Polling models with renewal arrivals: a new method to derive heavy-traffic asymptotics

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

February 22, 2007

Abstract

We consider asymmetric cyclic polling systems with an arbitrary number of queues, general service-time distributions, zero switch-over times, gated service at each queue, and with general renewal arrival processes at each of the queues. For this classical model, we propose a new method to derive closed-form expressions for the expected delay at each of the queues when the load tends to 1, under proper heavy-traffic (HT) scalings. In the literature on polling models, rigorous proofs of HT limits have only been obtained for polling models with Poisson-type arrival processes, whereas for renewal arrivals HT limits are based on conjectures [6, 7, 15]. Therefore, the main contribution of this paper lies in the fact that we propose a new method to rigorously prove HT limits for a class of non-Poisson-type arrivals. The results are remarkably simple and provide new fundamental insight and reveal explicitly how the expected delay at each of the queues depends on the system parameters, and in particular on the interarrival-time distributions at each of the queues. The results also suggest simple approximations for the expected delay in stable polling systems. Numerical results show that the approximations are highly accurate when the system load is roughly 90% or more.

Keywords: polling systems, renewal arrivals, heavy traffic, delay

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) 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 [13] for an overview. Since the late 1960s polling systems have received much attention in the literature, see [27, 18, 19, 20] for overviews of the available results. The vast majority of the papers assume that the arrival processes at the queues are Poisson. In many applications, however, the interarrival times are not exponentially distributed. Therefore, in this paper we study polling models in which the arrival process at each of the queues is a renewal process, i.e., in which the interarrival times are not necessarily exponentially distributed. Since the existing analysis techniques rely on the assumption of Poisson-type arrivals, hardly any exact results on the waiting-time and queue-length distributions are available for renewal arrivals.

In particular, this paper focuses on the heavy-traffic (HT) behavior of polling models with renewal arrivals, i.e., when the load tends to one. The motivation for studying HT asymptotics is that in many cases they lead to strikingly simple expressions for queue-length and waiting-time distributions, often even in closed form [6, 7, 15], whereas their counterparts for arbitrary values of the load can at best be obtained via numerical techniques that become highly computationally-intensive as the load increases (e.g., [2, 12]). In this way, HT asymptotics not only explicitly show how the waiting-time performance of the system depends on the system parameters leading to significant insights in the behavior of the system, but also form an excellent basis for developing simple approximations for the waiting times (distributions, moment, tail probabilities) for stable systems. In fact, such approximations based on HT limits have been found to be remarkably accurate in many cases, even for moderate load (e.g., [15]).

Motivated by the attractiveness of HT asymptotics, several approaches have been proposed to obtain HT limits for polling systems. For models with Poisson arrivals, rigorous proofs for HT limits can be obtained for models that possess a multi-type branching process (MTBP) structure (cf. [16] for details). By exploring the branching structure of the model, Van der Mei and co-authors [14, 21, 22, 23, 24, 25] explore the recursive relations of the Descendant Set Approach

(DSA) [9] to derive closed-form expressions for the asymptotic delay distribution in HT for polling systems with an MTBP-structure in a general parameter setting, both for cyclic and periodic server routing. Kudoh *et al.* [11] 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 HT limits of the first two moments of the delay for systems with an arbitrary number of queues. Kroese [10] uses the theory of age-dependent branching processes to study the HT behavior of continuous polling systems and shows that the steady-state number of waiting customers has approximately a gamma-distribution. Recently, Van der Mei and Winands [26] have proposed a different technique based on the so-called Mean Value Analysis (MVA) introduced in [29] to derive asymptotic expressions for the expected delay in the classical model with gated and exhaustive service at each of the queues. A fundamentally different approach to obtain HT limits is taken by Coffman *et al.* [6, 7], who use a HT averaging principle to study a two-queue model with exhaustive service at both queues and show that, under HT 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 conjecture that the analysis can be extended to systems with more than two queues. Olsen and Van der Mei [15] adopt the approach in [6, 7] to formulate conjectures about the asymptotic waiting-time distribution in polling models with exhaustive and gated services, and with renewal arrivals; numerical results support the validity of the results.

A remarkable observation is that HT limits for models with more than two queues have only been rigorously proven for models driven by Poisson - or compound-Poisson [25] - processes, models that do satisfy the MTBP-structure (see references above), whereas HT limits for models with renewal arrivals, which generally violate the MTBP structure, have only been obtained on the basis of conjectures [6, 7, 15]. For this reason, in this paper we consider perhaps one of the simplest polling models that does violate the MTBP structure, and propose a new method to derive rigorous proofs for HT asymptotics. More specifically, we consider a gated polling model with a general number of queues, general service-time distributions, zero switch-over times (see also Remark 3.2) under the assumption of general renewal arrivals. For this model, we use a result by Bertsimas and Mourtzinou [1], who derive a set of linear equations for the variance of the cycle times for

gated polling models with renewal arrivals in HT. Taking the proper HT limits of this set, in combination with the conservation law for the GI/G/1 queue in HT, we derive explicit closed-form expressions for the expected delay in HT for the model under consideration. As a by-product, the results reveal several interesting insensitivity properties, and suggest simple approximations for the expected delay in stable polling systems with renewal arrivals. Numerical results with simulations show that the approximations are highly accurate when the load is roughly 90% or more.

The contribution of the present paper is three-fold. First, the main contribution is that we present a new approach to derive rigorous proofs for HT asymptotics for a class of polling models with non-Poisson-type arrivals on hence violate the MTBP-structure, which is an important methodological contribution that opens up a range of challenges for further generalizations. Second, the results provide new insight in the impact of the burstiness of the arrival process on the delay incurred at each of the queues. Third, we use these results to propose simple closed-form approximations for the mean delay in stable systems, and show that these approximations, which allow for back-of-the-envelope calculations, are highly accurate when the load is 90% or more. These observations make the contribution of the present paper evident, both from a methodological and application point-of-view.

The remainder of this paper is organized as follows. In Section 2 the model is described and an expression is given for the scaled expected delay in HT, which is the main result of this paper. In Section 3 a rigorous proof of the result is given, and several asymptotic insensitity properties of the expected delay with respect to the system parameters are formulated. The results also suggest simple and fast approximations for the mean delay for stable systems. In Section 4 the accuracy of the approximations are discussed. Finally, in Section 5 we address a number of challenges for further research.

2 Model description and notation

Consider a system consisting of $N \geq 2$ stations $Q_1, \ldots, Q_N$, each with an infinite-sized buffer. A single server visits the queues in cyclic order, where he applies the *gated* service policy, i.e., when the server polls a queue, he serves all, and only, customers found at the polling instant. Type-i customers arrive at Q_i according to a renewal arrival process, defined by the distribution of the

interarrival times A_i; the arrival rate at Q_i is denote by $\lambda_i := 1/E[A_i]$. The total arrival rate is denoted by $\Lambda = \sum_{i=1}^{N} \lambda_i$. The service time of a type-i customer is a random variable B_i, with finite k-th moment $b_i^{(k)}, k = 1, 2$. The k-th moment of the service time of an arbitrary customer is denoted by $b^{(k)} = \sum_{i=1}^{N} \lambda_i b_i^{(k)}/\Lambda$, $k = 1, 2$. The load offered to Q_i is $\rho_i = \lambda_i b_i^{(1)}$, and the total offered load is equal to $\rho = \sum_{i=1}^{N} \rho_i$. The switch-over times are assumed to be negligible. All interarrival times and service times are assumed to be mutually independent and independent of the state of the system. A necessary and sufficient condition for the stability of the system is $\rho < 1$ (cf. [8]). Throughout, for each variable x that is a function of ρ, we denote its values *evaluated at* $\rho = 1$ by $\hat{x}$. Furthermore, we use the notation that $h(x) \backsim g(x)$ as $x \uparrow a$ means that $\lim_{\rho \uparrow a} h(x)/g(x) = 1$. Finally, for compactness of presentation, 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.

Let W_i be the delay incurred by an arbitrary customer at Q_i, defined as the time between the arrival of a customer at a station and the moment at which he starts to receive service. Our main interest is in the behavior of the mean delay $E[W_i]$ in HT, i.e., as ρ tends to 1. It goes without saying that, in HT, 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 $\rho = 1$, for $i = 1, 2, \ldots, N$,

$$\mathbb{E}[W_i] = \frac{\mathbb{E}[W_i^*]}{1 - \rho} + o((1 - \rho)^{-1}), \qquad \rho \uparrow 1, \tag{1}$$

where $g(x) = o(f(x))$ means that $g(x)/f(x) \to 0$ as $x \uparrow 1$. More colloquially, we can say that $\mathbb{E}[W_i^*]$, which is referred to as the *mean asymptotic scaled delay* at queue i, indicates the rate at which $\mathbb{E}[W_i]$ tends to infinity as $\rho \uparrow 1$. For the validity of the statement that $\mathbb{E}[W_i]$ has a first-order pole at $\rho = 1$, we refer to Remark 3.3.

The main result of the paper is the following.

Theorem 1 (Main result)

For $i = 1, 2, \ldots, N$,

$$E[W_i^*] = \lim_{\rho \uparrow 1} \; (1 - \rho) E\left[W_i\right] = \frac{(1 + \hat{\rho}_i)\sigma^2}{2 \sum_{j=1}^{N} \hat{\rho}_j (1 + \hat{\rho}_j)}, \tag{2}$$

with

$$\sigma^2 := \sum_{i=1}^{N} \hat{\lambda}_i \left(Var[B_i] + \hat{\rho}_i^2 Var[\hat{A}_i] \right). \tag{3}$$

Here, the limit is taken such that the arrival rates are increased, while keeping both the ratios between the arrival rates and the service-time distributions fixed. Notice that in the case of Poisson arrivals we have $\sigma^2 = b^{(2)}/b^{(1)}$.

3 Analysis

This section, which provides a rigorous proof of the main result of the paper (Theorem 1), consists of three parts. In Subsection 3.1, we present a HT analysis for systems with Poisson arrivals. Although the primary goal of this subsection is to lay the foundation of Subsection 3.2 - where we analyse HT asymptotics for systems with renewal arrivals - it provides, as by-product, an alternative derivation for the expected asymptotic delay in the case of Poisson arrivals. The last subsection discusses some properties in terms of the dependence of the mean asymptotic scaled delay with respect to the system parameters.

3.1 Poisson arrivals

Throughout the present subsection, the assumption is made that the arrivals follow Poisson processes. We start our investigation in this case with defining an i-cycle C_i to be the time between two successive polling instants at Q_i. Using simple balance arguments the mean delay at Q_i can be expressed in terms of the first two moments of C_i as follows, for $i = 1, 2, \ldots, N$,

$$E[W_i] = \frac{1+\rho_i}{2} \left(\frac{Var[C_i]}{E[C_i]} + E[C_i] \right). \tag{4}$$

Before we start the analysis, we have to spend some words on the behavior of the cycles in systems without setup times. That is, in such systems each time the system becomes empty, the server will execute, in the limit, an infinite number of cycles and, thus, the number of cycles with zero lengths tends to zero. In order not to be diverted by such effects, we assume, for the time being, that the system possesses at least one strictly positive (deterministic) setup time with mean r. Borst and Boxma [3] proved the continuity of the delay distribution between models with and without setup

times implying that we are allowed, at the end of the analysis, to take the limit $r \downarrow 0$ in order to find the mean delay in the zero setup time model under consideration.

It is well-known that the mean cycle lengths $E[C_i]$ are independent of the queue involved and are given by, for $i = 1, 2, \ldots, N$,

$$E[C_i] = \frac{r}{1-\rho}. \tag{5}$$

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., for $i = 1, 2, \ldots, N$,

$$\rho E[C_i] = E[C_i] - r. \tag{6}$$

Unfortunately, the variance of the cycle lengths $Var[C_i]$ $(i = 1, \ldots, N)$ are, in general, analytically intractable and do depend on the queue involved. Sarkar and Zangwill [17] present the following set of N linear equations, from which the N unknowns $Var[C_i]$ can be computed numerically *under any traffic intensity* $\rho < 1$, for $i = 1, 2, \ldots, N$,

$$\begin{aligned} &\left(\frac{1+2\rho_i-\rho_i^3}{2(1+\rho_i)} - \sum_{l=1}^{i-1} F_{i,l}^{(i)} - \sum_{l=i+1}^{N} E_{l,i}^{(i)} \right) Var[C_i] \\ -&\left(\frac{1}{2(1+\rho_i)} + \sum_{l=1}^{i-1} F_{i,l}^{(i+1)} + \sum_{l=i+1}^{N} E_{l,i}^{(i+1)} \right) Var[C_{i+1}] \\ &- \sum_{k \neq i, i+1} \left(\sum_{l=1}^{i-1} F_{i,l}^{(k)} + \sum_{l=i+1}^{N} E_{l,i}^{(k)} \right) Var[C_k] \quad = \quad \frac{H_i \rho_i}{1+\rho_i} + \sum_{l=1}^{i-1} F_{i,l}^{(0)} + \sum_{l=i+1}^{N} E_{l,i}^{(0)}, \end{aligned} \tag{7}$$

where the coefficients $E_{i,j}^{(k)}$ and $F_{i,j}^{(k)}$ are defined in Appendix A. In the above set, the constant H_i is defined as, for $i = 1, 2, \ldots, N$,

$$H_i = \lambda_i E[C_i] b_i^{(2)}. \tag{8}$$

For further reference, it is convenient to have an explicit expression for the row sums of the corresponding coefficient matrix at our disposal as well, for $i = 1, 2, \ldots, N$,

$$\frac{2\rho_i - \rho_i^3}{2(1+\rho_i)} - \sum_{k=1}^{N} \left(\sum_{l=1}^{i-1} F_{i,l}^{(k)} + \sum_{l=i+1}^{N} E_{l,i}^{(k)} \right) = \rho_i (1-\rho). \tag{9}$$

The complexity of the above set of equations prevents from solving it explicitly for general traffic settings. In HT, however, we can find asymptotically exact closed-form expressions.

To start this HT analysis, we introduce $Var[C_i^*]$, the variance of the asymptotic scaled i-cycle, as follows, for $i = 1, 2, \ldots, N$,

$$Var[C_i^*] = \lim_{\rho \uparrow 1} \ (1-\rho)^2 Var[C_i]. \tag{10}$$

The fact that $E[W_i]$ has a first-order pole at $\rho = 1$ in conjunction with Equations (4) and (5) has the immediate implication that $Var[C_i^*]$ has a second-order pole at $\rho = 1$ and, thus, the above limit is well-defined. In order to find the asymptotic quantities $Var[C_i^*]$, we multiply both sides of (7) by $(1-\rho)$ and let ρ tend to 1. Either by elementary, but tedious, row and column operations or by quoting from Van der Mei and Winands [26] we, subsequently, observe that $Var[C_i^*]$ is *independent* of i (in contrast to $Var[C_i]$ which in general does depend on the queue involved). Finally, it is easily seen that the righthand side of (7) vanishes in the limit.

Thereupon, the following scaled set in HT consisting of one single equation can be obtained, for $i = 1, 2, \ldots, N$,

$$\left(\frac{2\hat{\rho}_1 - \hat{\rho}_1^3}{2(1+\hat{\rho}_1)} - \sum_{k=1}^{N} \sum_{l=2}^{N} E_{l,1}^{(k)} \right) Var[C_i^*] = 0. \tag{11}$$

Calling upon (9) for $i = 1$ shows that (11) forms a homogenous set with an infinite number of non-degenerate solutions, i.e., for $i = 1, 2, \ldots, N$,

$$Var[C_i^*] = c, \tag{12}$$

with $c \in \mathbb{R}$ some unknown scaling factor. Using the fact that both $E[C_i^*]$ and $Var[C_i^*]$ are independent of i yields, with the help of (a scaled version of) Identity (4), for $i, j = 1, 2, \ldots, N$,

$$\frac{E[W_i^*]}{E[W_j^*]} = \frac{1+\hat{\rho}_i}{1+\hat{\rho}_j}. \tag{13}$$

In the limit of $r \downarrow 0$, a unique solution of these mean asymptotic scaled delays can be obtained by exploiting the continuity of the mean delay (see Borst and Boxma [3]) and by adding a single

non-homogenous equation, i.e., a scaled version of the conservation law (see, e.g., [5]), i.e.,

$$\sum_{i=1}^{N} \hat{\rho}_i E[W_i^*] = \frac{\sigma^2}{2}, \tag{14}$$

which yields, for $i = 1, 2, \ldots, N$,

$$E[W_i^*] = \frac{(1+\hat{\rho}_i)\sigma^2}{2\sum_{j=1}^{N} \hat{\rho}_j(1+\hat{\rho}_j)}, \tag{15}$$

which completes our analysis of the Poisson case by recalling that in this case σ^2 is defined as $b^{(2)}/b^{(1)}$. We want to remark that this result is in agreement with results of Coffman *et al.* [6], Van der Mei and Levy [21] and Van der Mei and Winands [26]. In the next subsection, the analysis is extended to the case of renewal arrivals, but, first, we close this subsection with a remark.

Remark 3.1. Although numerics show that the set presented by Sarkar and Zangwill [17] always possesses a unique solution for stable systems, i.e., $\rho < 1$, we did not come across an explicit justification of this observation.

3.2 Renewal arrivals

The present subsection is devoted to the HT analysis of polling systems with general renewal arrivals and, thus, arrivals are no longer assumed to be Poisson but instead follow a renewal process. Analoge to the Poisson case, we again assume, for the time being, that the total (deterministic) setup time r in a cycle is larger than zero.

Our analysis relies on results of Bertsimas and Mourtzinou [1], who state that the equations describing the physics of the system with renewal arrivals in HT are (almost) identical to the ones for the system with Poisson arrivals under any traffic intensity. More specifically, they show that the mean delay $E[W_i]$ in case of renewal arrivals in the limit of ρ tending to 1 is given by, for $i = 1, 2, \ldots, N$,

$$E[W_i] \backsim (1+\hat{\rho}_i)\left(\frac{Var[C_i]}{E[C_i]} + E[C_i]\right) + \frac{(c_{A_i}^2 - 1)b_i^{(1)}}{2}, \tag{16}$$

where an i-cycle C_i is defined identically to the Poisson case and where $c_{A_i}^2$ is the squared coefficient of the interarrival time for queue i. Equation (16) has to be compared with its Poisson counterpart (4), where the latter holds under any traffic intensity.

First of all, one can observe that the mean cycle lengths $E[C_i]$ once more satisfy the remarkably simple form as given by (5). Bertsimas and Mourtzinou [1] prove that, in case of HT, the unknown variables $Var[C_i]$ again satisfy the set of equations formed by (7), i.e., as $\rho \uparrow 1$ for $i = 1, 2, \ldots, N$,

$$\begin{aligned} &\left(\frac{1 + 2\hat{\rho}_i - \hat{\rho}_i^3}{2(1+\hat{\rho}_i)} - \sum_{l=1}^{i-1} F_{i,l}^{(i)} - \sum_{l=i+1}^{N} E_{l,i}^{(i)} \right) Var[C_i] \\ -&\left(\frac{1}{2(1+\hat{\rho}_i)} + \sum_{l=1}^{i-1} F_{i,l}^{(i+1)} + \sum_{l=i+1}^{N} E_{l,i}^{(i+1)} \right) Var[C_{i+1}] \\ &- \sum_{k \neq i, i+1} \left(\sum_{l=1}^{i-1} F_{i,l}^{(k)} + \sum_{l=i+1}^{N} E_{l,i}^{(k)} \right) Var[C_k] \quad \backsim \quad \frac{H_i \hat{\rho}_i}{1+\hat{\rho}_i} + \sum_{l=1}^{i-1} F_{i,l}^{(0)} + \sum_{l=i+1}^{N} E_{l,i}^{(0)}, \end{aligned} \tag{17}$$

where the coefficients $E_{i,j}^{(k)}$ and $F_{i,j}^{(k)}$ are defined in Appendix A. The only minor difference, compared to the Poisson case, is that the constant H_i is now defined as, for $i = 1, 2, \ldots, N$,

$$H_i = \hat{\lambda}_i E[C_i] \left(Var[B_i] + \hat{\rho}_i^2 Var[\hat{A}_i] \right). \tag{18}$$

It is important to stress that the above set is not applicable for stable systems, where, in case of renewal arrivals, the construction of such a set is still an open problem.

As remarked before, the structure of the set (17) does not allow for a closed-form solution and, in order to find the dominating terms of $Var[C_i]$ in HT, we scale this set again by $(1 - \rho)$ and let ρ tend to 1. Due to the fact that the (scaled) set for renewal arrivals is completely identical to the corresponding set for Poisson arrivals (except for the slightly different definition of the constant H_i), exactly the same conclusions as in the preceding subsection can be drawn. That is, in the limit of $\rho \uparrow 1$, the unknowns $Var[C_i^*]$ are independent of the queue involved which renders a homogeneous set consisting of one single equation with an infinite number of non-degenerate solutions, i.e., $0 \times Var[C_i^*] = 0$, which implies that, for $i = 1, 2, \ldots, N$,

$$Var[C_i^*] = c, \tag{19}$$

with $c \in \mathbb{R}$ some unknown scaling factor. Proceeding along the lines of the analysis in the Poisson case, we obtain by using the scaled counterpart of Identity (16), in conjunction with the fact that

both $E[C_i^*]$ and $Var[C_i^*]$ do not depend on i, for $i, j = 1, 2, \ldots, N$,

$$\frac{E[W_i^*]}{E[W_j^*]} = \frac{1+\hat{\rho}_i}{1+\hat{\rho}_j}. \tag{20}$$

These mean asymptotic scaled delays can be scaled properly, in the limit of $r \downarrow 0$, using a scaled version of the conservation law in HT (see, e.g., [1]),

$$\sum_{i=1}^{N} \hat{\rho}_i E[W_i^*] = \frac{\sigma^2}{2}, \tag{21}$$

which yields, for $i = 1, 2, \ldots, N$,

$$E[W_i^*] = \frac{(1+\hat{\rho}_i)\sigma^2}{2\sum_{j=1}^{N} \hat{\rho}_j(1+\hat{\rho}_j)}, \tag{22}$$

and, thus, the proof of the main result of the paper is completed after noting that σ^2 is defined in Equation (3). The approach elucidated in the present section is the first one allowing for a rigorous proof of HT asymptotics in polling systems with a general number of queues. In this context, it is important to remark that our results are in agreement with the conjectures of [6, 7, 15]. We close this subsection with some remarks.

Remark 3.2. In the model defined in Section 2 it is assumed that the setup times are zero. This assumption is realistic for many application areas (e.g., computer-communication systems), but may be unrealistic for other applications (e.g., manufacturing, maintenance). The reason for assuming the switch-over times to be zero is mainly technical: the set of linear equations in [1] for the variance of the scaled cycle times determines the unknowns $Var[C_i^*]$ $(i = 1, 2, \ldots, N)$ up to a scaling constant; for the case of zero switch-over times this scaling constant follows directly from the work-conservation law, while for the case of non-zero switch-over times such a normalizing relation is only available for Poisson arrivals - the so-called pseudo-conservation law which is based on the principle of work decomposition [4] - but not for renewal arrivals.

Remark 3.3. In Section 2, we have assumed that the mean delay incurred at each of the queues, considered as function of ρ, has a first order pole at $\rho = 1$, although (to the best of the authors knowledge) a rigorous proof of this assumption in case of renewal arrivals has not been published in the open literature. However, the results presented here actually prove the validity of this assumption. In fact, the normalizing relation (21) dictates that $\sum_{i=1}^{N} \rho_i E[W_i]$ has a first-order

pole at $\rho = 1$. Moreover, Equations (19) and (5) together imply that, for $i, j = 1, 2, \ldots, N$,

$$\lim_{\rho \uparrow 1} \frac{E[W_i]}{E[W_j]} = \frac{1 + \hat{\rho}_i}{1 + \hat{\rho}_j}, \tag{23}$$

which in turn implies that $E[W_i]$, $i = 1, 2, \ldots, N$, indeed has a first-order pole at $\rho = 1$.

Remark 3.4. The present paper focuses on the mean delay as performance measure of interest both for compactness of presentation as well as for the fact that these delay figures are in many applications the most important performance measures. The scope of applicability of our approach, goes, however, beyond this measure. For example, by exploiting the fact that the scaled variance of the cycle lengths are independent of the queue involved in combination with the (scaled) Equations (55) and (58) in [1] readily leads to the observation that the correlations between successive visit times in gated systems with renewal arrivals converge to one as the load tends to one. It is important to stress that the analysis in this paper proves this observation in systems with and without setup times. For a discussion of the importance of these correlation terms as performance metrics, we refer to [26]. An intuitive explanation for this result can be found in Coffman *et al.* [6, 7], who prove a HT averaging principle 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 averaging principle says that, in HT, 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 such, the results rigorously proven in this paper support the validity of the partially-conjectured results in [6, 7].

3.3 Implications

Theorem 1 reveals the following properties about the dependence of the mean asymptotic scaled delay with respect to the system parameters.

Corollary 1 (Insensitivity)

For $i = 1, 2, \ldots, N$, the mean asymptotic scaled delay $E[W_i^*]$,

(1) *depends on the interarrival-time distributions only through* σ^2*, defined in Equation* (3)*;*

(2) *is independent of the visit order;*

(3) *depends on the second moments of the service-time distributions only through* $b^{(2)}$*, i.e., the second moment of the service time of an arbitrary customer.*

Corollary 1 is known to be not generally valid for stable systems, i.e., for $\rho < 1$, where the individual interarrival-time distributions, the visit order and the individual second moments of the service-time distributions *do* have an impact on the mean waiting times. Hence, Corollary 1 shows that the influence of these parameters on the mean delays vanishes when the load tends to unity, and as such can be viewed as lower-order effects in HT.

4 Approximation

Theorem 1 suggests the following approximation for $E[W_i]$ in stable polling systems, for $i = 1, 2, \ldots, N$, $\rho < 1$,

$$E[W_i^{(app)}] := \frac{1}{1-\rho} \left\{ \frac{1+\hat{\rho}_i}{2\sum_{j=1}^{N} \hat{\rho}_j(1+\hat{\rho}_j)} \left(\sum_{i=1}^{N} \hat{\lambda}_i \left(Var[B_i] + \hat{\rho}_i^2 Var[\hat{A}_i] \right) \right) \right\}. \tag{24}$$

To assess the accuracy of the approximation in (24), in terms of "How high should the load be for the approximation to be accurate?", we have performed numerical experiments to test the accuracy of the approximations for different values of the load of the system. The relative error of the approximation of $E[W_i]$ is defined as follows, for $i = 1, 2, \ldots, N$,

$$\Delta\% := \text{abs}\left(\frac{E[W_i^{(app)}] - E[W_i]}{E[W_i]} \right) \times 100\%. \tag{25}$$

Of course, a wide variety of cases could be examined: different number of queues, choice of interarrival-time distributions and their parameters, choice of service-time distributions and their parameters, etcetera. Since our goal is, however, only to give a flavor of the behavior of the approximation, we confine ourselves to two of the most basic systems. First, we consider a symmetric gated polling model with exponential service times with mean 1. Interarrival times are taken to be deterministic, Erlang_2 or exponential. The latter case, which is included as a benchmark, actually allows for an exact closed-form representation of the mean delay, since - due to the work-conserving nature of the gated service policy - this expected delay figure is identical to the

corresponding quantity in the M/G/1 queue (see, e.g., [5]), for $i = 1, 2, \ldots, N$, $\rho < 1$,

$$E[W_i] = \frac{\rho b^{(2)}/b^{(1)}}{2(1-\rho)}. \tag{26}$$

Subsequently, observing that $E[W_i^{(app)}]$, in this case, reduces as follows, for $i = 1, 2, \ldots, N$, $\rho < 1$,

$$E[W_i^{(app)}] = \frac{b^{(2)}/b^{(1)}}{2(1-\rho)}, \tag{27}$$

yields an explicit expression of the relative error as a function of the total load,

$$\Delta\% := \left(\frac{1}{\rho} - 1\right) \times 100\%. \tag{28}$$

For the other two arrival processes, we obtain the exact values via simulations based on the simulation code described in [28]. Each simulation run is sufficiently long such that the widths of the 95% confidence intervals are smaller than 1% of the predicted value. Table 1 shows the exact and approximated (obtained via (24)) values of $E[W_i]$ for different values of the load (note that these mean delays are independent of i due to symmetry). The results in Table 1 demonstrate that the relative error of the approximations indeed tends to zero as the load tends to 1, as expected on the basis of Theorem 1. Moreover, the results show that the approximation converges to the limit rather quickly when $\rho \uparrow 1$. Roughly, the results are accurate when the load is 90% or more, which demonstrates the applicability of the asymptotic results for practical HT scenarios.

In the second case, we want to study the effect of asymmetry in the system. Therefore, we consider an asymmetric two-queue polling system with Poisson arrivals, where the ratios between the arrival rates are $3:1$ and where the service times follow exponential distributions with mean equal to 1 for both queues. In Table 2 the results of this case are summarized, from which we can again conclude that the approximation is accurate when the total load is 90% or more.

It goes without saying that other (larger) instances can be evaluated just as easily, but we have omitted them for reasons of presentation. Furthermore, it is not inconceivable that the approximations can be refined, but since the primary goal of this paper has been the rigorous proof of HT limits such refinements are beyond the scope of the paper.

	Deterministic			Erlang_2			Exponential		
ρ	$E[W_i^{(app)}]$	$E[W_i]$	$\Delta\%$	$E[W_i^{(app)}]$	$E[W_i]$	$\Delta\%$	$E[W_i^{(app)}]$	$E[W_i]$	$\Delta\%$
0.80	2.50	2.02	23.8	3.75	2.91	28.9	5.00	4.00	25.0
0.85	3.33	2.84	17.3	5.00	4.13	21.1	6.67	5.67	17.6
0.90	5.00	4.49	11.4	7.50	6.65	12.8	10.00	9.00	11.1
0.95	10.00	9.43	6.0	15.00	14.15	6.0	20.00	19.00	5.3
0.98	25.00	24.41	2.4	37.50	36.72	2.1	50.00	49.00	2.0

Table 1. Exact and approximated values for $E[W_i]$ for different values of the load. (Case 1).

	Queue 1			Queue 2		
ρ	$E[W_i^{(app)}]$	$E[W_i]$	$\Delta\%$	$E[W_i^{(app)}]$	$E[W_i]$	$\Delta\%$
0.80	5.38	4.25	26.6	3.85	3.25	18.5
0.85	7.18	6.04	18.9	5.13	4.55	12.7
0.90	10.77	9.63	11.8	7.69	7.13	7.9
0.95	21.54	20.39	5.6	15.38	14.84	3.6
0.98	53.85	52.70	2.2	38.46	37.93	1.4

Table 2. Exact and approximated values for $E[W_i]$ for different values of the load (Case 2).

5 Topics for Further Research

In this paper we have proposed a new method to rigorously prove HT limits for the expected delay in a polling model with gated service at each queue and with general renewal arrivals. The results presented may be generalized in several directions. First, it is interesting to see to what extent the results can be generalized to include non-zero switch-over times. In this context, note that in that case the work-conserving property used is violated, and that the pseudo-conservation law (PCL) that is known to hold for the case of Poisson arrivals [4] is no longer generally valid (see also Remark 3.2). Therefore, extension of the results in [4] to renewal arrivals, under HT assumptions, addresses a challenging area for further research. Second, the approach may be used to handle other types of service policies. We expect that the method can easily be extended to exhaustive service policies at all queues, although the branching property in [16] no longer holds for renewal arrivals. Finally, the proposed method might be extended to derive HT results for the complete waiting-time distributions, rather than for the means only. To this end, decomposition results similar to those on [1] may form an excellent basis to actually prove the conjectures formulated in [6, 7, 15], opening up a very challenging area for further research.

Acknowledgment

The authors wish to thank Marcel van Vuuren for his assistance in using the simulation program used in Section 4.

A Appendix

The coefficients $E_{i,j}^{(k)}$ and $F_{i,j}^{(k)}$ in (7) and (17) are recursively defined as, for $k = 0, 1, \ldots, N$,

$$E_{i,j}^{(0)} = (a_i - \rho_i e_j)E_{i-1,j}^{(0)} - a_i f_j E_{i-1,j+1}^{(0)} + f_j E_{i,j+1}^{(0)} + \frac{H_{i-1}\rho_i}{a_{i-1}\rho_{i-1}}, \quad \text{for} \quad i-j=2, \tag{29}$$

$$E_{i,j}^{(k)} = (a_i - \rho_i e_j)E_{i-1,j}^{(k)} - a_i f_j E_{i-1,j+1}^{(k)} + f_j E_{i,j+1}^{(k)}, \quad \text{for} \quad i-j=2, \tag{30}$$

$$E_{i,j}^{(k)} = (a_i - \rho_i e_j)E_{i-1,j}^{(k)} - a_i f_j E_{i-1,j+1}^{(k)} + f_j E_{i,j+1}^{(k)}, \quad \text{for} \quad i-j\geq 3, \tag{31}$$

$$F_{i,j}^{(k)} = (a_i - \rho_i e_j)F_{i-1,j}^{(k)} - a_i f_j F_{i-1,j+1}^{(k)} + f_j F_{i,j+1}^{(k)}, \quad \text{for} \quad i-j\geq 2, \tag{32}$$

$$\tag{33}$$

with initial conditions, for $j = 1, 2, \ldots, N$,

$$E_{j,j}^{(0)} = H_j, \tag{34}$$

$$E_{j,j}^{(k)} = \begin{cases} \rho_j^2, & k = j, \\ 0, & \text{else}, \end{cases} \tag{35}$$

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

$$E_{j+1,j}^{(0)} = \frac{H_j\rho_{j+1}}{1+\rho_j}, \tag{36}$$

$$E_{j+1,j}^{(k)} = \begin{cases} \frac{\rho_j(1+2\rho_j)\rho_{j+1}}{2(1+\rho_j)}, & k = j+1, \\ +\frac{\rho_j\rho_{j+1}}{2(1+\rho_j)}, & k = j, \\ 0, & \text{else}. \end{cases} \tag{37}$$

For $j = 1, 2, \ldots, N$,

$$F_{j,j}^{(0)} = \frac{H_j\rho_j}{1+\rho_j}, \tag{38}$$

$$F_{j,j}^{(k)} = \begin{cases} \frac{\rho_j(1+2\rho_j+2\rho_j^3)}{2(1+\rho_j)}, & k = j, \\ -\frac{\rho_j}{2(1+\rho_j)}, & k = j+1, \\ 0, & \text{else}. \end{cases} \tag{39}$$

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

$$F^{(0)}_{j+1,j} = \frac{e_j \rho_{j+1}}{1+\rho_j} H_j + \frac{f_j \rho_{j+1}}{1+\rho_{j+1}} H_{j+1}, \tag{40}$$

$$F^{(k)}_{j+1,j} = \begin{cases} \frac{e_j \rho_j (1+2\rho_j)\rho_{j+1}}{2(1+\rho_j)}, & k = j, \\ + \frac{e_j \rho_j \rho_{j+1}}{2(1+\rho_j)} + \frac{f_j \rho_{j+1}(1+2\rho_{j+1}+2\rho^3_{j+1})}{2(1+\rho_{j+1})}, & k = j+1, \\ - \frac{f_j \rho_{j+1}}{2(1+\rho_{j+1})}, & k = j+2, \\ 0, & \text{else.} \end{cases} \tag{41}$$

Finally, the constants a_i, e_i and f_i are defined as, respectively, for $i = 1, 2, \ldots, N$,

$$a_i = \frac{\rho_i(1+\rho_{i-1})}{\rho_{i-1}}, \qquad e_i = \frac{\rho_i}{1+\rho_i} \qquad \text{and} \quad f_i = \frac{1}{a_{i+1}}. \tag{42}$$

References

[1] D. Bertsimas and G. Mourtzinou (1997). Multiclass queueing systems in heavy traffic: an asymptotic approach based on distributional and conservation laws. Oper. Res. 45, 470-487.

[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] S.C. Borst and O.J. Boxma (1997). Polling models with and without switchover times. Oper. Res. 45, 536-543.

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

[5] J.W. Cohen (1969). *The Single Server Queue*. North-Holland Publishing Company, Amsterdam.

[6] 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.

[7] 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.

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

[9] 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.

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

[11] 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.

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

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

[14] 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.

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

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

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

[18] 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.

[19] 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.

[20] 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.

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

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

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

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

[25] R.D. van der Mei (2002). Waiting-time distributions in polling systems with simultaneous batch arrivals. Ann. of Oper. Res. 113, 157-173.

[26] R.D. van der Mei and E.M.M. Winands (2006). Mean value analysis for polling models in heavy traffic. Proc. int. conf. on Performance Evaluation Methodologies and Tools, ValueTools (Pisa).

[27] V.M. Vishnevskii and O.V. Semenova (2006). Mathematical methods to study the polling systems. Automation and Remote Control 67, 173-220.

[28] M. van Vuuren and E.M.M. Winands (2006). Iterative approximation of k-limited polling systems. To appear in Queueing Systems.

[29] E.M.M. Winands, I.J.B.F. Adan and G.J. van Houtum (2006). Mean value analysis for polling systems. Queueing Systems 54, 45-54.