 3, \ldots, N$,

$$\mathbb{E}[R_{\theta_{i,1}}] = \mathbb{E}[\bar{L}_{i,i}]\mathbb{E}[B_i] + \frac{\mathbb{E}[S_{i+1}]}{\mathbb{E}[\theta_{i,1}]} \mathbb{E}[R_{S_{i+1}}] + \frac{\rho_i \mathbb{E}[C]}{\mathbb{E}[\theta_{i,1}]} (\mathbb{E}[R_{B_i}] + \mathbb{E}[S_{i+1}]), \tag{28}$$

$$\begin{aligned} \mathbb{E}[R_{\theta_{i,j}}] = {} & \frac{q_{i,1}}{q_{i,j}} \left(\mathbb{E}[R_{\theta_{i,1}}] \prod_{n=1}^{j-1} (1+\rho_{i+n}) + \sum_{n=1}^{j-1} (\mathbb{E}[S_{i+n+1}] + \mathbb{E}[\widetilde{L}_{i+n,i}]\mathbb{E}[B_{i+n}]) \prod_{m=n+1}^{j-1} (1+\rho_{i+m}) \right) \\ & + (1 - \frac{q_{i,1}}{q_{i,j}}) \mathbb{E}[R_{\theta_{i+1,j-1}}]. \end{aligned} \tag{29}$$

Similar to the exhaustive case, the set (26) - (29) does not allow for a closed-form solution, but we can obtain an explicit solution in the limit of $\rho \uparrow 1$. Multiplying both sides of (26) - (29) by $(1-\rho)$ and letting $\rho \uparrow 1$ yields the corresponding set of equations in heavy traffic for $i = 1, 2, \ldots, N$ and $j = 1, 2, \ldots, N$,

$$\frac{\sum_{n=i}^{i+j-1} \hat{\rho}_n \mathbb{E}[\widetilde{L}^*_{i,n}]}{\sum_{m=i}^{i+j-1} \hat{\rho}_m} = \hat{\lambda}_i \mathbb{E}[R_{\theta^*_{i,j}}], \tag{30}$$

$$(1-\hat{\rho}_i) \sum_{n=1}^{N} \hat{\rho}_n \mathbb{E}[\widetilde{L}^*_{i,n}] + \hat{\rho}_i \mathbb{E}[\bar{L}^*_{i,i}] = \hat{\lambda}_i \mathbb{E}[R_{\theta^*_{i,N}}], \tag{31}$$

with for $i = 1, 2, \ldots, N$ and $j = 2, 3, \ldots, N$,

$$\mathbb{E}[R_{\theta^*_{i,1}}] = \mathbb{E}[\bar{L}^*_{i,i}]\mathbb{E}[B_i], \tag{32}$$

$$\mathbb{E}[R_{\theta_{i,j}}] = \frac{\hat{\rho}_i}{\sum_{n=i}^{i+j-1}\hat{\rho}_n}\left(\mathbb{E}[R_{\theta^*_{i,1}}]\prod_{n=1}^{j-1}(1+\hat{\rho}_{i+n}) + \sum_{n=1}^{j-1}\mathbb{E}[\widetilde{L}^*_{i+n,i}]\mathbb{E}[B_{i+n}]\prod_{m=n+1}^{j-1}(1+\hat{\rho}_{i+m})\right) + \frac{\sum_{n=i+1}^{i+j-1}\hat{\rho}_n}{\sum_{n=i}^{i+j-1}\hat{\rho}_n}\mathbb{E}[R_{\theta^*_{i+1,j-1}}]. \tag{33}$$

Similar to the exhaustive case, the set (30) - (33) can be solved up to some unknown scaling factor $c \in \mathbb{R}$.

Theorem 5.1. *The solution of the set* (30) - (33) *is given by*

$$\mathbb{E}[\bar{L}^*_{i,i}] = c\hat{\lambda}_i, \qquad i = 1, 2, \ldots, N, \tag{34}$$

$$\mathbb{E}[\widetilde{L}^*_{i,i}] = c\hat{\lambda}_i\hat{\rho}_i, \qquad i = 1, 2, \ldots, N, \tag{35}$$

$$\mathbb{E}[\widetilde{L}^*_{i,i+n}] = c\hat{\lambda}_i(2\sum_{m=0}^{n-1}\hat{\rho}_{i+m} + \hat{\rho}_{i+n}), \qquad i = 1, 2, \ldots, N, \quad n = 1, 2, \ldots, N-1, \tag{36}$$

with $c \in \mathbb{R}$.

Once more, via a scaled version of the pseudo-conservation law (3) the unknown scaling factor c can be found.

Theorem 5.2. *The quantity* c *is given by*

$$c = \frac{1 + \beta\delta\mathbb{E}[S]}{2\beta\delta},$$

where $\beta = \frac{\mathbb{E}[B]}{\mathbb{E}[B^2]}$ *and* $\delta = 1 + \sum_{i=1}^{N}\hat{\rho}_i^2$.

5.2 Performance measures

Delay: As in the exhaustive case, the results of the previous subsection yields a closed-form expression for the mean asymptotic scaled delay $\mathbb{E}[W^*_i]$ as shown in the corollary below.

Corollary 5.3. *For* $i = 1, 2, \ldots, N$,

$$\mathbb{E}[W^*_i] = (1 + \hat{\rho}_i)\frac{1 + \beta\delta\mathbb{E}[S]}{2\beta\delta}.$$

Correlations: Along the lines of the exhaustive case, we can also derive results for the scaled asymptotic covariance $\mathbb{C}\mathrm{ov}[\theta^*_i, \theta^*_{i+n}]$ and the asymptotic correlation $\mathbb{C}\mathrm{or}[\theta^*_i, \theta^*_{i+n}]$ of the visit periods θ_i and θ_{i+n} under heavy traffic.

Corollary 5.4. *For* $i = 1, 2, \ldots, N$ *and* $n = 1, 2, \ldots, N-1$, *we have*

$$\mathbb{C}ov[\theta^*_i, \theta^*_{i+n}] = \frac{\hat{\rho}_i\hat{\rho}_{i+n}}{\beta\delta}\mathbb{E}[S],$$

$$\mathbb{C}or[\theta^*_i, \theta^*_{i+n}] = 1.$$

Finally, we want to stress that also in the gated case, as in the exhaustive case, various properties of the performance measures and corresponding approximations can be derived. We omit, however, these results in the interest of space.

Remark 5.5. The assumption in Section 2 that all queues are served according to either the exhaustive or the gated policy is not essential in the analysis, and was only made for compactness of the presentation. In fact, the analysis can be extended without further complication to models with mixtures of gated and exhaustive service, i.e., where some of the queues are served according to the exhaustive policy and some by the gated strategy. Other straightforward extensions are for example models with simultaneous batch arrivals, non-cyclic periodic server routing, and discrete-time models.

6 Specific application area for correlations between visit times

Our motivation to study the correlations between the successive visit times is based on the specific application area of base-stock policies in inventory control. In this section this area of application is discussed in some detail, and can be viewed as an interesting extension to the survey in [16], which is merely focused on applications in computer-communications systems.

Consider a production-inventory system with one single production capacity for multiple products, in which there is an infinite stock space for each product and raw material is always available. Demands for the various products arrive according to stationary and mutually independent stochastic processes. Demand that cannot be satisfied directly from stock is backlogged until the product becomes available after production. The individual products are produced in a make-to-stock fashion with possibly stochastic production times. A possibly stochastic set-up time occurs before the start of the production of a product. Finally, only one product can be produced at a time. This setting is referred to as the stochastic economic lot scheduling problem (SELSP) (see Winands *et al.* [39], for a survey).

In many firms encountering the SELSP, cyclic base-stock policies are used for the control of the inventory of each product, which work as follows (see, e.g., Federgruen and Katalan [7, 8]):

1. the products are produced according to a fixed production sequence;
2. when the machine starts production of a product, it will continue production until a predefined base-stock level has been reached.

Now, the production facility, where the production orders queue up, can be represented as a polling model by identifying each product with a queue and the demand process of a product with the arrival process at the corresponding queue (cf. [7, 8]). Notice that the visit times in the polling system resemble the length of production runs in the production facility.

The SELSP is a common problem in process industries, where the utilization of capacity is typically extremely high. Thus, the heavy-traffic results of the present paper can be used to get fundamental new insights into the behavior and performance of base-stock policies in process industries. In particular, our heavy-traffic results show that the correlations among production runs for base-stock policies are relatively high even in moderate traffic load. This not only implies that the cycle lengths are highly variable, but also that the system may drift from average behavior for a significant period of time. Both effects may lead to higher inventory levels and costs at the production facility itself, are undesirable from an organizational point of view and hamper short-term decision making. Below, we elaborate further on these idiosyncrasies of base-stock policies revealed by our heavy-traffic results.

First, the high correlations between production runs result in very long cycles from time to time, which increases the amount of safety stocks needed at the production facility and rises the concomitant holding costs. Therefore, one may conjecture that exhaustive base-stock policies are not the most effective strategies in production situations where capacity utilizations are high as well

as that it may be desirable that production runs, and thus the cycle lengths, are bounded in these environments. Winands *et al.* [38] show that, in a highly idealized mathematical setting of the SELSP, that bounding production runs of (low-priority) products may indeed lead to considerable cost reductions.
Second, the high variance in cycle lengths caused by the correlations provides breeding ground for the conjecture that base-stock policies do not lead to stability, regularity and discipline on the work floor. These properties are desirable from an organizational point of view, since they facilitate maintenance scheduling, workforce planning, purchasing of raw material, scheduling of subsequent processes and shipment of finished products (see, e.g., Chapter 1 of Van Nyen [35]). Schmidt *et al.* [24] report on a real-life case, where they actually observe the organizational flaws of exhaustive lot sizing policies in a make-to-order production environment. By replacing the exhaustive policy in the plant by a strategy, which stabilizes the cycle lengths, many direct and indirect improvements could be observed. Our heavy-traffic results can be seen as theoretical justification of the lack of stability and discipline of base-stock policies as observed in practice.
Third, the strong correlations among the length of production runs in heavy traffic prove that the performance of base-stock policies in terms of, e.g., delivery times or work-in-progress (WIP) fluctuates strongly over time which may hinder short-term decision making. That is, the actual performance of the system is better than average for some periods of time, but for other periods the performance is below average. As Stoop [26] mentions, in the latter periods managers tend to make nervous myopic decisions in an attempt to reach average performance as quickly as possible, which may result in additional costs and lower long-term performance.

7 Model extensions

Although the results discussed in this paper can be generalized directly to models with mixed gated/exhaustive service policies, simultaneous batch arrivals and periodic server routing without significantly complicating the analysis (see Remark 5.5), application of the MVA-based approach to obtain heavy-traffic results for a variety of interesting models extensions require further analysis. First, it is not yet fully clear if - and how - the MVA-based analysis can be extended to general branching-type service policies (such as binomial-gated, fraction-exhaustive and the like), even though the asymptotics for the mean waiting times are known [30]. Second, a interesting type of extension is to include random or Markovian server routing. Interestingly, for this type of models the buffer-occupancy technique to determine the moments of the joint queue length at polling instants at a queue still works [37], whereas the MTBP-structure of the model is violated, because the immigration of customers becomes dependent on the realization of the visit order (cf. [12, 22] for more details). Although simple expressions for the mean delay seem to be hard to obtain, there is still hope for deriving heavy-traffic asymptotics for the correlations between the successive visit times. This opens up a challenging area for further research. Finally, the results on correlation between the visit times discussed in this paper raise challenging questions about extension of the results to limited-type service policies. Although polling models with K-limited service policies, where the server may only serve up to K_i customers during a visit to queue i, are notoriously difficult and not leave any hope for exact waiting-time asymptotics, it may be possible to derive useful approximations for the mean waiting times and the correlations between the visit times under heavy-traffic assumptions. Intuitively, one would expect that the duration of the visit times becomes degenerate under heavy-traffic assumptions, which gives some hope for asymptotic analysis of the duration and correlations of the visit times, addressing a very challenging area for further research.

Acknowledgements

The authors wish to thank Will Bertrand for fruitful discussions, which led to Section 6 of the present paper.

References

[1] J.P.C. Blanc (1992). An algorithmic solution of polling models with limited service disciplines. IEEE Trans. Commun. 40, 1152-1155.

[2] J.P.C. Blanc (1993). Performance analysis and optimization with the power-series algorithm. In: *Performance Evaluation of Computer and Communication Systems*, eds. L. Donatiello and R. Nelson (Springer-Verlag, Berlin), 53-80.

[3] O.J. Boxma and W.P. Groenendijk (1987). Pseudo conservation laws in cyclic-service systems. J. Appl. Prob. 24, 949-964.

[4] G. Choudhury and W. Whitt (1996). Computing transient and steady state distributions in polling models by numerical transform inversion. Perf. Eval. 25, 267-292.

[5] E.G. Coffman, A.A. Puhalskii and M.I. Reiman (1995). Polling systems with zero switch-over times: a heavy-traffic principle. Ann. Appl. Prob. 5, 681-719.

[6] E.G. Coffman, A.A. Puhalskii and M.I. Reiman (1998). Polling systems in heavy-traffic: a Bessel process limit. Math. Oper. Res. 23, 257-304.

[7] A. Federgruen and Z. Katalan (1996). The stochastic economic lot scheduling problem: cyclical base-stock policies with idle times. Mgmt. Sc. 42, 783-796.

[8] A. Federgruen and Z. Katalan (1998). Determining production schedules under base-stock policies in single facility multi-item production systems. Oper. Res. 46, 883-898.

[9] M.J. Ferguson (1986). Computation of the variance of the waiting times for Token Rings. IEEE J. Sel. Areas Commun. 4, 775-782.

[10] J.C. Fransoo (1992). Demand management and production control in process industries. Inter. J. of Oper. and Prod. Mgmt 12, 187-196.

[11] C. Fricker and M.R. Jaïbi (1994). Monotonicity and stability of periodic polling models. Queueing Systems 15, 211-238.

[12] A.G. Konheim, H. Levy and M.M. Srinivasan (1994). Descendant set: an efficient approach for the analysis of polling systems. IEEE Trans. Commun. 42, 1245-1253.

[13] D.P. Kroese (1997). Heavy traffic analysis for continuous polling models. J. Appl. Prob. 34, 720-732.

[14] S. Kudoh, H. Takagi and O. Hashida (1996). Second moments of the waiting time in symmetric polling systems. J. Oper. Res. Soc. Japan 43, 306-316.

[15] K.K. Leung (1991). Cyclic service systems with probabilistically-limited service. IEEE J. Sel. Areas Commun. 9, 185-193.

[16] H. Levy and M. Sidi (1991). Polling models: applications, modeling and optimization. IEEE Trans. Commun. 38, 1750-1760.

[17] D.M. Markowitz, M.I. Reiman and L.M. Wein (2000). The stochastic economic lot scheduling problem: heavy traffic analysis of dynamic cyclic policies. Oper. Res. 48, 136-154.

[18] D.M. Markowitz and L.M. Wein (2001). Heavy traffic analysis of dynamic cyclic policies: a unified treatment of the single machine scheduling problem. Oper. Res. 49, 246-270.

[19] T.L. Olsen and R.D. van der Mei (2003). Periodic polling systems in heavy-traffic: distribution of the delay. J. Appl. Prob. 40, 305-326.

[20] T.L. Olsen and R.D. van der Mei (2005). Periodic polling systems in heavy-traffic: renewal arrivals. OR Letters 33, 17-25.

[21] M.I. Reiman and L.M. Wein (1998). Dynamic scheduling of a two-class queue with setups. Oper. Res. 46, 532-547.

[22] J.A.C. Resing (1993). Polling systems and multitype branching processes. Queueing Systems 13, 409-426.

[23] D. Sarkar and W.I. Zangwill (1989). Expected waiting time for nonsymetric cyclic queueing systems - Exact results and applications. Mgmt. Sc. 35, 1463-1474.

[24] E. Schmidt, M. Dada, J. Ward and D. Adams (2001). Using cyclic planning to manage capacity at ALCOA. Interfaces 31, 16-27.

[25] M.M. Srinivasan, H. Levy and A.G. Konheim (1993). The individual station technique for the analysis of cyclic polling models. Nav. Res. Logist. 73, 79-101.

[26] P.P.M. Stoop (1996). *Performance Management in Manufacturing, A Method for Short Term Performance Evaluation and Diagnosis*. Ph.D. Thesis, Eindhoven University of Technology.

[27] H. Takagi (1990). Queueing analysis of polling models: an update. In: *Stochastic Analysis of Computer and Communication Systems*, ed. H. Takagi (North-Holland, Amsterdam), 267-318.

[28] H. Takagi (1997). Queueing analysis of polling models: progress in 1990-1994. In: *Frontiers in Queueing: Models, Methods and Problems*, ed. J.H. Dshalalow (CRC Press, Boca Raton, FL), 119-146.

[29] H. Takagi (2000). Analysis and application of polling models. In: *Performance Evaluation: Origins and Directions*, eds. G. Haring, C. Lindemann and M. Reiser, (Lecture Notes in Computer Science 1769, Springer, Berlin), 423-442.

[30] R.D. van der Mei and H. Levy (1997). Polling systems in heavy traffic: exhaustiveness of the service policies. Queueing Systems 27, 227-250.

[31] R.D. van der Mei and H. Levy (1998). Expected delay in polling systems in heavy traffic. Adv. Appl. Prob. 30, 586-602.

[32] R.D. van der Mei (1999). Polling systems in heavy traffic: higher moments of the delay. Queueing Systems 31, 265-294.

[33] R.D. van der Mei (1999). Distributions of the delay in polling systems in heavy traffic. Perf. Eval. 38, 133-148.

[34] R.D. van der Mei (2000). Polling systems with switch-over times under heavy load: moments of the delay. Queueing Systems 36, 381-404.

[35] P.L.M. van Nyen (2005). *The Integrated Control of Production-Inventory Systems*. Ph.D. Thesis, Eindhoven University of Technology.

[36] M. van Vuuren and E.M.M. Winands (2006). Iterative approximation of k-limited polling systems. Report, Eindhoven University of Technology.

[37] J.A. Weststrate (1992). *Analysis and Optimization of Polling Models*. Ph.D. thesis, Tilburg University.

[38] E.M.M. Winands, I.J.B.F. Adan and G.J. van Houtum (2005). A two-queue model with alternating limited service and state-dependent setups. Proceedings of Analysis of Manufacturing Systems - Production Management, Zakynthos, 200-208.

[39] E.M.M. Winands, I.J.B.F. Adan and G.J. van Houtum (2005). The stochastic economic lot scheduling problem: a survey. BETA WP-133, Beta Research School for Operations Management and Logistics, Eindhoven.

[40] E.M.M. Winands, I.J.B.F. Adan and G.J. van Houtum (2006). Mean value analysis for polling systems. To appear in Queueing Systems.