Closed-Form Waiting Time Approximations for Polling Systems*

M.A.A. Boon[†]
marko@win.tue.nl

E.M.M. Winands [‡]
emm.winands@few.vu.nl

I.J.B.F. Adan[†]
iadan@win.tue.nl

A.C.C. van Wijk[§]
a.c.c.v.wijk@tue.nl

October 21, 2009

Abstract

A typical polling system consists of a number of queues, attended by a single server in a fixed order. The present study derives closed-form approximations for the mean waiting times and mean marginal queue lengths of polling systems with renewal arrival processes, which can be computed by simple calculations. The results of the present research may be very suitable for the design and optimisation phase in many application areas, such as telecommunication, maintenance, manufacturing and transportation.

Keywords: Polling, waiting times, queue lengths, approximation

1 Introduction

Polling systems are queueing systems consisting of multiple queues, visited by a single server - typically in a fixed, cyclic order. They find their origin in many real-life applications, e.g. (computer) communication, production and manufacturing environments, traffic and transportation. For a good literature overview of polling systems and their applications, we refer to surveys of, e.g., Takagi [15], Levy and Sidi [9], and Vishnevskii and Semenova [18]. When studying literature on polling systems, it rapidly becomes apparent that the computation of the distributions and moments of the waiting times and marginal queue lengths is very cumbersome. Closed form expressions do not exist, and even when one specifies the number of queues and solves the set of equations that leads to the mean waiting times, the obtained expressions are still too lengthy and complicated to interpret directly. Numerical procedures, both approximate and exact, have been developed in the past to compute these performance

*The research was done in the framework of the BSIK/BRICKS project, and of the European Network of Excellence Euro-NF.

[†]EURANDOM and Department of Mathematics and Computer Science, Eindhoven University of Technology, P.O. Box 513, 5600MB Eindhoven, The Netherlands

[‡]Department of Mathematics, Section Stochastics, VU University, De Boelelaan 1081a, 1081HV Amsterdam, The Netherlands

[§]Department of Industrial Engineering & Innovation Sciences and Department of Mathematics and Computer Science, Eindhoven University of Technology, P.O. Box 513, 5600MB Eindhoven, The Netherlands

measures. However, these methods have several drawbacks. Firstly, they are not transparent and act as a kind of black box. It is, for instance, rather difficult to study the impact of parameters like the occupation rate and the service level. Secondly, these procedures are computationally complex and hard, if not impossible, to implement in a standard spreadsheet program commonly used on the work floor. Finally, the vast majority of standard methods focusses on Poisson arrival processes, which may not be very realistic in many application areas. In the present paper we study polling systems in which the arrival streams are not (necessarily) Poisson, i.e., the interarrival times follow a general distribution. The goal is to derive closed-form approximate solutions for the mean waiting times and mean marginal queue lengths, which can be computed by simple spreadsheet calculations.

Our approach in developing an approximation for the mean waiting times uses novel developments in polling literature. Recently, a heavy traffic (HT) limit has been developed for the mean waiting times as the system becomes saturated [17]. In the present paper we derive an approximation for the light traffic (LT) limit, i.e. as the load goes down to zero, which is exact for Poisson arrivals. The main idea is to create an interpolation between the LT limit and the HT limit. This interpolation yields good results, and has several nice properties, like satisfying the Pseudo Conservation Law (PCL), and being exact for symmetric systems with Poisson arrivals and in many limiting cases. These properties are described in more detail in the present paper. In polling literature, several alternative approximations have been developed before, most of which assume Poisson arrivals. For polling systems with Poisson arrivals and gated or exhaustive service, the best results, by far, are obtained by an approximation based on the PCL (see, e.g., [2, 4, 7]). Fischer et al. [5] study an approximation for the mean waiting times in polling systems, which is also based on an interpolation between (approximate) LT and HT limits. Their approach, however, is applied to a system with Poisson arrivals and time-limited service. Hardly any closed-form approximations exist for non-Poisson arrivals. The few that exist, perform well in specific limiting cases, e.g., under HT conditions [10, 17], or if switch-over times become very large [21, 22], but performance deteriorates rapidly if these limiting conditions are abandoned. We show in an extensive numerical study that the quality of our approximation can be compared to the PCL approximation for systems with Poisson arrivals, but provides good results as well for systems with renewal arrivals.

Because of its simple form, the approximation function is very suitable for optimisation purposes. Although only the mean waiting times of systems with exhaustive or gated service are studied, the results can be extended to higher moments and general branching-type service disciplines. Polling systems with polling tables and/or batch service can also be analysed in a similar manner.

The structure of the present paper is as follows: the next section introduces the model and the required notation, and states the main result. Section 3 illustrates how this main result is obtained, while Section 4 provides results on the accuracy of the approximation for a large set of combinations of input parameter values. The last section discusses further research topics and possible extensions of the model.

2 Model description and main result

The model under consideration is a polling system consisting of N queues, $Q_1, \ldots, Q_N$, with renewal arrival processes. Indices throughout the present paper are understood to be modulo N: Q_{N+1} actually refers to Q_1. Whenever a server switches from Q_i to Q_{i+1}, a random switch-over time S_i is incurred. The generic service requirement of a customer arriving in Q_i, also referred to as a type i customer, is denoted by the random variable B_i. We make the usual independence assumptions for polling systems;

the interarrival times, service times and switch-over times are all independent. The moment at which the server switches from one queue to the next queue, is determined by the *service discipline* of the queue that is being served. In the present paper we focus on polling systems in which each queue is either served according to the *gated* service discipline, which states that during the course of a visit of the server to Q_i, only those type i customers are served that were present at the beginning of that visit, or according to the *exhaustive* service discipline, which means that the server keeps on serving type i customers until Q_i is empty, before switching to Q_{i+1}.

We regard several variables as a function of the load ρ in the system. Scaling is done by keeping the service time distributions fixed, and varying the interarrival times. For each variable x that is a function of the load in the system, ρ, its value evaluated at $\rho = 1$ is denoted by $\hat{x}$. For $\rho = 1$, the generic interarrival time of the stream in Q_i is denoted by $\hat{A}_i$. Reducing the load ρ is done by scaling the interarrival times, i.e., taking the random variable $A_i := \hat{A}_i/\rho$ as generic interarrival time at Q_i. After scaling, the load at Q_i becomes $\rho_i = \rho\frac{\mathbb{E}[B_i]}{\mathbb{E}[\hat{A}_i]}$. The (scaled) rate of the arrival stream at Q_i is defined as $\lambda_i = 1/\mathbb{E}[A_i]$. Similarly, we define arrival rates $\hat{\lambda}_i = 1/\mathbb{E}[\hat{A}_i]$, and proportional load at Q_i, $\hat{\rho}_i = \frac{\rho_i}{\rho}$ ("proportional" because $\sum_{i=1}^N \hat{\rho}_i = 1$). The system is assumed to be stable, so ρ is varied between 0 and 1.

We use B to denote the generic service requirement of an arbitrary customer entering the system, with $\mathbb{E}[B^k] = \frac{\sum_{i=1}^N \hat{\lambda}_i \mathbb{E}[B_i^k]}{\sum_{j=1}^N \hat{\lambda}_j}$ for any integer $k > 0$, and $S = \sum_{i=1}^N S_i$ denotes the total switch-over time in a cycle. Finally, the (equilibrium) residual length of a random variable X is denoted by X^{res}, with $\mathbb{E}[X^{res}] = \frac{1}{2}\mathbb{E}[X^2]/\mathbb{E}[X]$.

We now present the main result of this paper, which is a closed-form approximation formula for the mean waiting time $\mathbb{E}[W_i]$ of a type i customer as a function of ρ:

$$\mathbb{E}[W_{i,app}] = \frac{K_{0,i} + K_{1,i}\rho + K_{2,i}\rho^2}{1-\rho}, \qquad i = 1, \ldots, N. \tag{2.1}$$

The constants $K_{0,i}$, $K_{1,i}$, and $K_{2,i}$ depend on the input parameters and the service discipline. If all queues receive *exhaustive* service, the constants become:

$$K_{0,i} = \mathbb{E}[S^{res}], \tag{2.2}$$

$$K_{1,i} = \hat{\rho}_i\big(\mathbb{E}[\hat{A}_i]\hat{g}_i(0) - 1\big)\mathbb{E}[B_i^{res}] + \mathbb{E}[B^{res}] + \hat{\rho}_i\big(\mathbb{E}[S^{res}] - \mathbb{E}[S]\big) - \frac{1}{\mathbb{E}[S]}\sum_{j=0}^{N-1}\sum_{k=0}^{j}\hat{\rho}_{i+k}\mathbb{V}\mathrm{ar}[S_{i+j}], \tag{2.3}$$

$$K_{2,i} = \frac{1-\hat{\rho}_i}{2}\left(\frac{\sum_{j=1}^N \hat{\lambda}_j\left(\mathbb{V}\mathrm{ar}[B_j] + \hat{\rho}_j^2\mathbb{V}\mathrm{ar}[\hat{A}_j]\right)}{\sum_{j=1}^N \hat{\rho}_j(1-\hat{\rho}_j)} + \mathbb{E}[S]\right) - K_{0,i} - K_{1,i}. \tag{2.4}$$

If all queues receive *gated* service, we get:

$$K_{0,i} = \mathbb{E}[S^{res}], \tag{2.5}$$

$$K_{1,i} = \hat{\rho}_i\big(\mathbb{E}[\hat{A}_i]\hat{g}_i(0) - 1\big)\mathbb{E}[B_i^{res}] + \mathbb{E}[B^{res}] + \hat{\rho}_i\mathbb{E}[S^{res}] - \frac{1}{\mathbb{E}[S]}\sum_{j=0}^{N-1}\sum_{k=0}^{j}\hat{\rho}_{i+k}\mathbb{V}\mathrm{ar}[S_{i+j}], \tag{2.6}$$

$$K_{2,i} = \frac{1+\hat{\rho}_i}{2}\left(\frac{\sum_{j=1}^{N}\hat{\lambda}_j\left(\mathbb{V}\mathrm{ar}[B_j] + \hat{\rho}_j^2\mathbb{V}\mathrm{ar}[\hat{A}_j]\right)}{\sum_{j=1}^{N}\hat{\rho}_j(1+\hat{\rho}_j)} + \mathbb{E}[S]\right) - K_{0,i} - K_{1,i}. \tag{2.7}$$

The term $\hat{g}_i(t)$ is the density of $\hat{A}_i$, the interarrival times at $\rho = 1$. This term is discussed in more detail in the next section, but for practical purposes it is useful to know that $\mathbb{E}[\hat{A}_i]\hat{g}_i(0)$ can be very well approximated by

$$\mathbb{E}[\hat{A}_i]\hat{g}_i(0) \approx \begin{cases} 2\frac{cv_{A_i}^2}{cv_{A_i}^2+1} & \text{if } cv_{A_i}^2 > 1, \\ \left(cv_{A_i}^2\right)^4 & \text{if } cv_{A_i}^2 \leq 1, \end{cases}$$

where $cv_{A_i}^2$ is the squared coefficient of variation (SCV) of A_i (and, hence, also of $\hat{A}_i$). Note that this simplification results in an approximation that requires only the first two moments of each input variable (i.e., service times, switch-over times, and interarrival times).

Remark 2.1 In case of Poisson arrivals, the constants $K_{1,i}$ and $K_{2,i}$ simplify considerably. E.g., for exhaustive service they simplify to:

$$K_{1,i}^{Poisson} = \mathbb{E}[B^{res}] + \hat{\rho}_i\big(\mathbb{E}[S^{res}] - \mathbb{E}[S]\big) - \frac{1}{\mathbb{E}[S]}\sum_{j=0}^{N-1}\sum_{k=0}^{j}\hat{\rho}_{i+k}\mathbb{V}\mathrm{ar}[S_{i+j}],$$

$$K_{2,i}^{Poisson} = (1-\hat{\rho}_i)\left(\frac{\mathbb{E}[B^{res}]}{\sum_{j=1}^{N}\hat{\rho}_j(1-\hat{\rho}_j)} + \frac{\mathbb{E}[S]}{2}\right) - K_{0,i} - K_{1,i}^{Poisson}.$$

The derivation of this approximative formula for the mean waiting time is the topic of the next section. An approximation for the mean *queue length* at Q_i, $\mathbb{E}[L_i]$ is obtained by application of Little's Law to the *sojourn time* of type i customers, i.e. the waiting time plus the service time. As a function of ρ, we have

$$\mathbb{E}[L_{i,app}] = \rho\frac{\mathbb{E}[W_{i,app}] + \mathbb{E}[B_i]}{\mathbb{E}[\hat{A}_i]}.$$

3 Idea behind the approximation

In the present section, we explain the idea behind approximation (2.1) for $\mathbb{E}[W_i]$. The restrictions that we impose on our approximation, are firstly that the formula should be closed-form, and easy to implement, since these are necessities for optimisation purposes and implementation in a spreadsheet. Secondly, we want the approximation to capture the light traffic limit, i.e. $\rho \downarrow 0$, and high traffic limit, i.e. $\rho \uparrow 1$, behaviour in an exact way. Based on these restrictions, we have chosen the form

$$\mathbb{E}[W_{i,app}] = \frac{K_{0,i} + K_{1,i}\rho + K_{2,i}\rho^2}{1-\rho}, \qquad i = 1, \ldots, N.$$

It is proved in [17] that capturing the HT behaviour in an exact way, requires the $(1-\rho)$ term in the denominator. This term is not surprising at all, because the mean waiting times of practically all queueing systems show this behaviour (the best known exception is an $M/G/1$ queue with shortest remaining processing time policy [13]). The motivation for taking a polynomial in the numerator of (2.1) can be found in several other approximations based on interpolation between LT and HT limits. E.g., Reiman and Simon [11] (see also [14]), and Whitt [19] use this approach to develop approximations for the mean waiting time in, respectively, queueing systems with Poisson input and $GI/G/1$ queues. A second-order polynomial fulfills the need for simplicity, *and* is sufficient to obtain an approximation which is exact for the two limiting situations (and, as is shown in Subsection 3.4, in many other limiting cases).

The remainder of this section is devoted to finding the constants $K_{0,i}$, $K_{1,i}$, and $K_{2,i}$. The requirement for the interpolation is an approximation for $\mathbb{E}[W_i]$ in light traffic. No such approximation exists in existing literature, so the next subsection is devoted to finding one. We can use this LT expression to find constants $K_{0,i}$ and $K_{1,i}$ in (2.1). The last unknown in the interpolation, $K_{2,i}$, is obtained using the HT limit of the mean waiting time, which has been found quite recently [17].

3.1 Light traffic

The mean waiting times in the polling model under consideration in light-traffic, have been studied in Blanc and Van der Mei [1], under the assumption of Poisson arrivals. They obtain expressions for the mean waiting times in light traffic that are exact up to (and including) first-order terms in ρ. These expressions have been found by carefully inspecting numerical results obtained with the Power-Series Algorithm, but no proof is provided. In the present section we shall not only prove the correctness of the light-traffic results in a system with Poisson arrivals, but also use them as base for an approximation for the mean waiting times in polling systems with renewal interarrival times. The key ingredient to the LT analysis of a polling system, is the well-known Fuhrmann-Cooper decomposition [6]. It states that in a vacation system with Poisson arrivals the queue length of a customer is the sum of two independent random variables: the number of customers in an isolated $M/G/1$ queue, and the number of customers during an arbitrary moment in the vacation period. The distributional form of Little's Law [8] can be used to translate this result to waiting times. Since no independence is required between the length of a vacation and the length of the preceding visit period, this decomposition also holds for polling systems with Poisson arrivals. We introduce V_i to denote the length of a visit period to Q_i, and I_i to denote the length of the intervisit period, i.e. the time that the server is away between two successive visits to Q_i. Using C_i to denote the cycle time, starting at a visit beginning to Q_i, we have $\mathbb{E}[V_i] = \rho_i \mathbb{E}[C_i]$ and $\mathbb{E}[I_i] = (1-\rho_i)\mathbb{E}[C_i]$. It is well-known that the mean cycle time in polling systems, unlike higher moments, does not depend on the starting point: $\mathbb{E}[C_i] = \mathbb{E}[C] = \frac{\mathbb{E}[S]}{1-\rho}$.

The Fuhrmann-Cooper decomposition, applied to the mean waiting time, results in:

$$\text{exhaustive:} \qquad \mathbb{E}[W_i] = \mathbb{E}[W_{i,M/G/1}] + \mathbb{E}[I_i^{res}], \tag{3.1}$$

$$\text{gated:} \qquad \mathbb{E}[W_i] = \mathbb{E}[W_{i,M/G/1}] + \mathbb{E}[I_i^{res}] + \frac{\mathbb{E}[V_i I_i]}{\mathbb{E}[I_i]}. \tag{3.2}$$

For our approximation, we assume that this decomposition also holds for renewal arrival processes in light traffic. Determining the LT limit of the mean waiting time, $\mathbb{E}[W_i^{LT}]$, in a polling system with exhaustive or gated service is based on the following two-step approach. The first step is to find the LT limit of $\mathbb{E}[W_{i,GI/G/1}]$, the mean waiting time of a $GI/G/1$ queue with only type i customers in

isolation, $i = 1, \ldots, N$. The second step is determining $\mathbb{E}[I_i^{res}]$, the mean residual intervisit time of Q_i, and $\frac{\mathbb{E}[V_i I_i]}{\mathbb{E}[I_i]}$, the mean visit time of Q_i given that it is being observed at a random epoch during the following intervisit time.

For the LT limit of the mean waiting time in a $GI/G/1$ queue, we use Whitt's result (Equation (16) in [19]), which gives:

$$\lim_{\rho_i \downarrow 0} \frac{\mathbb{E}[W_{i,GI/G/1}]}{\rho_i} = \frac{1 + cv_{B_i}^2}{2} \mathbb{E}[\hat{A}_i] \hat{g}_i(0) \mathbb{E}[B_i], \tag{3.3}$$

where $cv_{B_i}^2$ is the SCV of the service times, and $\hat{g}_i(t)$ is the density of the interarrival times $\hat{A}_i$. For practical purposes, it may be more convenient to express $\hat{g}_i(0)$ in terms of the density of A_i, the generic interarrival time of Q_i in the scaled situation.The relation between the density of the scaled interarrival times A_i ($= \hat{A}_i/\rho$), denoted by $g_i(t)$, and the density of $\hat{A}_i$, $\hat{g}_i(t)$, is simply: $g_i(t) = \rho \hat{g}_i(\rho t)$. This means that the term $\mathbb{E}[\hat{A}_i]\hat{g}_i(0)$ can be rewritten as

$$\mathbb{E}[\hat{A}_i]\hat{g}_i(0) = \mathbb{E}[A_i] g_i(0).$$

Because of this equality, in the remainder of the paper we might use either notation. Since determining $\mathbb{E}[\hat{A}_i]\hat{g}_i(0)$ is a required step in the computation of our approximation for $\mathbb{E}[W_i]$, we give some practical examples.

Example 1 If the scaled interarrival times A_i are exponentially distributed with parameter $\lambda_i := 1/\mathbb{E}[A_i]$, we have $g_i(t) = \lambda_i e^{-\lambda_i t}$. This implies that $\mathbb{E}[A_i] g_i(0) = 1$.

Example 2 In this example we assume that A_i follows a H_2 distribution with balanced means. The SCV of A_i is denoted by $cv_{A_i}^2$. The density of this hyper-exponential distribution is (see, e.g., [16])

$$g_i(t) = p\mu_1 e^{-\mu_1 t} + (1-p)\mu_2 e^{-\mu_2 t},$$

with

$$\begin{aligned} p &= \frac{1}{2}\left(1 + \sqrt{\frac{cv_{A_i}^2 - 1}{cv_{A_i}^2 + 1}}\right), \\ \mu_1 &= \frac{1}{\mathbb{E}[A_i]}\left(1 + \sqrt{\frac{cv_{A_i}^2 - 1}{cv_{A_i}^2 + 1}}\right), \\ \mu_2 &= \frac{1}{\mathbb{E}[A_i]}\left(1 - \sqrt{\frac{cv_{A_i}^2 - 1}{cv_{A_i}^2 + 1}}\right). \end{aligned}$$

This leads to $\mathbb{E}[A_i] g_i(0) = 1 + \frac{cv_A^2 - 1}{cv_A^2 + 1} = 2\frac{cv_A^2}{cv_A^2 + 1}$.

Example 3 Now we assume that the interarrival times follow a mixed Erlang distribution. The density of the scaled interarrival times is:

$$g_i(t) = p\frac{\mu^{k-1} t^{k-2}}{(k-2)!} e^{-\mu t} + (1-p)\frac{\mu^k t^{k-1}}{(k-1)!} e^{-\mu t},$$

i.e., a mixture of an Erlang($k-1$) and an Erlang(k) distribution with

$$\begin{aligned} k &= \left\lceil \frac{1}{cv_{A_i}^2} \right\rceil, \\ p &= \frac{k\, cv_{A_i}^2 - \sqrt{k(1+cv_{A_i}^2) - k^2\, cv_{A_i}^2}}{1+cv_{A_i}^2}, \\ \mu &= \frac{k-p}{\mathbb{E}[A_i]}. \end{aligned}$$

If $k > 2$, this leads to $\mathbb{E}[A_i]\, g_i(0) = 0$.

The distributions in Examples $1-3$ are typical distributions to be used in a two-moment fit if the SCV of the interarrival times is respectively 1, greater than 1, and less than 1 (cf. [16]). The examples illustrate how $\mathbb{E}[A_i]g_i(0)$ can be computed if the density of the (scaled) interarrival times is known. If no information is available about the complete density, but the first two moments of A_i are known, Whitt suggests to use the following approximation for $\mathbb{E}[A_i]g_i(0)$:

$$\mathbb{E}[A_i]g_i(0) = \begin{cases} 2\frac{cv_{A_i}^2}{cv_{A_i}^2+1} & \text{if } cv_{A_i}^2 > 1, \\ \left(cv_{A_i}^2\right)^4 & \text{if } cv_{A_i}^2 \leq 1, \end{cases}$$

where $cv_{A_i}^2$ is the squared coefficient of variation of the interarrival times of Q_i. This approximation is exact for $cv_{A_i}^2 > 1$, if the interarrival time distribution is a hyper-exponential distribution as discussed in Example 2. For $cv_{A_i}^2 \leq 1$, the approximation is rather arbitrary, but Example 3 shows that $\mathbb{E}[A_i]g_i(0)$ becomes small (or even zero) very rapidly as $cv_{A_i}^2$ gets smaller.

Summarising, the LT limit of a $GI/G/1$ queue (ignoring $\mathcal{O}(\rho_i^2)$ terms and higher) is:

$$\mathbb{E}[W_{i,GI/G/1}^{LT}] = \rho_i\, \mathbb{E}[A_i]g_i(0)\mathbb{E}[B_i^{res}]. \tag{3.4}$$

For Poisson arrivals ($\mathbb{E}[A_i]g_i(0) = 1$), it is known that $\mathbb{E}[W_{i,M/G/1}] = \frac{\rho_i}{1-\rho_i}\mathbb{E}[B_i^{res}] = \rho_i\mathbb{E}[B_i^{res}] + \mathcal{O}(\rho_i^2)$, which is consistent with our approximation.

The second step in determining the LT limit of the mean waiting time of a type i customer in a polling system, is finding the LT limits of $\mathbb{E}[I_i^{res}]$, the mean residual *intervisit time* of Q_i, and (for gated service only) $\frac{\mathbb{E}[V_i I_i]}{\mathbb{E}[I_i]}$, the mean visit time V_i given that it is observed from the following intervisit time I_i. In this LT analysis we need to focus on first order terms only. Noting the fact that $I_i = S_i + V_{i+1} + S_{i+1} + \cdots + V_{i+N-1} + S_{i+N-1}$, we condition on the moment at which I_i is observed. We distinguish between two cases. The moment of observation either takes place during a visit time, or during a switch-over time:

$$\begin{aligned} \mathbb{E}[I_i^{LT,res}] = &\sum_{j=1}^{N-1} \frac{\mathbb{E}[V_{i+j}]}{\mathbb{E}[I_i]} \mathbb{E}[I_i^{LT,res} | \text{observed during } V_{i+j}] \\ &+ \sum_{j=0}^{N-1} \frac{\mathbb{E}[S_{i+j}]}{\mathbb{E}[I_i]} \mathbb{E}[I_i^{LT,res} | \text{observed during } S_{i+j}]. \end{aligned} \tag{3.5}$$

Observation during visit time. The probability that a random observation epoch takes place during a visit time, say V_j, is $\frac{\mathbb{E}[V_j]}{\mathbb{E}[I_i]}$, for any $j \neq i$. However, we are only interested in order ρ terms, so this probability simplifies to

$$\frac{\mathbb{E}[V_j]}{\mathbb{E}[I_i]} = \frac{\rho_j \mathbb{E}[C]}{(1-\rho_i)\mathbb{E}[C]} = \rho_j + \mathcal{O}(\rho^2).$$

The fact that this probability is $\mathcal{O}(\rho)$, implies that all further $\mathcal{O}(\rho)$ terms in $\mathbb{E}[I_i^{LT,res}|\text{observed during } V_j]$ can be ignored, because in LT we focus on first order terms only.

The length of the residual intervisit time is the length of the residual visit period of type j customers, V_j^{res}, plus all switch-over times $S_j + \cdots + S_{i-1}$, plus all visit times $V_{j+1} + \cdots + V_{i-1}$. The first term simplifies to $\mathbb{E}[V_j^{res}] = \mathbb{E}[B_j^{res}] + \mathcal{O}(\rho)$. The terms $\mathbb{E}[V_k|\text{observed from } V_j]$, $k = j+1, \ldots, i-1$, in light traffic, are all $\mathcal{O}(\rho)$. Summarising, the mean residual intervisit period when observed during V_j is simply a mean residual service time $\mathbb{E}[B_j^{res}]$, plus all mean switch-over times $\mathbb{E}[S_j + \cdots + S_{i-1}]$, plus $\mathcal{O}(\rho)$ terms:

$$\mathbb{E}[I_i^{LT,res}|\text{observed during } V_j] = \mathbb{E}[B_j^{res}] + \sum_{k=j}^{i-1} \mathbb{E}[S_k] + \mathcal{O}(\rho). \tag{3.6}$$

Observation during switch-over time. We continue by determining the mean residual intervisit period, conditioned on a random observation epoch during a switch-over time, say S_j, $j = 1, \ldots, N$. The probability that such an epoch takes place during S_j, is

$$\frac{\mathbb{E}[S_j]}{\mathbb{E}[I_i]} = \frac{\mathbb{E}[S_j]}{(1-\rho_i)\mathbb{E}[C]} = \frac{\mathbb{E}[S_j]}{\mathbb{E}[S]} \frac{1-\rho}{1-\rho_i} = \frac{\mathbb{E}[S_j]}{\mathbb{E}[S]}(1-\rho+\rho_i) + \mathcal{O}(\rho^2).$$

It becomes apparent from this expression that things get slightly more complicated now, because order ρ terms in the conditional residual intervisit time may no longer be neglected. The residual intervisit time now consists of the residual switch-over time S_j^{res}, plus the switch-over times $S_j + \cdots + S_{i-1}$, plus all visit periods $V_{j+1} + \cdots + V_{i-1}$. The length of a visit period V_k, for $k > j$, is the sum of the busy periods of all type k customers that have arrived during $S_i, \ldots, S_{j-1}$, S_j^{past}, S_j^{res}, and $S_{j+1}, \ldots, S_{k-1}$. By S_j^{past} we denote the elapsed switch-over time during which the intervisit period is observed, which has the same distribution as the residual switch-over time S_j^{res}. Compared to an observation during a visit time, it is more difficult to determine the conditional mean length of a busy period $\mathbb{E}[V_k|\text{observed during } S_j]$ under LT. We use a heuristic approach, which is exact if the arrival process of type k customers is Poisson, and approximate it by:

$$\mathbb{E}[V_k|\text{observed during } S_j] \approx \rho_k \left(\sum_{l \neq j} \mathbb{E}[S_l] + \mathbb{E}[S_j^{past}] + \mathbb{E}[S_j^{res}] \right) + \mathcal{O}(\rho^2), \qquad k = j+1, \ldots, i-1.$$

If A_k is exponentially distributed, the above expression is exact. Nevertheless, numerical experiments have shown that this approximative assumption has no or at least negligible impact on the accuracy of the approximated mean waiting times. Summarising:

$$\begin{aligned}\mathbb{E}[I_i^{LT,res}|\text{observed during } S_j] \approx{}& \sum_{k=i}^{j-1} \mathbb{E}[S_k]\Big(\sum_{l=j+1}^{i+N-1} \rho_l\Big) + \mathbb{E}\left(S_j^{past}\right)\Big(\sum_{k=j+1}^{i+N-1} \rho_k\Big) + \mathbb{E}\left(S_j^{res}\right)\Big(1 + \sum_{k=j+1}^{i+N-1} \rho_k\Big) \\ &+ \sum_{k=j+1}^{i+N-1} \mathbb{E}[S_k]\Big(1 + \sum_{l=j+1}^{i+N-1} \rho_l\Big) + \mathcal{O}(\rho^2).\end{aligned} \tag{3.7}$$

The expression for I_i^{res} under light traffic conditions now follows from substituting (3.6) and (3.7) in (3.5). The result can be rewritten to:

$$\begin{aligned}
\mathbb{E}[I_i^{LT,res}] \approx & \sum_{j=i+1}^{i+N-1} \rho_j \mathbb{E}[B_j^{res}] + \sum_{j=i+1}^{i+N-1} \rho_j \sum_{k=j}^{i+N-1} \mathbb{E}[S_k] \\
& + \sum_{j=i}^{i+N-1} \frac{1}{2\mathbb{E}[S]} \left[\mathbb{E}\left(S_j^2\right) \left(1 - \rho + \rho_i + 2 \sum_{k=j+1}^{i+N-1} \rho_k\right) \right] \\
& + \frac{1}{\mathbb{E}[S]} \left[\sum_{k=i}^{j-1} \mathbb{E}[S_j]\mathbb{E}[S_k]\Big(\sum_{l=j+1}^{i+N-1} \rho_l\Big) + \sum_{k=j+1}^{i+N-1} \mathbb{E}[S_j]\mathbb{E}[S_k]\Big(1 - \rho + \rho_i + \sum_{l=j+1}^{i+N-1} \rho_l\Big) \right] \\
& + \mathcal{O}(\rho^2) \\
= & \sum_{j=i+1}^{i+N-1} \rho_j \mathbb{E}[B_j^{res}] + \sum_{j=i+1}^{i+N-1} \rho_j \sum_{k=j}^{i+N-1} \mathbb{E}[S_k] \\
& + (1 - \rho + \rho_i)\mathbb{E}[S^{res}] + \frac{1}{\mathbb{E}[S]} \sum_{j=i}^{i+N-1} \sum_{k=i}^{i+N-1} \mathbb{E}[S_j S_k]\Big(\sum_{l=j+1}^{i+N-1} \rho_l\Big) + \mathcal{O}(\rho^2) \\
= & \, \mathbb{E}[S^{res}] + \rho \mathbb{E}[B^{res}] - \rho_i \mathbb{E}[B_i^{res}] + \rho_i \big(\mathbb{E}[S^{res}] - \mathbb{E}[S]\big) - \frac{1}{\mathbb{E}[S]} \sum_{j=0}^{N-1} \sum_{k=0}^{j} \rho_{i+k} \mathbb{V}\mathrm{ar}[S_{i+j}] \\
& + \mathcal{O}(\rho^2), \qquad (3.8)
\end{aligned}$$

for $i = 1, \ldots, N$. The last step in (3.8) follows after some straightforward (but tedious) rewriting.

The Fuhrmann-Cooper decomposition of the mean waiting time for customers in a polling system with *gated* service (3.2), also requires computing $\frac{\mathbb{E}[V_i I_i]}{\mathbb{E}[I_i]}$ under LT conditions. Here, again, we have to resort to using a heuristic and use $\frac{\mathbb{E}[V_i I_i]}{\mathbb{E}[I_i]} = \rho_i \mathbb{E}[S] + \mathcal{O}(\rho^2)$, because this value is *exact* in the case of Poisson arrivals. Intuitively this term can be explained by observing that the only thing that changes for gated service, compared to exhaustive service, is that type i customers arriving during V_i are not served until the next cycle. As we have seen before, the probability of a type i arrival taking place during V_i is $\rho_i + \mathcal{O}(\rho^2)$. The mean residual cycle, observed from a random epoch in V_i, is $\mathbb{E}[C_i^{res}|\text{observed during } V_i] = \mathbb{E}[S] + \mathcal{O}(\rho)$. Combined, this gives $\frac{\mathbb{E}[V_i I_i]}{\mathbb{E}[I_i]} = \rho_i \mathbb{E}[S] + \mathcal{O}(\rho^2)$, in the case of Poisson arrivals. If the arrival process is not Poisson, this is not exact, but we use it as an approximation.

Having made all required preparations, we are ready to formulate the main result of the present subsection. Under light traffic, an approximation for the mean waiting time of a type i customer in a polling model with general arrivals and respectively exhaustive and gated service in Q_i, is:

$$\begin{aligned}
\mathbb{E}[W_i^{LT,exh}] \approx & \, \mathbb{E}[S^{res}] + \rho_i (\mathbb{E}[\hat{A}_i]\hat{g}_i(0) - 1)\mathbb{E}[B_i^{res}] + \rho \mathbb{E}[B^{res}] + (\rho - \rho_i)\left(\mathbb{E}[S] - \mathbb{E}[S^{res}]\right) \\
& + \frac{1}{\mathbb{E}[S]} \sum_{k=i+1}^{i+N-1} \rho_k \sum_{j=i}^{k-1} \mathbb{V}\mathrm{ar}[S_j] + \mathcal{O}(\rho^2), \qquad i = 1, \ldots, N, \qquad (3.9) \\
\mathbb{E}[W_i^{LT,gated}] \approx & \, \mathbb{E}[W_i^{LT,exh}] + \rho_i \mathbb{E}[S], \qquad (3.10)
\end{aligned}$$

where $\hat{g}_i(t)$ is the density of the interarrival times of type i customers at $\rho = 1$. Equation (3.9) follows

from substitution of (3.4) and (3.8) in

$$\mathbb{E}[W_i] \approx \mathbb{E}[W_{i,GI/G/1}] + \mathbb{E}[I_i^{res}], \qquad i = 1, \ldots, N. \tag{3.11}$$

For Poisson arrivals, (3.9) en (3.10) are exact. The LT limit for polling systems with Bernoulli service (and Poisson arrivals) has been experimentally found in [1] and, indeed, it can be shown that their result for exhaustive service, which is a special case of Bernoulli service, agrees with our result after substituting $\mathbb{E}[\hat{A}_i]\hat{g}_i(0) = 1$ in (3.9).

3.2 Heavy traffic

The mean delay in a polling system with renewal arrivals in HT, i.e. as ρ tends to 1, has been analysed in [17], where the following result has been obtained:

$$\mathbb{E}[W_i^{HT}] = \frac{\omega_i}{1-\rho} + o((1-\rho)^{-1}), \qquad \rho \uparrow 1. \tag{3.12}$$

Obviously, in HT, all queues become unstable and, thus, $\mathbb{E}[W_i]$ tends to infinity for all i. The rate at which $\mathbb{E}[W_i]$ tends to infinity as $\rho \uparrow 1$ is indicated by ω_i, which is referred to as the *mean asymptotic scaled delay* at queue i, and depends on the service discipline. For exhaustive service,

$$\omega_i = \frac{1-\hat{\rho}_i}{2}\left(\frac{\sigma^2}{\sum_{j=1}^N \hat{\rho}_j(1-\hat{\rho}_j)} + \mathbb{E}[S]\right), \qquad i = 1, \ldots, N,$$

with

$$\sigma^2 := \sum_{i=1}^N \hat{\lambda}_i \left(\mathbb{V}\mathrm{ar}[B_i] + \hat{\rho}_i^2 \mathbb{V}\mathrm{ar}[\hat{A}_i]\right).$$

Here, the limits are taken such that the arrival rates are increased, while keeping the service-time distributions fixed, and keeping the distributions of the interarrival times A_i $(i = 1, \ldots, N)$ fixed up to a common scaling constant ρ. Notice that in the case of Poisson arrivals we have $\sigma^2 = \mathbb{E}[B^2]/\mathbb{E}[B]$.

For gated service, we have

$$\omega_i = \frac{1+\hat{\rho}_i}{2}\left(\frac{\sigma^2}{\sum_{j=1}^N \hat{\rho}_j(1+\hat{\rho}_j)} + \mathbb{E}[S]\right).$$

3.3 Interpolation

Now that we have the expressions for the mean delay in both LT and HT, we can determine the constants $K_{0,i}$, $K_{1,i}$, and $K_{2,i}$ in approximation formula (2.1). We simply impose the requirements that approximation (2.1) results in the same mean waiting time for $\rho = 0$ as the LT limit, and for $\rho \uparrow 1$ as the HT limit. Since (3.9) (and (3.10) for gated service) has been determined up to the first order of ρ terms, we also add the requirement that the derivative with respect to ρ, taken at $\rho = 0$, of our approximation is equal to the derivative of the LT limit. A more formal definition of these requirements is presented below:

$$\begin{aligned}
\mathbb{E}[W_{i,app}]\big|_{\rho=0} &= \mathbb{E}[W_i]\big|_{\rho=0}, \\
\frac{\mathrm{d}}{\mathrm{d}\rho}\mathbb{E}[W_{i,app}]\big|_{\rho=0} &= \frac{\mathrm{d}}{\mathrm{d}\rho}\mathbb{E}[W_i]\big|_{\rho=0}, \\
(1-\rho)\mathbb{E}[W_{i,app}]\big|_{\rho=1} &= (1-\rho)\mathbb{E}[W_i]\big|_{\rho=1}.
\end{aligned}$$

This leads to (2.1) as approximation for $\mathbb{E}[W_i]$ in a polling system with general arrivals. Constants $K_{0,i}$, $K_{1,i}$, and $K_{2,i}$ are defined in (2.2)–(2.4) for systems with exhaustive service, or (2.5)–(2.7) for gated service.

3.4 Special cases

The approximation for the mean waiting time of a type i customer, $\mathbb{E}[W_{i,app}]$, has several nice properties discussed in the remainder of this subsection.

Pseudo-conservation law. A well-known result in polling literature, is the *pseudo-conservation law*, derived by Boxma and Groenendijk [3] using the concept of work decomposition. This law gives the following exact expression for the weighted sum of the mean waiting times in a polling system with Poisson arrivals:

$$\sum_{i=1}^{N} \rho_i \mathbb{E}[W_i] = \frac{\rho^2}{1-\rho}\mathbb{E}[B^{res}] + \rho\mathbb{E}[S^{res}] + \frac{\mathbb{E}[S]}{2}\frac{\rho^2 - \sum_{j=1}^{N}\rho_j^2}{1-\rho} + \sum_{j=1}^{N}\mathbb{E}[Z_{jj}], \tag{3.13}$$

where $\mathbb{E}[Z_{jj}]$ denotes the mean amount of work left behind in Q_j at the completion of a visit of the server to Q_j. It is shown in [3] that $\mathbb{E}[Z_{jj}]$ is the only term that depends on the service discipline. For exhaustive service $\mathbb{E}[Z_{jj}] = 0$, for gated service $\mathbb{E}[Z_{jj}] = \rho_j^2 \frac{\mathbb{E}[S]}{1-\rho}$. It can be shown that our approximation satisfies the pseudo-conservation law in the case of Poisson arrivals: if $\mathbb{E}[\hat{A}_i]\hat{g}_i(0) = 1$ for $i = 1, \ldots, N$, then $\sum_{i=1}^{N}\rho_i\mathbb{E}[W_{i,app}]$ also equals the right-hand side of (3.13). The derivation consists of basic, but cumbersome, algebraic manipulations only, and is therefore omitted. We only mention a helpful intermediate result: $\sum_{i=1}^{N}\sum_{k=i+1}^{i+N}\sum_{j=i}^{k-1}\rho_i\rho_k = N\sum_{i=1}^{N}\sum_{k=i}^{N}\rho_i\rho_k$, so $\sum_{i=1}^{N}\sum_{k=i+1}^{i+N}\sum_{j=i}^{k-1}\rho_i\rho_k\mathbb{V}\text{ar}[S_j] = \frac{1}{2}\left(\rho^2 + \sum_{i=1}^{N}\rho_i^2\right)\mathbb{V}\text{ar}[S]$. Using this result, it follows that $\sum_{i=1}^{N}\rho_i K_{2,i} = 0$.

Light and heavy traffic. The light traffic limit of $\mathbb{E}[W_i]$, given by (3.9) for exhaustive service and by (3.10) for gated service, is exact for Poisson arrivals. The heavy traffic limit (3.12) of $\mathbb{E}[W_i]$ is even exact for renewal arrivals. An appropriate choice of constants $K_{0,i}$, $K_{1,i}$ and $K_{2,i}$ can reduce (2.1) to either (3.9), (3.10), or (3.12). Since the LT and HT limits have been used in the set of equations that determine the coefficients of the approximation, it goes without saying that $\mathbb{E}[W_{i,app}]$ is equal to (3.9) (or (3.10) for gated service) and (3.12), for $\rho \downarrow 0$ and $\rho \uparrow 1$ respectively. This implies that the LT limit of our approximation is exact for Poisson arrivals, and the HT limit is exact for general arrivals.

Symmetric system. If $\hat{\rho}_i = \frac{1}{N}$ for all $i = 1, \ldots, N$, all B_i have the same distribution, and the variances $\mathbb{V}\text{ar}[S_i]$ of all switch-over times are equal, then our approximation is exact if all interarrival

distributions are exponential. For exhaustive service, we obtain

$$\begin{aligned}
K_{1,i} &= \mathbb{E}[B^{res}] + \frac{N-1}{N}\mathbb{E}[S] - \left(2 - \frac{1}{N}\right)\mathbb{E}[S^{res}] + \frac{1}{\mathbb{E}[S]}\sum_{k=i+1}^{i+N-1}\hat{\rho}_k \sum_{j=i}^{k-1}\mathbb{V}\text{ar}[S_j] \\
&= \mathbb{E}[B^{res}] + \frac{N-1}{N}\mathbb{E}[S] - \left(2 - \frac{1}{N}\right)\mathbb{E}[S^{res}] + \frac{N-1}{N}\frac{\mathbb{V}\text{ar}[S]}{2\mathbb{E}[S]} \\
&= \mathbb{E}[B^{res}] + \left(1 - \frac{1}{N}\right)\frac{\mathbb{E}[S]}{2} - \mathbb{E}[S^{res}],
\end{aligned}$$

which means that $\mathbb{E}[W_{i,app}] = \mathbb{E}[W_{i,symm}]$ (because $K_{2,i} = 0$ in a symmetric system), with

$$\mathbb{E}[W_{i,symm}] = \frac{\rho}{1-\rho}\mathbb{E}[B^{res}] + \mathbb{E}[S^{res}] + \frac{\rho(1-\frac{1}{N})}{1-\rho}\frac{\mathbb{E}[S]}{2}.$$

Note that we do *not* require that the mean switch-over times $\mathbb{E}[S_i]$ are equal. One can verify that the same holds for gated service.

Single queue (vacation model). An immediate consequence of the fact that our approximation is exact in symmetric polling systems with Poisson arrivals, is that it also gives exact results for the mean waiting time of customers in a single-queue polling system with Poisson arrivals. A polling system consisting of only one queue, but with a switch-over time between successive visits to this queue, is generally referred to as a queueing system with multiple server vacations.

Large switch-over times. For S deterministic, $S \to \infty$, and, again, under the assumption of Poisson arrivals, it is proven in [20, 21] that $\frac{\mathbb{E}[W_i]}{S} \to \frac{1-\rho_i}{2(1-\rho)}$ for exhaustive service. It can easily be verified that our approximation has the same limiting behaviour:

$$\lim_{S\to\infty}\frac{\mathbb{E}[W_{i,app}]}{S} = \frac{1-\rho_i}{2(1-\rho)}.$$

For gated service, $\frac{\mathbb{E}[W_{i,app}]}{S} \to \frac{1+\rho_i}{2(1-\rho)}$, which is also the exact limit (see, e.g., [20]).

Miscellaneous other exact results. The approximation is also exact in several other cases, all with Poisson arrivals, when the parameter values are carefully chosen. The relations between the input parameters that yield exact approximation results become very complicated, especially in polling systems with more than two queues. We only mention one interesting example here: our approximation gives exact results for a two-queue polling system with exhaustive service and

$$\mathbb{E}[B_1] = \mathbb{E}[B_2], \mathbb{E}[S_1] = \mathbb{E}[S_2], cv^2_{A_1} = cv^2_{A_2}, cv^2_{B_1} = cv^2_{B_2}, cv^2_{S_1} = cv^2_{S_2}, \tag{3.14}$$

if the following constraint is satisfied:

$$\rho = \frac{1 + I^2_{A_i}}{2I_{A_i}} - \frac{cv^2_{S_i}}{1 + cv^2_{B_i}} \cdot \frac{\mathbb{E}[S_i]}{\mathbb{E}[B_i]}, \tag{3.15}$$

where $I_{A_i} = \frac{\hat{\rho}_1}{\hat{\rho}_2}$ is the ratio of the loads of the two queues. Obviously, if $I_{A_i} = 1$, the system is symmetric and our approximation gives exact results regardless of the other parameter settings.

4 Numerical study

4.1 Initial glance at the approximation

Before we study the accuracy of the approximation to a huge test bed of polling systems, we just pick a rather arbitrary, simple system to compare the approximation with exact results in order to get some initial insights. Consider a three-queue polling system with loads of Q_1, Q_2, and Q_3 divided as follows: $\hat{\rho}_1 = 0.1$, $\hat{\rho}_2 = 0.3$, and $\hat{\rho}_3 = 0.6$. All service times and switch-over times are exponentially distributed, with mean 1. The interarrival times have SCV $cv^2_{A_i} = 3$ for $i = 1, 2, 3$. In Figure 1 we plot the approximated mean waiting time of Q_2, $\mathbb{E}[W_{2,app}]$, versus the load of the system ρ. Since this system cannot be analysed analytically, we compare the approximated values with simulated values. Both in the approximation and in the simulation we fit a H_2 distribution as described in Example 2.

The errors are largest for Q_2, which is the reason why we chose this queue in particular in Figure 1. The most important information that this figure reveals, is that even though the accuracy of the approximation is worst for this queue (a relative error of -4.47% for $\rho = 0.7$), the shape of the approximation function is very close to the shape of the exact function, which makes it very suitable for optimisation purposes. The maximum relative errors of Q_1 and Q_3 are 3.10% and 2.90% respectively.

In order to get more insight in the numerical accuracy of the approximation for a huge variety of different parameter settings, we create a large test bed in the next subsection and compare the approximation with exact or simulated results. It turns out that the maximum relative errors for most of the polling systems are smaller than the one selected in the above example.

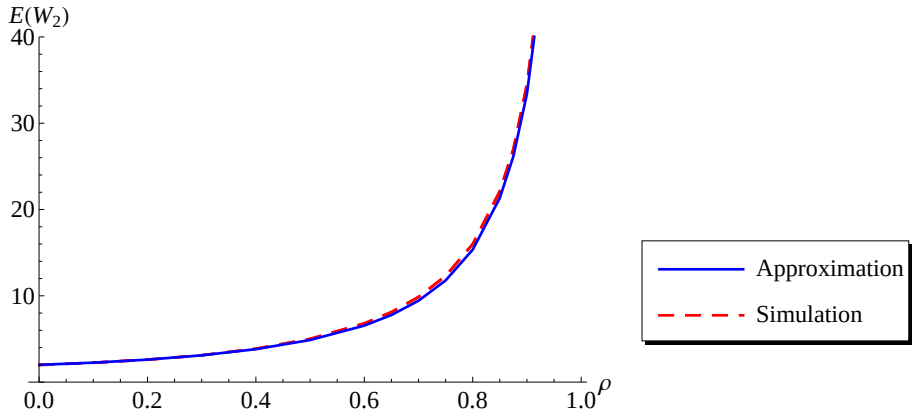

Figure 1: Approximated and simulated mean waiting time $\mathbb{E}[W_2]$ of Q_2 of the example in subsection 4.1.

4.2 Accuracy of the approximation

In the present section we study the accuracy of our approximation. We compare the approximated mean waiting times of customers in various polling systems to the exact values. The complete test bed of polling systems that are analysed, contains 2304 different combinations of parameter values, all listed in Table 1. We show detailed results for exhaustive service first, and discuss polling systems

with gated service at the end of this section. We have varied the load between 0.1 and 0.9 with steps of

Parameter	Notation	Values
Number of queues	N	2, 3, 4, 5
Load	ρ	0.1, 0.3, 0.5, 0.7, 0.9, 0.99
SCV interarrival times	$cv^2_{A_i}$	0.25, 1, 2
SCV service times	$cv^2_{B_i}$	0.25, 1
SCV switch-over times	$cv^2_{S_i}$	0.25, 1
Imbalance interarrival times	I_{A_i}	1, 5
Imbalance service times	I_{B_i}	1, 5
Ratio service and switch-over times	I_{S_i/B_i}	1, 5

Table 1: Test bed used to compare the approximation to exact results.

0.2, and included $\rho = 0.99$ to analyse the limiting behaviour of our approximation when the load tends to 1. The SCV of the interarrival times, $cv^2_{A_i}$, is varied between 0.25 and 2. In case of non-Poisson arrivals, i.e. $cv^2_{A_i} \neq 1$, the exact values have been established through extensive simulation because they cannot be obtained in an analytic way. In these simulations we fit a phase-type distribution to the first two moments of the interarrival times, as described in Examples 2 and 3. For service times and switch-over times, only SCVs of 0.25 and 1 are considered. SCVs greater than 1 are less common in practice and are discussed separately from the test bed later in this section. The imbalance in interarrival times and service times, I_{A_i} and I_{B_i}, is the ratio between the largest and the smallest mean interarrival/service time. The interarrival times are determined in such a way, that the overall mean is always 1, λ_1 is the largest and λ_N the smallest, and the steps between the λ_i are linear. E.g., for $N = 5$ and $I_{A_i} = 5$ we get $\lambda_i = 2 - i/3, i = 1, \ldots, 5$. The mean service times $\mathbb{E}[B_i]$ increase linearly in $i = 1, \ldots, N$, with $\mathbb{E}[B_N] = I_{B_i}\mathbb{E}[B_1]$ (so $\mathbb{E}[B_1]$ is the smallest mean service time). They follow from the relation $\sum_{i=1}^{N} \lambda_i \mathbb{E}[B_i] = \rho$. E.g., for $N = 5$, and $I_{A_i} = I_{B_i} = 5$ we get $\mathbb{E}[B_i]/\rho = 3i/35$. The last parameter that is varied in the test bed, is the ratio between the mean switch-over times and the mean service times, $I_{S_i/B_i} = \frac{\mathbb{E}[S_i]}{\mathbb{E}[B_i]}$. The total number of systems analysed is $4 \times 6 \times 3 \times 2^5 = 2304$. A system consisting of N queues results N mean waiting times, $\mathbb{E}[W_1], \ldots, \mathbb{E}[W_N]$, so in total these 2304 systems yield 8064 mean waiting times. The absolute relative errors, defined as $|o - e|/e$, where o stands for observed (approximated) value, and e stands for expected (exact) value, are computed for all these 8064 queues. Table 2 shows these relative errors (times 100%) categorised in bins of 5%. In this table, and in all other tables, results for systems with a different number of queues are displayed in separate rows. The reader should keep in mind that the statistics in each row are based on $\frac{1}{4} \times 2304 \times N$ absolute relative errors, where N is the number of queues used in the specified row. Table 2 shows that, e.g., 98.84% of the approximated mean waiting times in polling systems consisting of 3 queues deviate less than 5% from their true values. From Table 2 it can be concluded that the approximation accuracy increases with the number of queues in a polling system. More specifically, for systems with more than 2 queues, no approximation errors are greater than 10%, and the vast majority is less than 5%. The mean relative errors for $N = 2, \ldots, 5$ are respectively 2.18%, 0.93%, 0.70% and 0.57%. It is also noteworthy, that 193 out of the 2304 systems yield exact results. All of these 193 systems have Poisson input, and all of them – except for one – are symmetric. The only asymmetric case for which our approximation yields an exact result, happens to satisfy constraints (3.14) and (3.15).

In Table 3 the mean relative error percentages are shown for a combination of input parameter settings. The number of queues is always varied per row, while per column another input parameter is

varied. This way we can find in more detail which (combinations of) parameter settings result in large approximation errors. In Table 3(a) the load ρ is varied, and it can be seen that for a load of $\rho = 0.7$ the approximation is least accurate. E.g., the mean relative error of all approximated waiting times in polling systems consisting of 3 queues with a load of $\rho = 0.7$ is 1.69%. Table 3(b) shows the impact of the SCV of the interarrival times on the accuracy. Especially for systems with more than 2 queues the accuracy is very satisfactory, in particular for the case $cv^2_{A_i} = 1$. In Table 3(c) the impact of imbalance in a polling system on the accuracy is depicted, and, as could be expected, it can be concluded that a high imbalance in either service or interarrival times has a considerable, negative, impact on the approximation accuracy. Polling systems with more than 2 queues are much less bothered by this imbalance than polling systems with only 2 queues.

N	$0-5\%$	$5-10\%$	$10-15\%$	$15-20\%$
2	86.46	10.24	2.78	0.52
3	98.84	1.16	0.00	0.00
4	99.78	0.22	0.00	0.00
5	99.93	0.07	0.00	0.00

Table 2: Errors of the approximation applied to the 2304 test cases with exhaustive service, as described in Section 4, categorised in bins of 5%.

N	Load (ρ)					
	0.10	0.30	0.50	0.70	0.90	0.99
2	0.31	1.81	3.41	4.17	2.70	0.67
3	0.16	0.84	1.44	1.69	1.07	0.39
4	0.13	0.68	1.14	1.28	0.73	0.25
5	0.11	0.57	0.94	1.03	0.57	0.22

(a)

N	SCV interarrival times ($cv^2_{A_i}$)		
	0.25	1	2
2	2.27	1.76	2.50
3	1.36	0.52	0.92
4	1.13	0.29	0.69
5	0.97	0.19	0.56

(b)

N	Imbalance interarrival and service times			
	$I_{A_i} = 1, I_{B_i} = 1$	$I_{A_i} = 1, I_{B_i} = 5$	$I_{A_i} = 5, I_{B_i} = 1$	$I_{A_i} = 5, I_{B_i} = 5$
2	0.69	2.92	2.80	2.30
3	0.65	1.27	0.75	1.06
4	0.56	0.89	0.62	0.73
5	0.49	0.69	0.53	0.59

(c)

Table 3: Mean relative approximation error, categorised by number of queues (N) and total load of the system (a), SCV interarrival times (b), and imbalance of the interarrival and service times (c).

4.3 Miscellaneous other cases

More queues. In this subsection we discuss several cases that are left out of the test bed because they might not give any new insights, or because the combination of parameter values might be rarely found in practice. Firstly, we discuss polling systems with more than 5 queues briefly. Without listing the actual results, we mention here that the approximations become more and more accurate when letting N grow larger, and still varying the other parameters in the same way as is described in Table 1. For $N = 10$ already, all relative errors are less than 5%, with an average of less than 0.5% and it only gets smaller as N grows further.

More variation in service times and switch-over times. In the test bed we only use SCVs 0.25 and 1 for the service times and switch-over times, because these seem more relevant from a practical point of view. As the coefficient of variation grows larger, our approximation will become less accurate. E.g., for Poisson arrivals we took $cv^2_{B_i} \in \{2, 5\}$, $cv^2_{S_i} \in \{2, 5\}$, and varied the other parameters as in our test bed (see Table 1). This way we reproduced Table 2. The result is shown in Table 4 and indicates that the quality of our approximation deteriorates in these extreme cases. The mean relative errors for $N = 2, \ldots, 5$ are respectively 3.58%, 1.78%, 1.07%, and 0.77%, which is still very good for systems with such high variation in service times and switch-over times. For non-Poisson input, no investigations were carried out because the results are expected to show the same kind of behaviour.

N	$0-5\%$	$5-10\%$	$10-15\%$	$15-20\%$
2	74.22	14.84	6.51	2.08
3	89.76	7.29	2.08	0.69
4	94.53	4.56	0.91	0.00
5	97.71	2.19	0.10	0.00

Table 4: Errors of the approximation applied to the 768 test cases with Poisson arrival processes and high SCVs of the service times and switch-over times, categorised in bins of 5%.

Small switch-over times. Systems with small switch-over times, in particular smaller than the mean service times, also show a deterioration of approximation accuracy - especially in systems with 2 queues. In Figure 2 we show an extreme case with $N = 2$, service times and switch-over times are exponentially distributed with $\mathbb{E}[B_i] = \frac{9}{40}$ and $\mathbb{E}[S_i] = \frac{9}{200}$ for $i = 1, 2$, which makes the mean switch-over times 5 times *smaller* than the mean service times. Furthermore, the interarrival times are exponentially distributed with $\lambda_1 = 5\lambda_2$. In Figure 2 the mean waiting times of customers in both queues are plotted versus the load of the system. Both the approximation and the exact values are plotted. For customers in Q_1 the mean waiting time approximations underestimate the true values, which leads to a maximum relative error of -11.2% for $\rho = 0.7$ ($\mathbb{E}[W_{1,app}] = 0.43$, whereas $\mathbb{E}[W_1] = 0.49$). The approximated mean waiting time for customers in Q_2 is systematically overestimating the true value. The maximum relative error is attained at $\rho = 0.5$ and is 28.8% ($\mathbb{E}[W_{1,app}] = 0.41$, whereas $\mathbb{E}[W_1] = 0.52$). Although the relative errors are high in this situation, the absolute errors are still rather small compared to the mean service time of an individual customer. This implies that the mean *sojourn time* is already much better approximated. Nevertheless, this example illustrates one of the situations where our approximation gives unsatisfactory results.

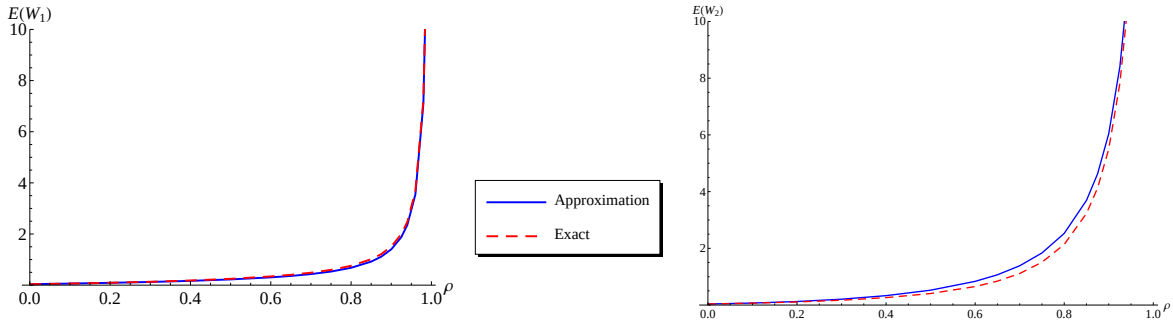

Figure 2: Approximated and exact mean waiting times for a two-queue polling system with small switch-over times.

4.4 Comparison with existing approximations

For non-exponential interarrival times hardly any good alternative approximations exist. In [10, 17] it is suggested to use the HT limit (3.12) as an approximation, but the accuracy is only found to be acceptable for $\rho > 0.8$. Another approximation for the mean waiting time in polling systems with non-exponential interarrival times uses the limit for $S \to \infty$ [21, 22]. This approximation is usable if either the total setup time in the system is large and the setup times have low variance, or the total setup time in the system is large and the system is in heavy traffic. The approximation discussed in the present paper is exact in all these limiting cases, but performs much better for systems under less extreme conditions. This makes our approximation the only one which can be applied under all circumstances.

For polling systems with Poisson arrivals, several alternative approximations have been developed in existing literature. The best one among them (see, e.g., [2, 4, 7]) uses the relation $\mathbb{E}[W_i] = (1 \pm \rho_i)\mathbb{E}[C_i^{res}]$, where C_i is the cycle time, starting at a visit *completion* to Q_i when service is exhaustive, and starting at a visit *beginning* for gated service. By $\pm$ we mean $-$ for exhaustive service, and $+$ for gated service. The mean residual cycle time, $\mathbb{E}[C_i^{res}]$, is assumed to be equal for all queues, i.e. $\mathbb{E}[C_i^{res}] \approx \mathbb{E}[C^{res}]$, and can be found by substituting $\mathbb{E}[W_i] \approx (1 \pm \rho_i)\mathbb{E}[C^{res}]$ in the pseudo-conservation law (3.13). We have used this PCL-based approximation to estimate the mean waiting times of all queues in the test bed described in Table 1, but taking only the 768 cases where $C_{A_i}^2 = 1$. Table 5 shows the mean relative errors for our approximation (a) and the PCL approximation (b), categorised in bins of 5% as was done before in Table 2. From these tables (and from other performed experiments that are not mentioned for the sake of brevity) it can be concluded that for $N > 2$ both approximation have almost the same accuracy, our approximation being slightly better for small values of ρ, and the PCL approximation being slightly better for high values of ρ (both methods are asymptotically exact as $\rho \uparrow 1$). However, for $N = 2$ our method suffers greatly from imbalance in the system, whereas the PCL approximation proves to be more robust.

4.5 Gated service

Until now we have only shown and discussed approximation results for polling systems with exhaustive service. The complete test bed described in Table 1 has also been analysed for polling systems where each queue receives gated service. As can be seen in Table 6, the overall quality of the approximation is good, but worse than for polling systems with exhaustive service. More details on the

N	$0-5\%$	$5-10\%$	$10-15\%$
2	89.32	9.11	1.56
3	100.00	0.00	0.00
4	100.00	0.00	0.00
5	100.00	0.00	0.00

(a)

N	$0-5\%$	$5-10\%$	$10-15\%$
2	96.09	2.86	1.04
3	99.31	0.69	0.00
4	100.00	0.00	0.00
5	100.00	0.00	0.00

(b)

Table 5: Errors of the approximation applied to the 768 test cases with Poisson input, categorised in bins of 5%. In (a) the percentages of mean relative errors in each bin are shown for our approximation, in (b) results are shown for the PCL approximation.

reason for these inaccuracies can be found in Table 7, which is the equivalent of Table 3 for gated service. Table 7(b) illustrates that there is now a huge difference between systems with Poisson arrivals, and systems with non-Poisson arrivals. For the cases with $cv^2_{A_i} = 1$, the approximation is extremely accurate, even for two-queue polling systems. The accuracy in cases with $cv^2_{A_i} \neq 1$ is worse, which is caused by the assumptions that are made to approximate the LT limit (3.10). Firstly, the decomposition (3.2) does not hold for non-Poisson arrivals, and secondly, the terms $\mathbb{E}[I_i^{res}]$ and $\frac{\mathbb{E}[V_i I_i]}{\mathbb{E}[I_i]}$ in this decomposition have only been approximated. For exhaustive service, these assumptions do not have much negative impact on the accuracy, but apparently, for gated service, they do. The mean relative errors for $N = 2, \ldots, 5$ queues are respectively 2.70%, 2.25%, 1.90%, and 1.63%. The imbalance of the mean interarrival and service times hardly influences the accuracy of the approximation, as can be concluded from Table 7(c).

If we consider the 768 cases with Poisson arrivals only, the mean relative errors of our approximation for $N = 2, \ldots, 5$ are respectively $0.34\%, 0.17\%, 0.10\%$, and 0.08%. This accuracy is even better than the one achieved by the PCL approximation.

N	$0-5\%$	$5-10\%$	$10-15\%$	$15-20\%$
2	82.55	12.33	2.95	1.56
3	85.42	10.53	3.13	0.81
4	88.85	8.46	2.43	0.26
5	92.22	6.60	1.15	0.03

Table 6: Errors of the approximation applied to the 2304 test cases with gated service, as described in Subsection 4.5, categorised in bins of 5%.

5 Further research topics

The research that is done in the present paper can be extended in many different directions. In this section we discuss some possibilities that we find most relevant.

N	Load (ρ)					
	0.10	0.30	0.50	0.70	0.90	0.99
2	2.64	4.55	4.31	3.10	1.25	0.37
3	2.03	3.78	3.68	2.68	1.04	0.30
4	1.62	3.14	3.13	2.32	0.92	0.28
5	1.35	2.67	2.71	2.03	0.81	0.21

(a)

N	SCV interarrival times ($cv^2_{A_i}$)		
	0.25	1	2
2	4.72	0.34	3.05
3	4.06	0.17	2.53
4	3.45	0.10	2.16
5	2.98	0.08	1.84

(b)

N	Imbalance interarrival and service times			
	$I_{A_i} = 1, I_{B_i} = 1$	$I_{A_i} = 1, I_{B_i} = 5$	$I_{A_i} = 5, I_{B_i} = 1$	$I_{A_i} = 5, I_{B_i} = 5$
2	2.76	2.64	2.81	2.59
3	2.28	2.25	2.27	2.21
4	1.93	1.91	1.90	1.87
5	1.64	1.66	1.64	1.58

(c)

Table 7: For gated service: mean relative approximation error, categorised by number of queues (N) and total load of the system (a), SCV interarrival times (b), and imbalance of the interarrival and service times (c).

Higher moments. Firstly, a logical follow-up step would be to use the same approach to find approximations for higher moments of the waiting time distribution as well. This might prove to be a hard exercise, since the LT limit of $\mathbb{E}[W_i^2]$ is unknown and, although its derivation might follow the same lines as in Section 3, it probably requires substantially more effort. Also, the HT limit of $\mathbb{E}[W_i^2]$ is unknown, although some research in this area has already been done and in [10] a strong conjecture is given for the limiting distribution of W_i as $\rho \uparrow 1$.

Another question that remains to be investigated, is the required form of the interpolation, as (2.1) is surely not adequate to approximate higher moments of $\mathbb{E}[W_i]$.

Other service disciplines. In the present paper, only exhaustive and gated service are discussed. In order to obtain results for polling systems with some queues receiving exhaustive service, and others receiving gated service, only minor modifications should be made, but we leave this to the reader. It would be more challenging to generalise the approximation to a wider variety of service disciplines. In particular, it would be nice to have one expression for the mean waiting time of customers in a queue with an arbitrary branching-type service discipline (cf. [12]). The *exhaustiveness* of a branching-type service discipline (cf. [20]) might appear in this expression. Gated and exhaustive are both branching type service disciplines, but are discussed separately in the present paper. The HT limit can most likely be established for arbitrary branching type service disciplines (see conjectures in [10]), so the question that remains is whether the LT limit can be found in a similar way.

Optimisation. One of the main reasons to choose (2.1) as form of the interpolation, besides its asymptotic correctness, is its simplicity. Having this exact and simple expression for the approximate mean waiting times, makes it very useful for optimisation purposes. In production environments, one can, for example, determine what the optimal strategy is to combine orders of different types (i.e., determine what queue customers should join). Because general arrivals are supported, one can determine optimal sizes of batches in which items are grouped and sent to a specific machine. The simplicity of (2.1) makes it possible for a manager to create a handy Excel sheet that can be used by operators to compute all kind of optimal parameter settings. No difficult computations are required at all, so a large variety of users can use the approximation.

In the present paper the accuracy of the approximation has been investigated and has been found to be very good in most situations. Another advantage of our approximation regarding optimisation purposes, is that the general shape of the approximated curve follows the exact curve very closely. Even in cases where the relative errors are rather large, like in Figure 1, the shape of the actual curves is still very well approximated. This means that plugging our approximation, instead of an exact expression if it had been available, in an optimisation function yields an optimum that should be close to the true optimum.

Polling Table. The interpolation based approximation can also be extended to polling systems where the visiting order of the queues is not cyclic. Waiting times in polling systems with so-called polling tables can be obtained in the same way as shown in the present paper. Both the LT and HT limits are not difficult to determine in this situation, and the interpolation follows directly from these limits.

Model. The form of the interpolation might be changed to improve the accuracy of approximations for cases that give less satisfactory results in the present form. E.g., one could try other functions than a second-order polynomial as numerator of (2.1). Alternatively, one could try to find a correction term which could be added to (2.1) to obtain better results for, e.g., two-queue polling systems. But most of all, if an *exact* LT limit of the mean waiting time in a polling system with non-Poisson arrivals could be found, the accuracy of the approximation in the case of gated service might be improved.

Acknowledgements

The authors wish to thank Onno Boxma for valuable discussions and for useful comments on earlier drafts of the present paper.

References

[1] J. P. C. Blanc and R. D. van der Mei. Optimization of polling systems with Bernoulli schedules. *Performance Evaluation*, 22:139–158, 1995.

[2] O. J. Boxma. Workloads and waiting times in single-server systems with multiple customer classes. *Queueing Systems*, 5:185–214, 1989.

[3] O. J. Boxma and W. P. Groenendijk. Pseudo-conservation laws in cyclic-service systems. *Journal of Applied Probability*, 24(4):949–964, 1987.

[4] D. Everitt. Simple approximations for token rings. *IEEE Transactions on Communications*, COM-34(7):719–721, 1986.

[5] M. J. Fischer, C. M. Harris, and J. Xie. An interpolation approximation for expected wait in a time-limited polling system. *Computers & Operations Research*, 27:353–366, 2000.

[6] S. W. Fuhrmann and R. B. Cooper. Stochastic decompositions in the $M/G/1$ queue with generalized vacations. *Operations Research*, 33(5):1117–1129, 1985.

[7] W. P. Groenendijk. Waiting-time approximations for cyclic-service systems with mixed service strategies. In *Proc. 12th ITC*, pages 1434–1441. North-Holland Publ. Co., Amsterdam, 1989.

[8] J. Keilson and L. D. Servi. The distributional form of Little's Law and the Fuhrmann-Cooper decomposition. *Operations Research Letters*, 9(4):239–247, 1990.

[9] H. Levy and M. Sidi. Polling systems: applications, modeling, and optimization. *IEEE Transactions on Communications*, 38:1750–1760, 1990.

[10] T. L. Olsen and R. D. van der Mei. Polling systems with periodic server routing in heavy traffic: renewal arrivals. *Operations Research Letters*, 33:17–25, 2005.

[11] M. I. Reiman and B. Simon. An interpolation approximation for queueing systems with Poisson input. *Operations Research*, 36(3):454–469, 1988.

[12] J. A. C. Resing. Polling systems and multitype branching processes. *Queueing Systems*, 13:409 – 426, 1993.

[13] L. E. Schrage and L. W. Miller. The queue $M/G/1$ with the shortest remaining processing time discipline. *Operations Research*, 14(4):670–684, 1966.

[14] B. Simon. A simple relationship between light and heavy traffic limits. *Operations Research*, 40(Supplement 2):S342–S345, 1992.

[15] H. Takagi. Queuing analysis of polling models. *ACM Computing Surveys (CSUR)*, 20:5–28, 1988.

[16] H. C. Tijms. *Stochastic models: an algorithmic approach*. Wiley, Chichester, 1994.

[17] R. D. van der Mei and E. M. M. Winands. A note on polling models with renewal arrivals and nonzero switch-over times. *Operations Research Letters*, 36:500–505, 2008.

[18] V. M. Vishnevskii and O. V. Semenova. Mathematical methods to study the polling systems. *Automation and Remote Control*, 67(2):173–220, 2006.

[19] W. Whitt. An interpolation approximation for the mean workload in a $GI/G/1$ queue. *Operations Research*, 37(6):936–952, 1989.

[20] E. M. M. Winands. *Polling, Production & Priorities*. PhD thesis, Eindhoven University of Technology, 2007.

[21] E. M. M. Winands. On polling systems with large setups. *Operations Research Letters*, 35: 584–590, 2007.

[22] E. M. M. Winands. Branching-type polling systems with large setups. *To appear in OR Spectrum*, 2009.