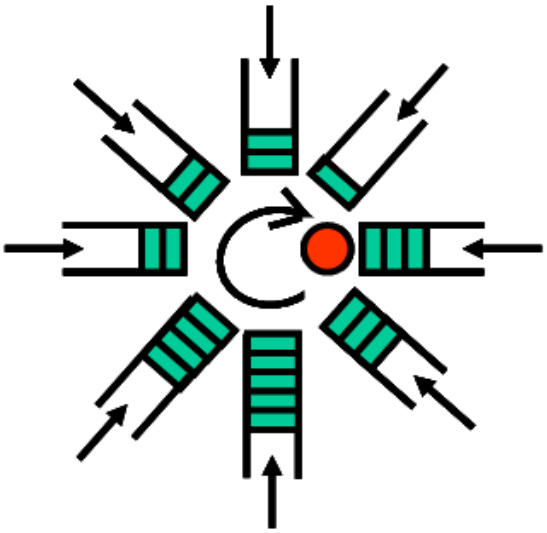
Internship report

Evaluation and optimization of polling systems



Author:
Jan-Pieter Dorsman

CWI
Centrum Wiskunde & Informatica

vrije Universiteit amsterdam

Supervisor CWI:
prof. dr. R.D. van der Mei

Supervisors VU:
dr. ir. E.M.M. Winands
A. Roubos, MSc

Internship report

Evaluation and optimization of polling systems

Author:
Jan-Pieter Dorsman

July 2010

Centrum Wiskunde & Informatica

Research Group PNA2
Science Park 123
1098 XG Amsterdam

Supervisor:
prof. dr. R.D. van der Mei

Vrije Universiteit Amsterdam

Faculty of Sciences
De Boelelaan 1081a
1081 HV Amsterdam

Supervisors:
dr. ir. E.M.M. Winands
A. Roubos, MSc

Preface

The end of the BMI (Business Mathematics and Informatics) Master program at the Vrije Universiteit in Amsterdam is marked by an internship carried out at an external business, industry or research facility. The present report contains the results of my internship done at the PNA2 (Probability & Stochastic Networks) research group of the Centrum Wiskunde & Informatica (CWI) in Amsterdam.

I would like to thank Rob van der Mei and Erik Winands for giving me the opportunity to undertake an internship at the CWI. Furthermore, their support and feedback are much appreciated. Thanks are also due to Alex Roubos, who provided valuable comments on earlier drafts of this internship report. I am indebted to Marko Boon, who placed parts of his sound polling simulation program at my disposal, which unquestionably saved a lot of work. Furthermore, I would like to express my gratitude to Rob van der Mei and Bert Zwart for offering me the opportunity to present the work of Chapter 2 at the Third Madrid Conference on Queueing Theory, held in Toledo, Spain. This has been an invaluable experience for me. Finally, I am grateful to all colleagues at the CWI for a great working atmosphere in general.

Amsterdam, July 2010

Abstract

Throughout this internship report, polling systems play a central role. Polling systems are queueing systems consisting of multiple queues, attended by a single server. The server can only serve one queue at a time. Whenever the server moves from one queue to another, a stochastic, non-zero switch-over time is incurred. The server never idles; even when there are no customers waiting in the whole system the server keeps roving between queues.
In the literature on these systems, often Poisson arrivals are assumed. In many applications, however, the Poisson assumption is not realistic. This motivates us to study polling models with renewal arrivals, which represent a far broader class of arrival streams.
Firstly, an approximation of the complete waiting time distribution in polling systems with renewal arrivals is derived. This approximation may act as a basis for design decisions within polling systems.
Secondly, polling systems with renewal arrivals and batch service are studied. Using characteristics of the obtained distributional approximation, the question is addressed how sizes of batches should be chosen in order to optimize the waiting time.

Contents

Chapter 1

Introduction

1.1 About the CWI

Founded in 1946, CWI is the national research center for Mathematics and Computer Science in the Netherlands. More than 170 full professors have come from CWI, of whom 120 still are active. CWI's strength is the discovery and development of new ideas, and the transfer of knowledge to academia and to Dutch and European industry. This results in importance for our economy, from payment systems and cryptography to telecommunication and the stock market, from public transport and internet to water management and meteorology.

With its 55 permanent research staff, 40 postdocs and 65 PhD students, CWI lies at the heart of European research in mathematics and computer science. Researchers at CWI are able to fully concentrate their efforts on their scientific work, and to build an international network of peers. More than half of the permanent research staff maintains close contact with universities as part-time professors. The personal and institutional research networks strengthen CWI's positions and serve as a magnet for attracting talent. CWI researchers come from more than 25 countries world-wide.

CWI was a birthplace of the world-wide internet. The national domain name cwi.nl was the first one ever issued anywhere. CWI helped with the development of the wing of the Fokker Friendship, chosen later as the most beautiful Dutch design of the 20th century. The popular language Python was invented at CWI, the language in which Google was developed. CWI applied combinatorial algorithms to the scheduling of the Dutch railway system. XML-databases were build to the needs of the Netherlands Forensic Institute and 3D visualization techniques to better detect cancer tumors.

1.2 Polling systems

The internship carried out at the PNA2 department within the CWI primarily encompassed research on polling systems. Consequently, polling systems will play a central role throughout this internship report. Polling systems are queueing systems consisting of multiple queues, attended by a single server. Each queue has its own customer arrival stream. When the server moves from one queue to another, a non-zero switch-over time is incurred in order to prepare for service at the next queue. For a graphical representation of a polling system in which the server switches through the queues in a cyclic manner, see Figure 1.1.
The analysis of polling systems is important in many real-life application areas, such as computer-communication systems, manufacturing systems and traffic systems. To give an insight into the applicability in these areas, we give an example of an application in each of these areas.

- Computer-communication: *Time-sharing computer systems* consist of multiple terminals, which share multi-drop lines in order to communicate with a central computer. To avoid transmission collisions, the central computer polls the terminals, one at a time, after which all pending data transactions of the polled terminal are handled.

- Manufacturing systems: The *stochastic economic lot scheduling problem* (SELSP) deals with the make-to-stock production of multiple standardized products on a single machine with limited capacity under random demands. This problem is analyzed extensively in [39].

- Traffic systems: The analysis of *signalized intersections* is the most obvious example of a traffic system application. Cars waiting in the several lanes represent customers waiting in queues and the green light is modelled by drawing the analogy with the server.

For more applications, see [41]. Because of the importance of analysis of polling systems in real-life applications, a huge body of literature on polling systems has developed since the late 1950s, starting with the papers of [21] and [22]. Literature overviews of polling systems can be found in [19, 30, 32].
Since the underlying models of the next chapters are slightly different — although they both can be classified as polling models —, we do not aim to give a general notation framework in this section; each of these chapters will have its own introduction and notation. Nevertheless, the several input parameters and characteristics of the polling model are to be discussed. By construction of the polling model, there are three stochastic components that are influential to the performance of the polling system.

Arrival processes. Customers arrive at their correspondent queues according to mutually independent arrival processes. In literature these arrival processes are

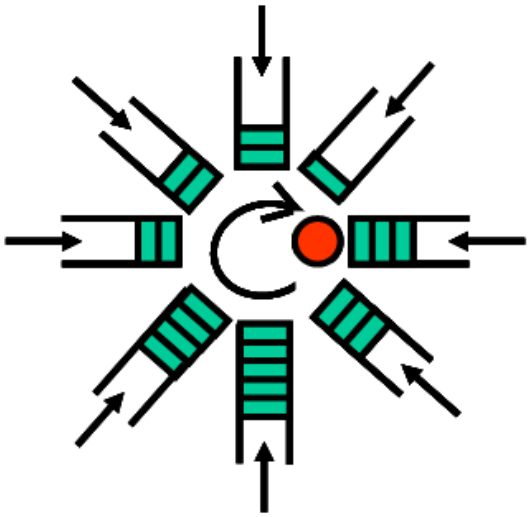


Figure 1.1: Polling system with cyclic service.

often assumed to be Poisson processes, i.e., the interarrival times of customers of the same type are exponentially distributed. However, the validity of the Poisson assumption may be very questionable in certain real-life applications. This is why throughout this report only renewal arrivals are assumed.

Service processes. Whenever a server *polls* a queue[1], it will commence a service period, in which a number of customers of the current queue is served. Each customer requires its own service time from the server. For customers of the same type, these service requirements are commonly assumed to be mutually independent, independent of the state of the system and identically distributed. Throughout the analysis following in this report, hardly any characteristics of the service time distributions are required other than the first two moments.

Switch-over processes. To switch from one queue to another, the server typically needs some time to prepare for offering service at the "new" queue. For each directed pair of queues, the switch-over times needed are assumed to be independent samples from some probability distribution. Again, we will generally require only the first two moments of these distributions.

Although these stochastic components are very important in the description of the polling model, they do not fully identify a polling system. The performance of such a system also depends, other than the number of queues, on the following variables.

[1]The moment when the server concludes the switch-over period leading to being set up for service at the current queue.

Buffer size. Since in most applications there is ample room for customers to wait for service in the system, the queues are usually assumed to have an infinite buffer size. Obviously this assumption does not coincide with reality, however it still is very applicable due to the fact that the buffer size is hardly ever a limiting factor.

Service discipline. The service discipline specifies how many customers are served by the server between two switch-over periods. Results found independently by [13] and [28] show that performance analysis is tractable if the service discipline satisfies a so-called branching property. Analysis will generally become hardly tractable otherwise. In such case, exact results on performance measures such as the waiting time are only available for some special cases, such as a symmetric or a two-queue system. The branching property is defined as follows:

Property 1.2.1. *If the server in a polling system with N queues arrives at queue i to find l_i customers there, then during the course of a server's visit, each of these l_i customers will effectively be replaced in an i.i.d. manner by a random population having probability generating function (PGF) $h_i(z) = h_i(z_1, \ldots, z_N)$, which can be any N-dimensional probability generating function.*

The most important of the service disciplines satisfying Property 1.2.1 are

- The *exhaustive* service discipline: when the server start service at a queue, it will continue service as long as there are customers waiting in the queue. It will stop service and commence a switch-over procedure if and only if there are no customers waiting in the current queue anymore.
- The *gated* service discipline: when the server starts service at a queue, it will only serve the customers that were present at the start of the service period before it switches to another queue.

Common service disciplines that do not satisfy Property 1.2.1 are

- The *customer-limited* service discipline: the server will continue serving customers until either the queue is empty or a certain prespecified number of customers have been served during the service period, whichever happens first.
- The *time-limited* service discipline: the server will continue serving customers until either the queue is empty or a certain prespecified amount of time has passed since the start of the service period, whichever happens first.

Server routing. The order in which the server polls the queues is determined by the routing mechanism. A distinction can be made between dynamic and static routing mechanisms. In dynamic routing mechanisms, the choice of the queue to be switched to is dependent on the state of the system, whereas it is not when

the server adopts a static routing mechanism. An example of a dynamic routing mechanism policy is the so called 'serve-longest-queue' policy, where the server will always switch to the queue having the most customers waiting.
In this report the focus however is on static routing mechanisms. Static routing mechanisms include

- The *cyclic* routing mechanism: the server switches through the queues in a cyclic manner. That is, the server will typically serve all queues repeatedly in the order in which the queues were numbered. The time between two departure epochs of the server at the same queue is then called a cycle.
- Routing according to a *polling table*: the server switches through the queues periodically according to some prespecified order. The cyclic routing mechanism may be seen as a specific case of polling table routing.
- The *Markovian* routing mechanism: randomness is introduced in the decision for the server which queue to switch to after a service period. Probabilities p_{ij} are introduced, which denote the probability that the server will proceed to queue j after having served queue i. Of course, we have that $\sum_{j=1}^{N} p_{ij} = 1$ for all i.

Queueing discipline. The order in which customers are served by the server within a service period is specified by the queueing discipline. Mostly a First-Come-First-Served (FCFS) discipline is assumed, which means that the customer with the least recent arrival time will be served first. It should be noted that the distribution of the length of a queue is independent of its queueing discipline, provided the queueing discipline does not depend on the service times. In case of Poisson arrivals, this also means that the mean waiting time is independent of the queueing discipline by Little's law. However, the waiting time distribution as a whole does depend on the queueing discipline at all times.

The three stochastic components together with these variables totally identify a single-server polling system. Note that each queue can have its own service and queueing discipline. We will conclude this section with a note on the stability of the polling system. In case of the exhaustive or gated service discipline and a cyclic routing mechanism, $\rho < 1$ and $\mathbb{E}[S] < \infty$ are necessary and sufficient conditions for stability [9]. Here, ρ is the load of the system (the sum of the arrival rates times the mean service times of each customer-type), and $\mathbb{E}[S]$ is the total switch-over time needed by the server during a complete cycle, which in practice is of course finite.
In case of Poisson arrivals, cyclic routing and quantity-limited service, the following equation is a necessary and sufficient condition for stability:

$$\rho + \mathbb{E}[S] \max_{i \in \{1,\ldots,N\}} \left(\frac{\lambda_i}{k_i} \right) < 1, \tag{1.1}$$

where λ_i and k_i are the mean arrival rate and the quantity limit of queue i respectively (cf. [12]).

1.3 Research objectives

A central role in analysis of polling systems is taken by the evaluation of the waiting time. Waiting time is often seen as an important performance measure of a queueing system. In addition, waiting time can become a critical issue in systems involving human customers, or in systems where perishable goods are produced. To this end, it is very helpful to be able to predict characteristics of the waiting time based on the parameters of the polling system.
In literature, often Poisson arrivals are assumed, mainly because of the reduction of analytic complexity this assumption comes with. However, the assumption of exponentially distributed interarrival times may not always be valid and realistic. Moreover, most of the literature focuses solely on the first moment of the waiting time distribution. Remarkably little attention has been paid to the evaluation of the complete distribution of the delay in the queues. This gives birth to the first research objective.

Research objective 1. *Development of a closed-form approximation of the complete waiting time distribution in polling systems under the assumption of renewal arrival processes.*

The assumption of renewal arrivals implies that the interarrival times of customers in each queue should be independent and identically distributed. However, they are not necessarily exponential.
For systems with renewal arrivals, expressions or numerical algorithms for the exact computation of waiting time characteristics do not exist in general. A closed-form approximation of the mean waiting time was derived in [2]. However, no closed-form approximations exist for the *complete* waiting time distribution. Therefore, we aim to obtain a closed-form approximation of the complete waiting time distribution. When such an approximation is found, it may act as a basis for design decisions within polling systems. For example, one may think of a model variation where customers do not arrive at the polling system directly, but first form batches, after which those batches enter the polling system as a whole. When assuming that the service requirements of such batches are independent of the number of underlying customers, one is confronted with the question how large batch sizes should be chosen. This issue gives rise to the second research objective.

Research objective 2. *Development of accurate and efficient methods to optimize the size of batches in polling systems with renewal arrivals and batch service.*

The methods obtained may be used for implementation in decision support systems on the one hand, and may give insights into how optimal batch sizes behave in the several polling system's parameters on the other hand.

1.4 Structure of the report

The present internship report is set up as follows. Chapter 2 contains a study that aims to fulfill the first research objective as stated in the previous section. In Chapter 3 the methods found by research based on the second research objective are explained. As the research topics of the latter two chapters are quite distinct and the models studied have subtle, but very important differences, both chapters contain their own brief introductions and literature studies, model descriptions, notations and suggestions for further research.

Chapter 2

Evaluation

In this chapter we aim to fulfill the first research objective as formulated in Section 1.3. The focus is on waiting time distributions in cyclic polling models with renewal arrivals, general service and switch-over times, and exhaustive service at each of the queues. The assumption of renewal arrivals prohibits an exact analysis and reduces the available analytic results to heavy traffic asymptotics, limiting results for large switch-over times and large numbers of queues, and some numerical algorithms. Motivated by this, the goal of the present study is to propose a new method for deriving simple closed-form approximations for the complete waiting time distributions that works well for arbitrary load values. Extensive simulation results show that the approximations are highly accurate over a wide range of parameter settings.

2.1 Introduction

In this chapter, the waiting time in cyclic polling systems with renewal arrival streams and exhaustive service is considered. Arrival streams are not (necessarily) Poisson, while the assumption of Poisson arrivals is made in most of the literature on polling systems. For systems with renewal arrivals, solutions for performance metrics, such as moments and distributions of waiting times do not exist in general. Exact expressions and algorithms that do exist, are only valid in certain limiting cases, e.g., when the load tends to one [26] or the total switch-over time in a cycle tends to infinity [40]. These expressions and algorithms can be used as an approximation in general, but practice shows that they often become inaccurate quickly as soon as the limiting condition is violated. Closed-form approximations are available for the mean waiting time [2], however there are none available that approximate the complete waiting time distribution. Even in case of Poisson arrivals, generally little attention is paid to the complete waiting time distribution rather than the mean waiting time. As an exception, assuming Poisson arrivals Choudhury and Whitt [5] propose an efficient numerical algorithm to calculate tail probabilities of the waiting times based on numerical transform inversion, for models that satisfy a

multi-type branching structure [28]. For models that violate the branching structure, more computationally intensive algorithms exist [1, 18]. A common drawback of these numerical algorithms is that they only give limited insight into how the waiting time distribution reacts to changes in the system parameters.

The goal of this chapter is to propose a new method for deriving a closed-form approximation of the waiting time distribution for arbitrary values of the load. The approach taken is that of combining known heavy traffic (HT) asymptotics for the waiting time distributions [23, 26], which are shown to work well when the system is heavily loaded, with a recently developed approximation for the mean waiting times by Boon et al. [2], which works well for the whole range of load values. This chapter presents the first *closed-form* approximation of the waiting time *distribution* in polling systems with renewal arrivals. This approximation is shown to be exact in the known limiting cases, and extensive experimentation with simulations shows that the approximation is highly accurate for a wide range of parameter settings. We emphasize that the strength of this combined approach lies in its striking simplicity and the fact that it leads to approximations in closed form, which opens up many possibilities for generalization of the approach to other polling models (e.g., with more general branching-type service policies, and with non-cyclic periodic server routing) and for optimization of the system performance.

In Section 2.2 the model and notation are introduced. In Section 2.3 the main result of this chapter, the distributional approximation is presented, and the idea behind it explained. Section 2.4 discusses several properties of the obtained approximation. In Section 2.5, the approximation is verified by an extensive simulation study. Finally, Section 2.6 discusses suggestions for further research.

2.2 Model description and notation

Consider a polling system consisting of $N \geq 1$ queues, $Q_1, \ldots, Q_N$, with an infinite-sized buffer at which customers arrive. Throughout this chapter, a queue index i is understood as $((i-1) \mod N)+1$, e.g., Q_{N+1} actually refers to Q_1. Customers in the different queues are waiting to be processed by a single server. At each queue, customers arrive according to a renewal process. Interarrival times of customers at Q_i are denoted by the random variable A_i, the customers are assumed to arrive at rate $\lambda_i = \frac{1}{\mathbb{E}[A_i]}$. The total arrival rate to the system is denoted by $\Lambda = \sum_{i=1}^{N} \lambda_i$. Within a queue, customers are served on a First-Come-First-Served (FCFS) basis. The service time of a type-i customer at Q_i is denoted by the random variable B_i with k^{th} moment $\mathbb{E}[B_i^k]$, and its waiting time in Q_i by the random variable W_i with k^{th} moment $\mathbb{E}[W_i^k]$, $k > 0$. We use B to denote the service time of an arbitrary customer entering the system, with $\mathbb{E}[B^k] = \sum_{i=1}^{N} \frac{\lambda_i}{\Lambda} \mathbb{E}[B_i^k]$.
Queues are served according to an *exhaustive* service discipline, i.e., a server will

not start switching to another queue before the customers in the current queue are all served, including the ones that arrived during the service period. Whenever a server has finished service at Q_i, it will switch to Q_{i+1}. We define a cycle at Q_i as the time between two successive departures of the server at Q_i. In order to switch from Q_i to Q_{i+1}, the server needs a switch-over time, of which the duration is denoted by the random variable S_i with k^{th} moment $\mathbb{E}[S_i^k]$, $k > 0$. $S = \sum_{i=1}^{N} S_i$ denotes the total switch-over time in a cycle. Throughout it is assumed that $\mathbb{E}[S] > 0$ and that all interarrival times, service times and switch-over times are mutually independent and independent of the state of the system. The load offered to Q_i is denoted by $\rho_i = \lambda_i \mathbb{E}[B_i]$, $1 \leq i \leq N$. The total load in the system is denoted by $\rho = \sum_{i=1}^{N} \rho_i$. A necessary and sufficient condition for the stability of the described system reads $\rho < 1$ [9]. The waiting time at Q_i is defined as the time between the arrival of an arbitrary customer in the system and the moment when he is taken into service.

Throughout, it may be convenient to scale a system such that a certain load is achieved. This scaling is done by keeping the service time distributions fixed and varying the rates of the renewal processes. In particular, it proves convenient to denote with $\hat{x}$ the value of each variable x that is a function of ρ evaluated at $\rho = 1$. In that case $\hat{A}_i$ denotes the interarrival time of Q_i customers evaluated at $\rho = 1$. Then, scaling to a load $\rho < 1$ is done by taking the random variable $A_i := \frac{\hat{A}_i}{\rho}$.

Finally, we introduce some notation. The residual length of a random variable X is denoted by X^{res}, with $\mathbb{E}[X^{res}] = \frac{\mathbb{E}[X^2]}{2\mathbb{E}[X]}$. The squared coefficient of variation (SCV) of a random variable X, $\frac{\mathrm{Var}[X]}{\mathbb{E}[X]^2}$, is denoted by c_X^2. We define $\sigma^2 = \sum_{i=1}^{N} \hat{\lambda}_i(\mathrm{Var}[B_i] + c_{A_i}^2 \mathbb{E}[B_i]^2)$ and $\delta = \sum_{j=1}^{N} \sum_{k=j+1}^{N} \hat{\rho}_j \hat{\rho}_k$. In the case of Poisson arrivals, the former can be simplified to $\sigma^2 = \mathbb{E}[B^2]/\mathbb{E}[B]$. We refer to a polling system as symmetric, when the queues in the polling system share the same interarrival time distributions, service time distributions and switch-over time distributions. Of course, a system is asymmetric when this condition is violated. Throughout, the notation $\xrightarrow{d}$ means convergence in distribution. Indicator functions are used in the form of $\mathbb{1}_{\{A\}}$, which evaluate to one if condition A holds, zero otherwise. Finally, when a random variable X is said to be gamma distributed with shape parameter α and inverse scale parameter μ, its density function is given by $f_X(x) = e^{-\mu x}\mu^\alpha x^{\alpha-1}\mathbb{1}_{\{x\geq 0\}}/\Gamma(\alpha)$, where $\Gamma(\alpha) = \int_0^\infty e^{-x}x^{\alpha-1}dx$.

2.3 Derivation of the approximation

The two key ingredients of the distributional approximation will be the HT diffusion approximation for the waiting time by Olsen and Van der Mei [26], and the mean waiting time approximation by Boon et al. [2] for a general load. The HT diffusion approximation will be refined such that its mean concides with the mean waiting

time approximation, while the diffusion approximation remains unchanged in the case of HT after refinement. The two ingredients are given first, after which the main result is derived and presented. Although not stated explicitly, it follows naturally from [26] that

$$(1-\rho)W_i \xrightarrow{d} UI_i, \quad \rho \uparrow 1, \tag{2.1}$$

where U is a uniformly distributed random variable on [0,1], and I_i a gamma distributed random variable with shape parameter α and inverse scale parameter μ_i as follows:

$$\alpha = \frac{2\mathbb{E}[S]\delta}{\sigma^2} + 1 \text{ and } \mu_i = \frac{2\delta}{(1-\hat{\rho}_i)\sigma^2}.$$

Let $\mathbb{E}[W_i]$ denote the mean waiting time of a type-i customer at Q_i, $1 \leq i \leq N$. Then, the work of Boon et al. [2] shows that an accurate approximation $\mathbb{E}[W_{i,Boon}]$ of $\mathbb{E}[W_i]$ as a function of ρ is as follows:

$$\mathbb{E}[W_{i,Boon}] = \frac{K_0 + K_{1,i}\rho + K_{2,i}\rho^2}{1-\rho}, \tag{2.2}$$

where the constants $K_0, K_{1,i}$ and $K_{2,i}$ depend on several parameters of the polling system at hand:

$$\begin{aligned}
K_0 &= \mathbb{E}[S^{res}], \\
K_{1,i} &= \hat{\rho}_i((c_{A_i}^2)^4 \mathbb{1}_{\{c_{A_i}^2 \leq 1\}} + 2\frac{c_{A_i}^2}{c_{A_i}^2+1}\mathbb{1}_{\{c_{A_i}^2 > 1\}} - 1)\mathbb{E}[B_i^{res}] + \mathbb{E}[B^{res}] \\
&\quad + \hat{\rho}_i(\mathbb{E}[S^{res}] - \mathbb{E}[S]) - \frac{1}{\mathbb{E}[S]}\sum_{j=0}^{N-1}\sum_{k=0}^{j}\hat{\rho}_{i+k}\mathrm{Var}[S_{i+j}], \\
K_{2,i} &= \frac{1-\hat{\rho}_i}{2}\Big(\frac{\sigma^2}{2\delta} + \mathbb{E}[S]\Big) - K_0 - K_{1,i}.
\end{aligned}$$

An explanation of the derivation of $\mathbb{E}[W_{i,Boon}]$ can be found in Appendix A.

Our distributional approximation will be a refinement of the HT diffusion approximation given in (2.1). We assume that the waiting time distribution of Q_i can be well approximated by a product of a uniform[0,1] and a gamma random variable with shape parameters α_{ia} and inverse scale parameter μ_{ia}, divided by $(1-\rho)$. The property as presented in (2.1) shows that in HT the waiting time distribution can indeed be decomposed in a uniform part and a gamma part. Hence, for a significant load, this assumption is a natural one. Therefore, when suitable expressions for α_{ia} and μ_{ia} are found, a distributional approximation is obtained.

Refinement is done using the first moment approximation presented in (2.2). We initially impose two requirements on the refinement:

1. *In HT the refined approximation must coincide with the diffusion approximation of [26], i.e.,*

$$\frac{\alpha}{\alpha_{ia}} \to 1 \quad \textit{and} \quad \frac{\mu_i}{\mu_{ia}} \to 1$$

when the load tends to one.

2. *The expectation of the refined approximation coincides with the mean waiting time approximation of [2].*

There is an infinite number of feasible combinations of α_{ia} and μ_{ia} that satisfy these two requirements, hence we add a third requirement:

3. *The SCV of the refined approximating distribution matches the SCV of the HT diffusion approximation by [26]. In other words, the shape of the refined diffusion approximation matches the shape of the HT diffusion approximation.*

Together with this third requirement, there is just one feasible set of parameters left, which ends the search for suitable parameters. This results in the main result of this chapter, the following approximation for the waiting time distribution in polling systems with renewal arrivals and $\rho < 1$:

$$\mathbb{P}[W_i < x] \approx \mathbb{P}[UI_{i,app} < (1-\rho)x], \tag{2.3}$$

where U is a uniformly distributed random variable on [0,1], and $I_{i,app}$ a gamma random variable with parameters

$$\alpha_{ia} = \alpha_a = \frac{2\mathbb{E}[S]\delta}{\sigma^2} + 1 \text{ and } \mu_{ia} = \frac{2\mathbb{E}[S]\delta + \sigma^2}{2\sigma^2(1-\rho)\mathbb{E}[W_{i,Boon}]}.$$

It can be verified that the k^{th} moment of the obtained distributional approximation can be expressed as follows, for $k \geq 1$,

$$\mathbb{E}[W_{i,app}^k] = \frac{1}{(1-\rho)^k}\frac{1}{k+1}\prod_{i=0}^{k-1}\frac{\alpha_a + i}{\mu_{ia}} = \frac{2^k\mathbb{E}[W_{i,Boon}]^k}{k+1}\prod_{i=1}^{k}\frac{2\mathbb{E}[S]\delta + i\sigma^2}{2\mathbb{E}[S]\delta + \sigma^2}, \tag{2.4}$$

with α_a and μ_{ia} as defined above.

We end this section with several remarks about the obtained approximation. In the following section additional analytical justification for the approximation is presented.

Remark 1 (Olsen's approximation). A refined diffusion approximation for the distribution of the waiting time in polling systems with Poisson arrivals was presented by Olsen [25]. The HT diffusion approximation used by [25] for Q_i consists

of a uniformly distributed random variable on [0,1] "times" a gamma distribution with shape parameter $\frac{\mathbb{E}[S]\sum_{i=1}^{N}\rho_i(\rho-\rho_i)}{\sum_{i=1}^{N}\lambda_i(\mathrm{Var}[B_i]+\mathbb{E}[B_i]^2)}+1$ and inverse scale parameter $\frac{(1-\rho)\sum_{i=1}^{N}\rho_i(\rho-\rho_i)}{(1-\rho_i)\sum_{i=1}^{N}\lambda_i(\mathrm{Var}[B_i]+\mathbb{E}[B_i]^2)}$. Note that in HT these parameters coincide with the ones used in (2.1). Refinement is done using an approximation of the mean delay obtained by [10] for Poisson arrivals. Suggested by this mean delay approximation, Olsen adds an extra factor of ρ in the shape parameter, such that it becomes $\frac{\mathbb{E}[S]\sum_{i=1}^{N}\rho_i(\rho-\rho_i)}{\rho\sum_{i=1}^{N}\lambda_i(\mathrm{Var}[B_i]+\mathbb{E}[B_i]^2)}+1=\frac{2\mathbb{E}[S]\delta}{\sigma^2}+1$. The inverse scale parameter is changed accordingly such that the approximation satisfies the mean delay approximation in [10]. One can verify that in case of Poisson arrivals, the shape parameters of Olsen's approximation and our approximation coincide. Hence, the distributional approximation as given in this section generalizes Olsen's approximation to systems with renewal arrivals, and the presented derivation of the main result of this chapter creates intuition and justification behind the distributional approximation of [25].

Remark 2 (Information availability). The derived waiting time distribution approximation (2.3) only requires the first two moments of the interarrival, service and switch-over time distributions as an input, whereas the complete waiting time distribution generally depends on their complete distributions, even for Poisson arrivals. This makes the approximations useful for practical purposes, because in reality information about more than the first two moments is often hard to get.

Remark 3 (Applicability). Yet another view is provided by the notion that the derived approximation gives a procedure to estimate the complete waiting time distribution based on the mean waiting time and aggregate measures for imbalance δ and variability σ^2. In this regard, it is important to note that the mean waiting time can easily be measured in real-life applications, in contrast to higher moments or tail probabilities.

2.4 Analytical results

In [2] it is shown that the first moment of the distributional approximation is in line with several known exact results, which gives support for the quality of the approximation. That is, for Poisson arrivals the first moment satisfies the well-known pseudo-conservation laws and is exact in symmetric systems, vacation queues and general systems in light traffic (LT), i.e., where the load tends to zero. Moreover, the first moment is proved to give exact results for systems with general renewal arrivals in the asymptotic regimes of HT or infinite switch-over times. In the present section, comparable results are proved for higher moments of the distributional approximation.

Heavy-traffic. By construction, the distributional approximation is exact in systems with general renewal arrivals in HT. This property is very desirable from a practical perspective, since the proper operation of a system is particularly critical when the system is heavily loaded.

Large switch-over times. In case of deterministic switch-over times, the waiting time is only dependent on the total switch-over time in a cycle S rather than the marginal switch-over times S_i (cf. [14]). A strong conjecture is presented in [40] that in this case the distribution of $\frac{W_i}{S}$ tends to a uniform distribution on $[0, \frac{1-\rho_i}{1-\rho}]$ as $S \to \infty$.
It turns out that the distributional approximation as presented satisfies this result. To this end, consider the k^{th} moment of $\frac{W_{i,app}}{S}$, $k > 0$ as $S \to \infty$. It can easily be verified that $\lim_{S\to\infty} \mathbb{E}[\frac{W_{i,Boon}}{S}] = \frac{1-\rho_i}{2(1-\rho)}$, and hence,

$$\lim_{S\to\infty} \mathbb{E}\left[\left(\frac{W_{i,app}}{S}\right)^k\right] = \frac{1}{k+1}\left(\frac{1-\rho_i}{1-\rho}\right)^k \lim_{S\to\infty} \prod_{i=1}^{k} \frac{2S\delta + i\sigma^2}{2S\delta + \sigma^2} = \frac{1}{k+1}\left(\frac{1-\rho_i}{1-\rho}\right)^k. \tag{2.5}$$

This expression exactly coincides with the finite k^{th} moment of a uniformly distributed random variable Y on $[0, \frac{1-\rho_i}{1-\rho}]$. Thus, the k^{th} moment of $\frac{W_{i,app}}{S}$ converges to the k^{th} moment of Y when S tends to infinity, $k \geq 1$. Under certain conditions (which are met here), this moment-wise implies convergence in distribution (cf. [6], Theorem 4.5.5). Therefore, the distributional approximation becomes exact in the case of deterministic switch-over times that tend to infinity.

Large number of queues. Another limiting case is a symmetric system where the number of queues tends to infinity. A polling model with a number of queues that tends to infinity, also called a continuous polling model, may for example be applicable in systems where the server is patrolling a certain route, and customers arrive at random positions anywhere along the route. Due to symmetry, we have that $\lambda_i = \frac{\Lambda}{N}$ and $S_i = \frac{S}{N}$, $1 \leq i \leq N$. When taking the limit of $N \to \infty$, Λ and S remain unchanged, which means that λ_i and S_i will both tend to 0. For symmetric polling systems with Poisson arrival streams and deterministic switch-over times, a limiting result for the second moment of the waiting time is obtained in [11] as $N \to \infty$:

$$\begin{aligned}\lim_{N\to\infty} \mathbb{E}[W_i^2] =& \frac{2\rho\mathbb{E}[B^3]}{3(2-\rho)(1-\rho)\mathbb{E}[B]} + \frac{\rho^2(3-\rho)\mathbb{E}[B^2]^2}{3(2-\rho)(1-\rho)^2\mathbb{E}[B]^2} \\ &+ \frac{\rho\mathbb{E}[S]\mathbb{E}[B^2]}{(1-\rho)^2\mathbb{E}[B]} + \frac{\mathbb{E}[S]^2}{3(1-\rho)^2}.\end{aligned} \tag{2.6}$$

Using the fact that $\lim_{N\to\infty} \mathbb{E}[W_{i,Boon}] = \frac{1}{1-\rho}(\frac{\mathbb{E}[S]}{2} + \rho\mathbb{E}[B^{res}])$, one can show that the approximation error is bounded when the number of queues tends to infinity. For the sake of easy understanding and conciseness, we present the following

analysis with the additional assumption that the service times are deterministic. For general service times, the same analysis can be done and the same conclusion can be drawn, however the formulas become a lot more cumbersome. In case of deterministic service times, one can obtain a closed-form expression for the second moment of the distributional approximation as $N \to \infty$:

$$\lim_{N\to\infty} \mathbb{E}[W_{i,app}^2] = \frac{1}{3(1-\rho)^2} \frac{2\mathbb{E}[B] + \mathbb{E}[S]}{\mathbb{E}[B] + \mathbb{E}[S]} \left(\rho\mathbb{E}[B] + \mathbb{E}[S]\right)^2. \tag{2.7}$$

Then, the limiting percentual absolute relative error of the approximated second moment becomes

$$\begin{aligned} &100\% \times \lim_{N\to\infty} \frac{|\mathbb{E}[W_{i,app}^2] - \mathbb{E}[W_i^2]|}{\mathbb{E}[W_i^2]} \\ &= 100\% \times \frac{(1-\rho)\mathbb{E}[B]}{\mathbb{E}[B] + \mathbb{E}[S]} \left| \frac{\mathbb{E}[S]^2 - \rho\mathbb{E}[B]^2}{\rho(\rho+1)\mathbb{E}[B]^2 + 3\rho\mathbb{E}[B]\mathbb{E}[S] + \mathbb{E}[S]^2} \right|. \end{aligned} \tag{2.8}$$

Taking the derivative of this expression with respect to ρ, and subsequently equating it to 0 shows that a (local) maximum is obtained in $\rho = \frac{\mathbb{E}[S]}{\mathbb{E}[B]}$ with a value of $100\% \times \frac{(\mathbb{E}[B]-\mathbb{E}[S])^2}{(\mathbb{E}[B]+\mathbb{E}[S])(\mathbb{E}[B]+5\mathbb{E}[S])}$. This value cannot become larger than 100% and for $\frac{\mathbb{E}[S]}{\mathbb{E}[B]} < 1$ becomes smaller rapidly as the durations of the switch-over times increase. To find the maximum of this expression in the domain $\rho = [0, 1]$, apart from $\rho = \frac{\mathbb{E}[S]}{\mathbb{E}[B]}$ when $\frac{\mathbb{E}[S]}{\mathbb{E}[B]} < 1$, the boundary ρ-values need to be regarded, $\rho = 0$ and $\rho = 1$. In practice, $\frac{\mathbb{E}[S]}{\mathbb{E}[B]}$ is mostly larger than 1, which implies there is no maximum in the interval $[0, 1]$ other than in the boundaries. As stated earlier in this section the approximation is exact for polling systems in HT. This leaves the value of $\rho = 0$. It can be verified that

$$\lim_{\rho\downarrow 0} \lim_{N\to\infty} \frac{|\mathbb{E}[W_{i,app}^2] - \mathbb{E}[W_i^2]|}{\mathbb{E}[W_i^2]} = 100\% \times \frac{\mathbb{E}[B]}{\mathbb{E}[B] + \mathbb{E}[S]}, \tag{2.9}$$

which also cannot become larger than 100%, and becomes smaller rapidly as the durations of the switch-over times increase.

From these results, one can conclude that the absolute value of the relative error never grows beyond 100%, and becomes smaller rapidly as ρ or $\mathbb{E}[S]$ becomes larger. Summarizing, we conclude that as $N \to \infty$, the approximation does not become exact, however the standard deviation of the exact waiting time distribution is nevertheless well approximated in a variety of polling systems.

2.5 Simulation study

In this section, we evaluate the accuracy of the approximation of the waiting time distribution as presented in Section 2.3. First, we regard a rather arbitrary polling

system and see how well the approximated and exact density functions coincide. The "exact" density function is determined by means of simulation. Then, we study the accuracy of the approximation in a wide range of parameter combinations by applying the approximation to a test bed containing 10368 polling systems and summarizing the errors of standard deviation and percentile approximations. Again, the "exact" standard deviations and percentiles are determined by means of simulation. Also in case of Poisson arrivals, where numerical methods exist to determine the exact distribution, we opt for simulation, since the determination of the exact values using numerical methods can be very cumbersome. All simulation results presented in this section are an average taken from a variable number of simulation runs with a length of at least 1,000,000 time units, such that the width of the confidence interval of the average is less than 1% of the value of the actual average.

2.5.1 Accuracy of the approximated density function

We consider a symmetric polling system with five queues. The load ρ equals 0.7, the SCV of the interarrival times at each queue are 0.25 each. All the service times and switch-over times are exponentially distributed with mean 1. Since there is no exact closed-form expression available for the waiting time distribution in this case, we compare the density function of the approximated distribution with the simulated density function for an arbitrary queue. To obtain the latter, a kernel estimation was made based on a huge set of simulated waiting time realizations. Both the approximated density function and the simulated density function are depicted in Figure 2.1. For the interarrival times, a gamma distribution was used with shape parameter 4 and inverse scale parameter 16.

Figure 2.1 shows that the shape of the exact waiting time distribution closely resembles the approximation. In other words, the shape of the waiting time distribution in heavy traffic closely matches the shape of the waiting time distribution for a general load. Hence, the figure suggests that the approximation is useful for approximating the density function, tail probabilities and other characteristics of the distribution. The next subsection will show that the approximation works well not only in this case, but also in a variety of other polling systems.

2.5.2 Accuracy of approximated percentiles and standard deviation

In this subsection we assess the accuracy of the approximation by evaluating errors in the approximation of the standard deviation and several percentiles. We regard the standard deviation and several percentiles of the approximated distribution and the exact distribution of the waiting time of the first queue in a large number of polling systems with exhaustive, cyclic service. The standard deviation and

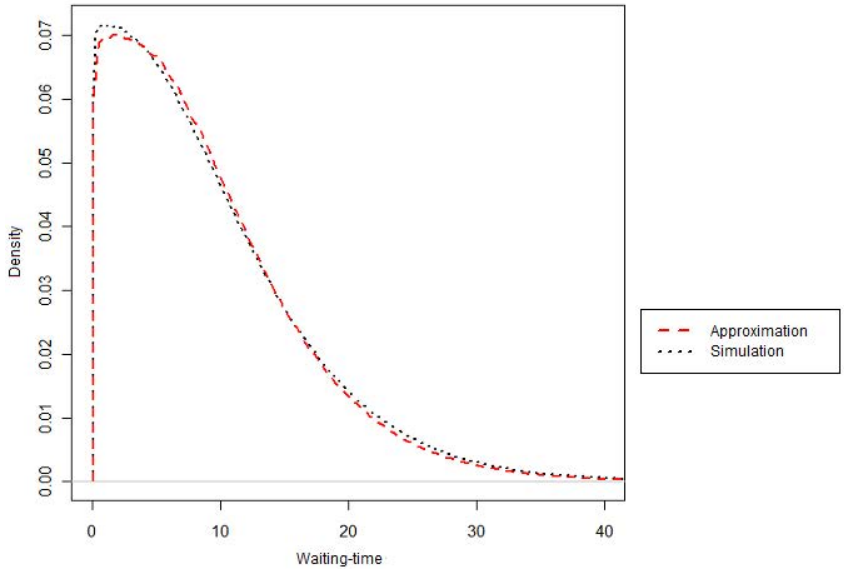


Figure 2.1: Approximated and simulated density function of the waiting time of an arbitrary queue in the example in Subsection 2.5.1.

percentiles of the exact distribution are determined by means of simulation. We first give a general impression of the accuracy, after which we try to explore what the impact of each of the parameters is on the accuracy.

The parameter values contained in the test bed can be found in Table 2.1. There is no difference between the queues within a particular polling system in terms of the service time distributions and the switch-over time distributions. All parameters are explained above, except for the last one displayed in the table. $Q_2, \ldots, Q_N$ take on the same amount of load each. The parameter p_1 denotes what amount of load is taken on by Q_1 relative to the other queues. For example, if the first queue takes half, twice or five times as much load as any other, p_1 becomes 0.5, 2 or 5 respectively. In case the system is symmetric, $p_1 = 1$. Since simulation needs complete distributions as input rather than just their first two moments, two-moment fits were deployed as described in Appendix B. For each polling system, the approximation error of the standard deviation and the approximation error of the 40^{th}, 50^{th}, 60^{th}, 70^{th}, 80^{th}, 90^{th} and 95^{th}-percentiles are calculated. The errors are measured in a percentual absolute relative way, i.e.,

$$\Delta\% = \frac{|a - s|}{s} \times 100\%, \tag{2.10}$$

where a denotes the approximated value by means of the distributional approximation as presented in Section 2.3. and s denotes the exact value, determined by means of simulation.

Notation	Parameter	Considered parameter values
N	Number of queues	$\{5, 10, 20\}$
ρ	Load	$\{0.5, 0.6, 0.7, 0.8, 0.9, 0.95\}$
$c^2_{A_i}$	SCV interarrival times	$\{0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75, 2\}$
$\mathbb{E}[B_i]$	Mean service times	$\{1\}$
$c^2_{B_i}$	SCV service times	$\{0, 1, 4\}$
$\mathbb{E}[S_i]$	Mean switch-over times	$\{0.2, 1, 10\}$
$c^2_{S_i}$	SCV switch-over times	$\{0, 1\}$
p_1	Measure of asymmetry	$\{0.5, 1, 2, 5\}$

Table 2.1: Parameter values of the test bed used in subsection 2.5.2.

p_1	Bins				
	0-5%	5-10%	10-15%	15-20%	20%+
0.5	69.98%	19.68%	5.02%	2.28%	3.05%
1	69.87%	19.33%	5.56%	2.39%	2.85%
2	67.67%	19.98%	6.17%	3.20%	2.97%
5	58.72%	22.07%	8.60%	4.86%	5.75%

Table 2.2: Mean standard deviation error of the approximation applied to the test bed, categorized in bins of 5%.

Table 2.2 and Table 2.3 show the errors made in approximating the standard deviation and in approximating the percentiles respectively in bins of 5%. This table suggests that the approximation performs best in case of $p_1 \leq 1$. Regardless, it is shown that the majority of the standard deviation approximations, and even more so the percentile approximation have errors lower than 5%.

In order to judge the impact of the system parameters on the performance of the approximation, the mean absolute relative errors of the standard deviation approximation and the percentile approximation are given in Table 2.4 and in Table 2.5 respectively, vertically categorized in the different rates of asymmetry, horizontally categorized in each of the relevant system parameters. Table 2.4(a) and Table 2.5(a) show that the distributional approximation becomes better when N increases. The same behavior of the approximation error in N is present in the approximations of [2, 25]. Table 2.4(b) together with Table 2.5(b) shows a surprising effect of the SCV of the service times on the performance of the approximation. While in case of $c^2_{B_i} = 0$ the standard deviation approximation error seems to grow with p_1, the same effect does not seem to happen when $c^2_{B_i} = 4$. Also, if $c^2_{B_i} = 1$ or $c^2_{B_i} = 4$, the approximation error of the standard deviation and the approximation error of the percentiles do not seem to react in the same way to changes in p_1. Tables 2.4(c) and 2.5(c) suggest that the distributional approximation becomes better when the variance of the switch-over times increases. Table 2.4(d) and Table 2.5(d) suggest that approximations become better as ρ approaches 1, i.e., as the system

p_1	Bins				
	0-5%	5-10%	10-15%	15-20%	20%+
0.5	82.25%	10.77%	3.26%	1.07%	2.65%
1	82.14%	10.98%	3.44%	1.03%	2.41%
2	82.25%	11.17%	3.38%	1.04%	2.16%
5	77.63%	13.32%	4.76%	2.00%	2.29%

Table 2.3: Mean percentile error of the approximation applied to the test bed, categorized in bins of 5%.

(a)

p_1	N		
	5	10	20
0.5	5.97	4.02	3.17
1	5.99	4.06	3.25
2	6.41	4.31	3.29
5	9.08	5.76	3.77

(b)

p_1	$c^2_{B_i}$		
	0	1	4
0.5	3.91	3.25	5.99
1	4.15	3.33	5.81
2	4.80	3.64	5.57
5	6.97	5.49	6.15

(c)

p_1	$c^2_{S_i}$	
	0	1
0.5	4.49	4.28
1	4.59	4.27
2	8.80	4.46
5	6.73	5.68

(d)

p_1	ρ					
	0.5	0.6	0.7	0.8	0.9	0.95
0.5	8.22	6.66	5.11	3.48	1.81	1.06
1	8.37	6.77	5.12	3.49	1.79	1.06
2	8.82	7.14	5.42	3.65	1.87	1.12
5	11.02	9.30	7.36	5.19	2.75	1.60

(e)

p_1	$\mathbb{E}[S_i]$		
	0.2	1	10
0.5	7.35	3.37	2.44
1	7.34	3.53	2.33
2	7.44	4.07	2.30
5	8.76	6.79	3.06

(f)

p_1	$c^2_{A_i}$							
	0.25	0.5	0.75	1	1.25	1.50	1.75	2
0.5	4.11	4.13	3.95	4.04	4.34	4.59	4.82	5.10
1	4.10	4.14	4.00	4.06	4.40	4.67	4.91	5.18
2	4.18	4.42	4.28	4.33	4.67	4.93	5.18	5.37
5	4.77	5.20	5.28	6.00	6.48	6.94	7.30	7.65

Table 2.4: Mean standard deviation error categorized by the value of p_1 vertically and the number of queues (a), the SCV of the service times (b), the SCV of the switch-over times (c), the total load (d), the mean switch-over time (e), and the SCV of the interarrival times (f) horizontally.

gets closer to HT. This is very plausible, since by construction the approximation is exact in HT as explained in Section 2.4. According to Table 2.4(e) and Table 2.5(e) approximations seem to become better as the switch-over times become bigger. This is in line with the result found in Section 2.4 that the approximations

(a)

p_1	N		
	5	10	20
0.5	5.77	3.27	1.99
1	5.45	3.23	2.05
2	5.10	3.18	2.02
5	5.87	3.50	2.15

(b)

p_1	$c^2_{B_i}$		
	0	1	4
0.5	2.15	2.40	6.48
1	2.28	2.27	6.19
2	2.44	2.31	5.56
5	3.35	3.03	5.14

(c)

p_1	$c^2_{S_i}$	
	0	1
0.5	3.87	3.49
1	3.77	3.39
2	3.58	3.29
5	4.14	3.54

(d)

p_1	ρ					
	0.5	0.6	0.7	0.8	0.9	0.95
0.5	5.99	5.79	4.46	2.89	1.37	0.79
1	5.86	5.65	4.28	2.75	1.30	0.79
2	5.67	5.44	4.05	2.56	1.22	0.75
5	6.14	5.86	4.60	3.10	1.55	0.91

(e)

p_1	$\mathbb{E}[S_i]$		
	0.2	1	10
0.5	7.50	2.19	1.34
1	6.98	2.31	1.45
2	6.43	2.42	1.46
5	6.14	3.64	1.74

(f)

p_1	$c^2_{A_i}$							
	0.25	0.5	0.75	1	1.25	1.5	1.75	2
0.5	4.20	3.59	3.25	3.31	3.48	3.65	3.86	4.09
1	3.97	3.41	3.14	3.29	3.44	3.61	3.78	3.97
2	3.53	3.13	3.04	3.30	3.45	3.56	3.67	3.80
5	3.15	3.26	3.52	3.80	4.00	4.18	4.32	4.50

Table 2.5: Mean percentile error categorized by the value of p_1 vertically and the number of queues (a), the SCV of the service times (b), the SCV of the switch-over times (c), the total load (d), the mean switch-over time (e), and the SCV of the interarrival times (f) horizontally.

becomes exact as the total switch-over time tends to infinity. Also, using (2.4) one can show that the moments of the distributional approximation become less dependent of σ^2 and δ as switch-over times become smaller, which gives support to the plausibility of the approximation becoming less reliable when the switch-over times become relatively small. Finally, both Table 2.4(f) and Table 2.5(f) show that the approximations' quality is dependent on the SCV of the interarrival times, but again an interaction effect with the value of p_1 is observed.
Table 2.6 shows the mean absolute relative error categorized per tested percentile. Generally, the 80% percentiles seem to be approximated best.

From the test bed results we can conclude that the approximation performs well over a wide range of parameter combinations. In case of extremely variable service times, low load and negligibly small switch-over times, the relative error

p_1	Percentile						
	40^{th}	50^{th}	60^{th}	70^{th}	80^{th}	90^{th}	95^{th}
0.5	5.90	5.23	4.15	2.88	1.72	2.29	3.57
1	5.73	5.00	3.93	2.71	1.67	2.37	3.63
2	5.42	4.59	3.60	2.50	1.67	2.49	3.78
5	6.29	4.87	3.40	2.11	1.88	3.42	4.92

Table 2.6: Mean absolute relative errors categorized in the several percentiles.

becomes worse. The worst-case scenarios found in the testbed in terms of absolute relative error are approximations of the 50^{th} percentile in systems with $N = 5$, $\rho = 0.5$, $c^2_{B_i} = 4$ and $\mathbb{E}[S_i] = 0.2$ having errors with an order of magnitude of 100%. However, in practice these characteristics are uncommon. For example, in production systems settings like $c^2_{B_i} = 4$ are hardly found due to the just-in-time philosophy, which dictates to reduce variability in, e.g., service times in order to reduce in-process inventory. Also, these systems are typically utilized beyond $\rho = 0.5$ to increase productivity, and switch-over periods are commonly longer than service periods. Moreover, in case of a low load and small switch-over times, although the relative error of the percentile approximations can be high, the absolute errors may still be rather small when compared to the order of longitude of service time durations. Therefore, the sojourn time distribution is already much better approximated in these situations.

2.6 Further research

The present study gives birth to a variety of directions for further research. Firstly, the distributional approximation for cyclic systems with exhaustive service may be generalized to models with branching-type service policies [28], non-cyclic periodic server routing [26] and other model variations. Secondly, the simple closed-form expression may act as a basis for design decisions within polling systems. Finally, one could attempt to improve the approximation by deriving an interpolation approximation for higher moments of the waiting time and, subsequently, fit a phase-type distribution. However, this would impel one to considerably extend the analysis of [2], while potentially losing the simple form of the current distributional approximation.

Chapter 3

Optimization

In this chapter we aim to fulfill the second research objective as formulated in Section 1.3. This research objective is motivated by flexible production facilities with batch services where the batch-processing times are independent of the batch size. In such systems, a key problem is to determine the optimal batch size. Motivated by this, we consider an N-queue batch-service polling system consisting of an *inner* part and an *outer* part. Type-i customers arrive at the outer system according to a renewal process and accumulate into a type-i batch. As soon as D_i customers have arrived, the batch is forwarded to the inner system where the batch is processed, and where the batch-service requirement is independent of the batch size D_i. For this type of models, we study the problem of determining the combination of batch sizes $\vec{D}^{opt} = (D_1^{opt}, \ldots, D_N^{opt})$ that minimizes a weighted sum of the mean waiting times in the outer and the inner system. A balance in the trade-off between the waiting times in the outer system V_i and the waiting time in the inner system W_i needs to be found: the larger the batch size, the lower the workload of the system and hence the average "inner waiting time", but the higher the average "outer waiting time" needed to accumulate D_i customers. This model does not allow for an exact analysis. Therefore, we present a numerical approach to this problem, and propose a closed-form approximation for the optimal combination of batch sizes (which is the main result of this chapter). Note that these two methods complement each other: the numerical approach works better in systems with a small number of queues, but the closed-form approximation requires significantly less computation time and performs almost equally well in systems with a large number of queues. As a by-product, we observe near-insensitivity properties of $\vec{D}^{opt}$, e.g., to higher moments of the interarrival and switch-over time distributions. Extensive experimentation shows that the obtained approximation is highly accurate.

3.1 Introduction

This chapter is motivated by practical flexible manufacturing systems where a production facility makes a large number of products. The nature of the production technology is such that products are processed in batches. Moreover, the time required to process such a batch depends only weakly on the size of the batch, because processing itself affects the entire batch at once. Common examples are an oven that heats multiple items at once, a paint bath which may paint several items at a time, the production of pharmaceuticals or the blending of gasoline (cf. [42]). Batch service also does not only have its applications in production facilities, but it is also applicable in the field of computer-communication systems, such as video tex systems and Time Division Multiple Access (TDMA) systems [20]. Management of production facilities is often challenged with a rather complex trade-off concerning the determination of batch sizes. Having smaller batch sizes implies that less product inventory is needed to have a batch formed and that arriving jobs requiring server attention spend a smaller amount of time in storage before they become part of a batch. Therefore, taking small batch sizes is often desirable on the one hand. However, having larger batch sizes implies that there are less batches that require processing, which reduces the workload of the production system. This reduction of workload thus translates into less waiting time for the batches as a whole. Therefore, taking larger batch sizes is desirable on the other hand.

In accordance with the applicability of batch services, there is a body of literature available on non-polling queueing systems with batch service, both from an evaluation perspective [16, 17], and a design perspective [7, 35, 42]. Over the past few decades, the analysis and optimization of polling systems also have been subject of intensive research efforts, see [19, 30, 32] for overviews. However, remarkably little attention has been paid to polling models in combination with batch service. As an exception, Boxma et al. [3] study the performance of a class of polling models with batch service when batches are always served integrally, and Vlasiou and Yechiali [33] study the case where the service of underlying jobs may be abandoned and pushed to the batch of the next visiting period when the current visit time is up. Optimal dynamic routing policies for these systems are studied, when the server has complete freedom of visits in [20] and when routing must be done in subsequent Hamiltonian tours [34]. In [27], the question is studied whether upon arrival the server should poll a station or idle until more customers have arrived at the station when the server assumes a cyclic routing mechanism. These studies mostly assume that the server can take in any number of customers for service at a time and that customers arrive according to Poisson arrival processes. For polling models with renewal arrivals and normal service, hardly any exact results are known, except for several asymptotic regimes, including limiting cases where the system is heavily loaded [26], the switch-over times are large [40], or where the number of queues grows to infinity. Faced by this, approximations have been

developed for the mean waiting time [2], recently extended to the complete waiting time distribution [8].

In the batch service models addressed above, there is no upper bound to the number of customers served during one visit of the server to a queue. We consider a model that is fundamentally different. We assume that at each queue, the server can only process batches of a queue-dependent, fixed number of customers at a time, which is more realistic in many production systems. More specifically, we study an N-queue batch-service polling system consisting of an inner part and an outer part. Type-i customers arrive at the outer system according to a renewal process and accumulate into a type-i batch. Thus, customers have to wait in the outer part until they become part of a fully accumulated batch. When D_i customers have accumulated in the outer system, the batch is forwarded to the inner system, which can be seen as a regular polling system. In the inner system the batch, and thus the underlying customers wait until the batch as a whole, is processed by the server. It is assumed that the batch service requirement is independent of the batch size D_i. For this model, we study the problem of determining the combination of batch sizes $\vec{D}^{opt} = (D_1^{opt}, \ldots, D_N^{opt})$ that minimizes a weighted sum of the total mean waiting times in the outer and the inner system. Note that by addressing this problem, one is confronted with a challenging trade-off. When batch sizes are increased, the workload of the system reduces, leading to the inner waiting time being shorter. However, it will take longer before a batch of size D_i is fully accumulated and sent to the inner system, which increases the outer waiting time.
In the absence of exact analysis, we present a numerical approach for this problem. Moreover, as the main result of this chapter we propose a closed-form approximation for the optimal combination of batch sizes that allows for back-of-the-envelope calculations, and which gives valuable insights in the dependence of the system performance with respect to the batch size. Extensive experimentation shows that the approximation is highly accurate. The methods considered are complementary. Assuming a small number of queues, the numerical method performs better than the closed-form approximation. However, the numerical method does not scale well in the number of queues, assuming larger numbers of queues will eventually result in the numerical method requiring infeasibly long computation times. Moreover, the closed-form approximation will perform increasingly well in that case, favouring the closed-form approximation whenever there is a large number of queues involved. As a by-product of the closed-form approximation, we observe near-insensitivity properties of the batch sizes. Firstly, the closed-form approximation suggests that the optimal batch sizes are insensitive to higher moments of the interarrival and switch-over time distributions. Secondly, it is suggested that the ratio of the optimal batch sizes of two types or queues is independent of any characteristic of other queues.

The structure of this chapter is as follows. In Section 3.2, the model is in-

troduced in detail and required notation is given. In Section 3.3 the problem is described in detail. Section 3.4 provides evaluation measures, after which a numerical approach to obtain near-optimal batch sizes is discussed and a closed-form approximation of the optimal batch sizes proposed in Section 3.5. In Section 3.6 validation of these approaches is performed by means of simulation, while Section 3.7 takes a look at the near-insensitivity properties suggested by the obtained closed-form approximation by means of numerical studies. Finally, Section 3.8 offers suggestions for further research.

3.2 Model description and notation

We consider the queueing system as depicted in Figure 3.1. This queueing system can be divided into an *outer part* and an *inner part*. The outer part has type-i customers entering the system according to a renewal arrival process, where interarrival times are denoted by the random variable EA_i. Type-i customers (denoted by yellow rectangles in Figure 3.1) are said to arrive at rate $\lambda_i = \frac{1}{\mathbb{E}[EA_i]}$ and have to wait in the outer system until D_i type-i customers are present in the outer part, where D_i is by nature a strictly positive integer. Whenever there are D_i customers present, these customers immediately form a type-i batch (denoted by teal rectangles in Figure 3.1) and this batch is immediately forwarded to the inner system as a type-i super-customer. The inner part of the queueing system is a typical polling system consisting of $N > 1$ queues, $Q_1, \ldots, Q_N$, each with an infinite-sized buffer at which batches arrive and wait until they are taken into service by a single server that is attending all queues. Throughout, a queue index i is understood as $((i-1) \mod N) + 1$, e.g., Q_{N+1} actually refers to Q_1. Interarrival times of type-i batches at Q_i are denoted by the random variable A_i and the type-i batches themselves are said to arrive at rate $\nu_i = \frac{1}{\mathbb{E}[A_i]}$. Note that $A_i = \sum_{i=1}^{D_i} EA_i$, which gives rise to the useful relations $\nu_i = \frac{\lambda_i}{D_i}$ and $\mathrm{Var}[A_i] = D_i \mathrm{Var}[EA_i]$.

In the polling system, batches are served based on a first-in-first-out queueing discipline. It should be noted that the length of $Q_1, \ldots, Q_N$, and in case the A_i are exponential also the mean waiting times are independent of the queueing discipline, provided that the queueing discipline is on its turn independent of the service times. The service time of a type-i customer is denoted by the random variable B_i, which is completely independent of D_i. That is, the number of customers present in the batch, or simply the batch size has no impact on the service requirement of the batch itself. Whenever a batch is being served, it is assumed that the underlying customers are all served simultaneously, such that there is no underlying customer in the batch that has its service requirement completed before the batch as a whole is served. The cost of having a type-i customer waiting anywhere in the system for one unit of time is denoted by c_i.
The server attends the queues according to an *exhaustive* service discipline, i.e., when attending Q_i, the server will commence moving to another queue if and only

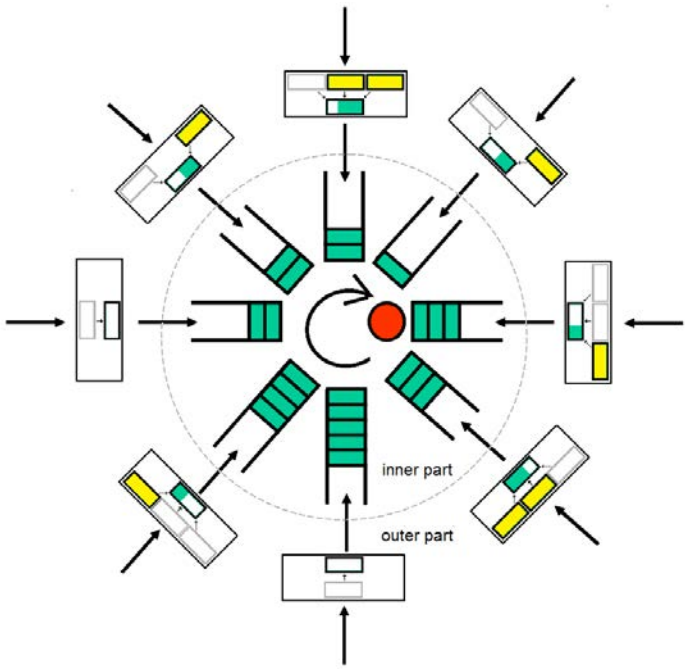


Figure 3.1: Graphical representation of the model under consideration.

if Q_i is completely empty. The server switches through the queues in a cyclic manner, which means that the server will switch to Q_{i+1} after the server has completed a service period at Q_i. In order to succesfully perform this switch, the server typically needs a switch-over time, of which the duration is denoted by a random variable S_i.
We define a cycle at Q_i as the time between two successive departures of the server at Q_i. Then, $S = \sum_{i=1}^{N} S_i$ denotes the total switch-over time in a cycle. Since batches arrive at Q_i at rate $\nu_i = \frac{1}{\mathbb{E}[A_i]}$, the load offered to Q_i is denoted by $\rho_i = \nu_i \mathbb{E}[B_i]$. The total load offered to the inner polling system is then denoted by $\rho = \sum_{i=1}^{N} \rho_i$. A necessary and sufficient condition for the stability of the inner system reads $\rho < 1$ [9]. Since no work is created or destroyed in the outer part of the system, this is also a necessary and sufficient condition for the stability of the described system as a whole.

Throughout, it may be convenient to scale the polling system such that a certain load is achieved. This scaling is done by keeping the service time distributions fixed and varying the rates of the renewal arrival processes of the customers. In particular, it proves convenient to represent with $\hat{x}$ the value of each variable x that is a function of ρ evaluated at $\rho = 1$. In that case $\hat{A}_i$ denotes the interarrival time of type-i batches evaluated at $\rho = 1$. Then, scaling to a load $\rho < 1$ is done by taking the random variable $A_i := \frac{\hat{A}_i}{\rho}$ (or in terms of customers, $EA_i := \frac{\widehat{EA}_i}{\rho}$).

Finally, we introduce some additional notation. The residual length of a non-negative random variable X with positive finite mean is denoted by X^{res}, with $\mathbb{E}[X^{res}] = \frac{\mathbb{E}[X^2]}{2\mathbb{E}[X]}$. The squared coefficient of variation (SCV) of a random variable X, $\frac{\text{Var}[X]}{\mathbb{E}[X]^2}$ is denoted by c_X^2. We define $\sigma^2 = \sum_{i=1}^{N} \hat{\nu}_i(\text{Var}[B_i] + c_{A_i}^2 \mathbb{E}[B_i]^2)$ and $\delta = \sum_{j=1}^{N} \sum_{k=j+1}^{N} \hat{\rho}_j \hat{\rho}_k$. Note that in case batches arrive at the polling system according to a Poisson process, the former can be simplified to $\sigma^2 = \mathbb{E}[B^2]/\mathbb{E}[B]$. We refer to the system as symmetric, whenever the interarrival distributions of all the customer types are equal, D_i is independent of its index and the queues in the polling system have identical service time distributions and switch-over time distributions. Of course, a system is asymmetric when any of these conditions is violated. Indicator functions are used in the form of $\mathbb{1}_{\{A\}}$, which evaluate to one if condition A holds, zero otherwise.

The main result of this chapter is the following closed-form approximation of the batch size vector $\vec{D}^{opt}$ minimizing the weighted sum of the waiting times of the customers:

$$E_{app}^{opt} = \sum_{i=1}^{N} \frac{\lambda_i \mathbb{E}[B_i]}{d_i} + \sqrt{2 \left(\sum_{i=1}^{N} \frac{c_i d_i}{\lambda_i} \right)^{-1} \left(\sum_{i=1}^{N} c_i \omega_{i,app} \right) \left(\sum_{i=1}^{N} \frac{\lambda_i \mathbb{E}[B_i]}{d_i} \right)}, \quad (3.1)$$

$$\vec{d} = (d_1, \ldots, d_N) = \left(1, \frac{\lambda_2}{\lambda_1} \sqrt{\frac{c_1 \mathbb{E}[B_2]}{c_2 \mathbb{E}[B_1]}}, \ldots, \frac{\lambda_N}{\lambda_1} \sqrt{\frac{c_1 \mathbb{E}[B_N]}{c_N \mathbb{E}[B_1]}} \right), \quad (3.2)$$

and

$$\vec{D}_{app}^{opt} = (D_{1,app}^{opt}, D_{2,app}^{opt}, \ldots, D_{N,app}^{opt}) = (d_1 E_{app}^{opt}, d_2 E_{app}^{opt}, \ldots, d_N E_{app}^{opt}), \quad (3.3)$$

where

$$\omega_{i,app} = \frac{1 - \hat{\rho}_i}{2} \left(\frac{\sum_{j=1}^{N} \hat{\nu}_j \text{Var}[B_j]}{2\delta} + \mathbb{E}[S] \right) \quad \text{for } i = 1, \ldots, N \quad (3.4)$$

and $D_{i,app}^{opt}$ represents an approximation of the optimal type-i batch size. Since these values may be fractional, they are rounded off to the nearest positive integer if that results in a stable system, otherwise they are rounded to the nearest strictly larger integer.

3.3 Problem description

The optimization problem studied in this chapter concerns the optimal choice of batch sizes. The waiting time of a type-i customer can be decomposed into two parts:

- V_i, the time a type-i customer has to wait upon arrival until its type-i batch is full.
- W_i, the time a type-i customer spends in Q_i as part of a batch waiting for attention of the server.

This decomposition illustrates a trade-off in the choice of batch sizes. When one decides to increase the size of a type-i batch D_i, V_i will obviously increase since customers have to wait longer on average until their batch is full. On the other hand, when batch sizes are increased, less batches need to be attended to by the server of the polling system. Consequently, the load ρ of the polling system will decrease, and W_i will become smaller. This motivates us to study the question how batch sizes should be chosen, such that the (weighted) type-averaged total waiting time of a customer is as small as possible.

Let us define a cost function as a function of $\vec{D} = (D_1, \ldots, D_N)$,

$$C(\vec{D}) = C(D_1, \ldots, D_N) = \sum_{i=1}^{N} c_i(\mathbb{E}[V_i] + \mathbb{E}[W_i]). \tag{3.5}$$

Then, the problem at hand encompasses the search of

$$\arg\min_{\vec{D}} C(D_1, \ldots, D_N) \tag{3.6}$$

for a given weight vector $\vec{c} = (c_1, c_2, \ldots, c_n)$, under the constraint that all the elements of $\vec{D}$ remain positive and integer-valued. Unfortunately, there is no exact expression available for $\mathbb{E}[W_i]$. For $\mathbb{E}[V_i]$ however, a closed-form exact expression in terms of $\vec{D}$ can be obtained through the conditional law of the unconscious statistician. Let E_{ij} be the event that an arriving type-i customer is the j-th taking place in the type-i batch currently being filled in the outer system. Then,

$$\mathbb{E}[V_i] = \sum_{j=1}^{D_i} \mathbb{P}[E_{ij}]\mathbb{E}[V_i|E_{ij}]. \tag{3.7}$$

It is evidently seen that $\mathbb{P}[E_{ij}] = \frac{1}{D_i}$ and $\mathbb{E}[V_i|E_{ij}] = \frac{D_i - j}{\lambda_i}$. Combining this, we have that $\mathbb{E}[V_i] = \frac{D_i - 1}{2\lambda_i}$. Therefore, the cost function equals

$$C(D_1, \ldots, D_N) = \sum_{i=1}^{N} c_i \left(\frac{D_i - 1}{2\lambda_i} + \mathbb{E}[W_i] \right). \tag{3.8}$$

Figure 3.2 gives the general form and a visual clue about how $\mathbb{E}[V_i] + \mathbb{E}[W_i]$ changes in D_i, $1 \leq i \leq N$. The grey lines show $\mathbb{E}[V_i]$ and $\mathbb{E}[W_i]$. These two components together determine the total waiting time of a type-i customer and clearly show the explained trade-off. The component $\mathbb{E}[W_i]$ in this figure was approximated using the result of [2], which is also presented and used in Section 3.4. The present study concerns the question how the batch size vector $\vec{D} = (D_1, D_2, \ldots, D_N)$ should be chosen such that $C(D_1, \ldots, D_N)$ is minimized.

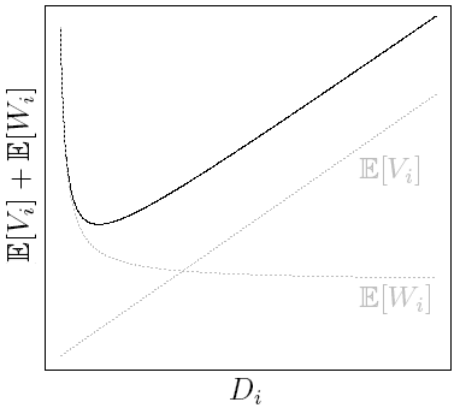


Figure 3.2: Illustration of the trade-off between small and large batch sizes in the total waiting time.

3.4 Evaluational tools

As concluded earlier, the total waiting time of a type-i customer can be decomposed in two parts, denoted by $\mathbb{E}[V_i]$ and $\mathbb{E}[W_i]$. An exact expression for $\mathbb{E}[V_i]$ is readily available and derived in the previous section. However, an exact expression for $\mathbb{E}[W_i]$, the waiting time of a customer in a polling system with renewal arrivals, does not exist. Exact evaluation of (3.8) is therefore not possible. A way to get around this is by using the well-performing approximation for $\mathbb{E}[W_i]$ found by Boon et al. [2], which was given and used earlier in 2.2 in the previous chapter. This approximation is shown to be exact in a variety of limiting cases and a very good approximation in general. For an explanation of the derivation of this approximation $\mathbb{E}[W_{i,Boon}]$, see Appendix A.
Using this approximation, one may use the following approximative cost function for evaluational purposes:

$$C_{Boon}(D_1, \ldots, D_N) = \sum_{i=1}^{N} c_i \left(\frac{D_i - 1}{2\lambda_i} + \mathbb{E}[W_{i,Boon}] \right). \tag{3.9}$$

Since $\mathbb{E}[W_i]$ is the only term in this function that is approximated, $C_{Boon}(D_1, \ldots, D_N)$ adopts the accuracy properties of $\mathbb{E}[W_{i,Boon}]$. That is, $C_{Boon}(D_1, \ldots, D_N)$ is exact in a variety of limiting cases, and a very accurate approximation of the cost in general. This approximative cost function can now be used for optimization purposes, as is done in Subsection 3.5.1.

3.5 Optimization

This section discusses a numerical approach to the optimization problem posed in Section 3.3 and moreover presents the main result of this chapter, a closed-form

approximation of the batch sizes minimizing $C_(D_1, \ldots, D_N)$. These two methods are complementary: for a small N, the numerical approach works significantly better than the closed-form approximation. However, the numerical approach does not scale well in N, such that it requires long computation times for a large N. The closed-form approximation however only requires a negligible amount of computation time and performs almost equally well in case the number of queues is large.

3.5.1 Numerical approach

When one wants to avoid a simulation-like method for the determination of optimal batch sizes, an expression for $\mathbb{E}[W_i]$ is needed. Taking $\mathbb{E}[W_{i,Boon}]$ as an approximation for $\mathbb{E}[W_i]$ results in the approximative cost function $C_{Boon}(D_1, \ldots, D_N)$ as given in Section 3.4. The vector $\vec{D}^{opt}_{Boon}$ minimizing this approximative cost function could then act as an approximation of the optimal batch size vector. Closed-form expressions for the batch sizes minimizing $C_{Boon}(D_1, \ldots, D_N)$ are exceptionally hard to derive and untractably complicated. However, numerical methods can be used to obtain this minimum.
Note that optimization problem (3.6) is a constrained non-linear problem, since the polling system should be stable at all times and all batch sizes should be positive. However, this can be converted to an unconstrained optimization problem by adding a term to the cost function which is uncomparably large whenever the constraint is violated, zero otherwise.
This procedure may lead to fractional optimal batch sizes. However, by nature, batch sizes must be integer. Each of the fractional values is therefore rounded to the nearest positive integer. If this results in an unstable system, each of the values is rounded upwards to the nearest larger integer.

Validation of this numerical approach is done in Section 3.6. The numerical approximation is shown to perform well over a wide range of system parameters. However, numerical methods such as these may be cumbersome to deploy and only give limited insight into how the waiting time and the optimal value for $\vec{D}$ react to changes of the system's parameters. The numerical approach performs well and does not require much computation time for low values of N. However, it does not scale well with the number of queues, since more queues implies more batch sizes to optimize. On its turn, more batch sizes translates in additional dimensions. Due to the curse of dimensionality the computation time needed may grow infeasibly long.

3.5.2 Closed-form approximation

Because of the limitations of a numerical approach, our present goal is to derive a closed-form approximation of the optimal batch sizes. To this end, we aim to derive a simply computable closed-form approximation of the optimal batch sizes

based on an approximative cost function. The cost function $C_{Boon}(D_1, \ldots, D_N)$ as derived in Section 3.4 defies the derivation of a closed-form solution. Therefore, we first derive an approximation for $\mathbb{E}[W_i]$ that is less complex than $\mathbb{E}[W_{i,Boon}]$, resulting in the approximative cost function $C_{app}(D_1, \ldots, D_N)$. Then, closed-form approximations for the optimal batch sizes will be derived. For the sake of clarity of presentation, we initially approach the problem in its one-dimensional form, more specifically under the restriction that $D_1 = D_2 = \ldots = D_N$. Note that this is a restriction that is often a natural one in practice. For example, think of an oven with a fixed number of baking slots. After we have solved the one-dimensional case, we consider the multi-dimensional problem. That is, we obtain the vector $\vec{D}_{app}^{opt}$ minimizing $C_{app}(D_1, \ldots, D_N)$ and use it as an approximation of the optimal batch sizes minimizing $C(D_1, \ldots, D_N)$.

Preliminaries

The function $C_{Boon}(D_1, \ldots, D_N)$ is too complex to gain closed-form approximations of the optimal batch sizes, which is the main aim of this chapter. We therefore introduce another, even simpler approximation of the mean batch waiting time of the form

$$\mathbb{E}[W_{i,app}] = \frac{a + b_i \rho}{1 - \rho}. \tag{3.10}$$

This form has a first order polynomial in the numerator rather than a second order polynomial as in $\mathbb{E}[W_{i,Boon}]$. Still, we require this approximation to be exact in LT and highly accurate in HT:

1. In LT, we have that $\mathbb{E}[W_i] = \mathbb{E}[S^{res}]$. Therefore we require $\mathbb{E}[W_{i,app}]|_{\rho=0} = a = \mathbb{E}[S^{res}]$.

2. The behavior of the mean waiting-time in HT has been analyzed in [23], where the following result has been obtained for the mean waiting time in HT:

$$\mathbb{E}[W_i] = \frac{\omega_i}{1 - \rho} + o((1 - \rho)^{-1}), \quad \rho \uparrow 1, \tag{3.11}$$

where ω_i can be thought of as a rate at which $\mathbb{E}[W_i]$ tends to infinity (in case of $\rho = 1$, the polling system is unstable). For exhaustive service, it is shown that

$$\omega_i = \frac{1 - \hat{\rho}_i}{2} \left(\frac{\sigma^2}{2\delta} + \mathbb{E}[S] \right). \tag{3.12}$$

where $\sigma^2 = \sum_{i=1}^{N} \hat{\nu}_i (\mathrm{Var}[B_i] + c_{A_i}^2 \mathbb{E}[B_i]^2)$. Since $c_{A_i}^2 = \frac{c_{EA_i}^2}{D_i}$ tends to 0 very rapidly as D_i increases, we may simplify σ^2 by $\sigma_{app}^2 = \sum_{i=1}^{N} \hat{\nu}_i \mathrm{Var}[B_i]$. We

opt to do so for reasons found below, such that we have that

$$\omega_{i,app} = \frac{1-\hat{\rho}_i}{2}\left(\frac{\sigma^2_{app}}{2\delta} + \mathbb{E}[S]\right). \tag{3.13}$$

This second requirement leads to $(1-\rho)\mathbb{E}[W_{i,app}]|_{\rho=1} = a + b_i = \omega_{i,app}$.

The above requirements result in the following approximation for the mean waiting time:

$$\mathbb{E}[W_{i,app}] = \frac{\mathbb{E}[S^{res}] + (\omega_{i,app} - \mathbb{E}[S^{res}])\rho}{1-\rho} = \mathbb{E}[S^{res}] + \frac{\omega_{i,app}\rho}{1-\rho}. \tag{3.14}$$

This leads to the definition of the approximative cost function

$$C_{app}(D_1, \ldots, D_N) = \sum_{i=1}^{N} c_i \left(\frac{D_i - 1}{2\lambda_i} + \mathbb{E}[W_{i,app}]\right). \tag{3.15}$$

Note that we have opted to approximate σ^2 and ω_i by σ^2_{app} and $\omega_{i,app}$. We use these approximations rather than the original expressions mainly because of a very interesting characteristic these approximations exhibit, which proves invaluable in the derivation of a closed-form approximation. When the ratios between batch sizes are known, these approximations can be evaluated while lacking knowledge of the batch sizes themselves. Accordingly, we have that the approximations retain their values, whenever batch sizes are changed in a way that the ratios between them remain the same. In σ^2_{app} we have the terms $\hat{\nu}_i$ and $\mathrm{Var}[B_i]$. The latter is by definition independent of the batch sizes, the former requires a bit more thought. We have that $\hat{\nu}_i = \frac{\hat{\lambda}_i}{D_i}$. When the batch sizes are changed such that the ratios between them remain the same, e.g., when $\vec{D} = (3, 6, 12)$ is changed to $(2, 4, 8)$ or $(6, 12, 24)$, but not to $(1, 2, 3)$, we have that the $\hat{\lambda}_i$ scale accordingly in order to keep the load evaluated at one, such that the $\hat{\nu}_i$ retain their values. Therefore, under the assumption of fixed ratios between the batch sizes, the $\hat{\nu}_i$ and therefore σ^2_{app} can be evaluated independently of the D_i themselves. This is not the case with σ^2, since the term $c^2_{A_i}$ is inversely proportional to D_i.
The "independence" of $\hat{\nu}_i$ also implies that $\hat{\rho}_i = \hat{\nu}_i\mathbb{E}[B_i]$ can be evaluated independently of the batch sizes under the same assumption. Therefore, $\omega_{i,app}$, which is dependent on $\hat{\rho}_i$ and σ^2_{app}, can also be fully identified using the ratios of the batch sizes, rather than the batch sizes themselves. The approximations prove to be rather accurate, since taking customers together in larger batches result in batch arrival processes converging to deterministic arrival processes, in which case these approximations are exact.

We conclude this section with the listing of two limiting characteristics of $\mathbb{E}[W_{i,app}]$:

- In case of deterministic switch-over times tending to infinity, a strong conjecture is presented in [40] stating that $\frac{W_i}{S}$ tends to a uniform distribution on $[0, \frac{1-\rho_i}{1-\rho}]$. This implies that $\frac{\mathbb{E}[W_i]}{S} \to \frac{1-\rho_i}{2(1-\rho)}$ when $S \to \infty$. It turns out that $\mathbb{E}[W_{i,app}]$ becomes exact under these circumstances, since

$$\lim_{S\to\infty} \frac{\mathbb{E}[W_{i,app}]}{S} = \frac{1-\rho_i}{2(1-\rho)} = \lim_{S\to\infty} \frac{\mathbb{E}[W_i]}{S}. \tag{3.16}$$

- For systems with $N = 1$ without vacations, i.e., a GI/G/1 queue, a widely used elementary approximation for $\mathbb{E}[W]$ — which is exact when assuming Poisson arrivals — reads

$$\mathbb{E}[W] \approx \frac{\rho\mathbb{E}[B](c_A^2 + c_B^2)}{2(1-\rho)}, \tag{3.17}$$

see for example (8) of [37]. In case of deterministic arrival streams and $N = 1$, $\mathbb{E}[W_{i,app}]$ coincides with this particular approximation. This is a desirable characteristic, since $c_{A_i}^2$ drops rapidly as D_i is increased.

One-dimensional problem

In this subsection, we solve the problem under the restriction that $D = D_1 = D_2 = \ldots = D_N$. In that case, the approximative cost function reduces to

$$C_{app}(D) = \sum_{i=1}^{N} \frac{c_i(D-1)}{2\lambda_i} + \sum_{i=1}^{N} c_i\mathbb{E}[S^{res}] + \frac{\sum_{i=1}^{N} c_i\omega_{i,app}\sum_{i=1}^{N}\lambda_i\mathbb{E}[B_i]}{D - \sum_{i=1}^{N}\lambda_i\mathbb{E}[B_i]}. \tag{3.18}$$

Since the ratios between the batch sizes are taken fixed in this one-dimensional case, $\omega_{i,app}$ behaves independently of the value of D, as illuminated earlier in the present section. This independency property allows a trivial derivation of the value D minimizing $C_{app}(D)$. In order to find this minimum, we evaluate $\frac{d}{dD}C_{app}(D)$ and equate the result to zero:

$$\frac{(D - \sum_{i=1}^{N}\lambda_i\mathbb{E}[B_i])^2\sum_{i=1}^{N}\frac{c_i}{2\lambda_i} - \sum_{i=1}^{N} c_i\omega_{i,app}\sum_{i=1}^{N}\lambda_i\mathbb{E}[B_i]}{(D - \sum_{i=1}^{N}\lambda_i\mathbb{E}[B_i])^2} = 0. \tag{3.19}$$

Solving this equation for D leads to the following optimal and feasible value for D:

$$D^{opt} = \sum_{i=1}^{N}\lambda_i\mathbb{E}[B_i] + \sqrt{2\left(\sum_{i=1}^{N}\frac{c_i}{\lambda_i}\right)^{-1}\left(\sum_{i=1}^{N} c_i\omega_{i,app}\right)\left(\sum_{i=1}^{N}\lambda_i\mathbb{E}[B_i]\right)}. \tag{3.20}$$

Since for $D \leq \sum_{i=1}^{N}\lambda_i\mathbb{E}[B_i]$ the system is unstable, and there are no boundaries to examine, the given expression for D^{opt} is the only one that minimizes $C_{app}(D)$. Because this expression can take fractional values, proper rounding is needed. We do this the same way as done in subsection 3.5.1, i.e., rounding to the nearest positive integer if that results in a stable system, rounding upwards to the nearest larger integer otherwise.

Multi-dimensional problem

In this subsection, the type-i batches are not necessarily equal for the different i, i.e., we relax the constraint that $D_1 = D_2 = \ldots = D_N$. In this case, the traditional approach would be to take partial derivatives of $C_{app}(D_1, \ldots, D_N)$ with respect to $D_1, D_2, \ldots, D_N$, and equate each of them to zero, resulting in a set of N equations with N unknowns. This however does not allow for easy solutions, since we do not have any assumptions on the ratios of the batch sizes. Lacking such an assumption, the property that $\omega_{i,app}$ can be evaluated independently of the D_i as illuminated earlier in this section does not hold, making analysis intractable. Therefore, we use a different approach. We will write $\vec{D} = \vec{d}E$, where $\vec{d} = (d_1, d_2, \ldots, d_N)$ is a constant N-dimensional vector. The elements of this vector can be thought of as proportional values of the optimal, *fractional* batch sizes before rounding. E.g., when $d_2/d_1 = 2$, then $D^{opt}_{2,frac}/D^{opt}_{1,frac} = 2$. When we have a vector of proportional values that is known to hold for the optimal value of $\vec{D}_{frac}$, the only task that remains is optimizing E, which then results in the optimal value of $\vec{D}$ itself. Since E is a mere one-dimensional constant, this is a one-dimensional optimization problem, where $\omega_{i,app}$ will act as a constant in E by its independency property. This allows for the same kind of analysis as carried out in Subsection 3.5.2.

Of course, first the vector $\vec{d}$ has to be identified. We will not derive a vector that is exact for any combination of input parameters, for the same reason we do not approach the matter as a multi-dimensional problem. Instead, we propose an approximation, which is exact in certain limiting cases. The problem described in Section 3.3 encompasses the search of the value of

$$\arg\min C(D_1, \ldots, D_N) = \arg\min_{\vec{D}} \sum_{i=1}^{N} c_i(\mathbb{E}[V_i] + \mathbb{E}[W_i]) \tag{3.21}$$

under the constraint that the system is stable, i.e., that the load in the polling system $\rho = \sum_{i=1}^{N} \nu_i \mathbb{E}[B_i]$ remains smaller than one. We now deploy an independence argument. Typically, optimal batch sizes are such that upon adopting those optimal batch sizes, ρ becomes very small. That is, the ρ_i are proportional to $\frac{1}{D_i}$, which means that for considerable batch sizes $\rho = \sum_{i=1}^{N} \rho_i$ is fairly moderate, which on its turn results in the system being near the state of LT. Since in LT it holds that $\mathbb{E}[W_i] = \mathbb{E}[S^{res}]$, we have that the mean waiting time is independent of the batch sizes in LT. This motivates us to use the ratios of the following problem as an approximation for $\vec{d}$, using the expression found for $\mathbb{E}[V_i]$ in Section 3.3:

$$\arg\min_{\vec{D}} \sum_{i=1}^{N} c_i(\mathbb{E}[V_i] + \mathbb{E}[S^{res}]) = \arg\min_{\vec{D}} \sum_{i=1}^{N} c_i \mathbb{E}[V_i] = \arg\min_{\vec{D}} \sum_{i=1}^{N} c_i \frac{D_i - 1}{2\lambda_i}, \tag{3.22}$$

under the constraint that $\sum_{i=1}^{N} \nu_i \mathbb{E}[B_i] < 1$. Since $\mathbb{E}[V_i]$ is monotonously decreasing in D_i, the optimal solution will be such that the slack of this constraint converges

to zero. In other words, in the optimum we have that $\sum_{i=1}^{N} \nu_i \mathbb{E}[B_i] + \epsilon = 1$, where $\epsilon \downarrow 0$. Although not conventionally correct, we will write $\sum_{i=1}^{N} \nu_i \mathbb{E}[B_i] = 1$ in the interest of easy notation, which gives birth to $D_N = \frac{\lambda_N \mathbb{E}[B_N]}{1 - \sum_{i=1}^{N-1} \nu_i \mathbb{E}[B_i]}$. Then, we can write this problem unconstrained as

$$\arg\min_{\vec{D}} \sum_{i=1}^{N-1} c_i \frac{D_i - 1}{2\lambda_i} + \frac{c_N}{2\lambda_N} \left(\frac{\lambda_N \mathbb{E}[B_N]}{1 - \sum_{i=1}^{N-1} \frac{\lambda_i \mathbb{E}[B_i]}{D_i}} - 1 \right). \tag{3.23}$$

In the optimal value of $\vec{D}$, we have that the derivative of the operand with respect to D_i equals zero, $1 \leq i < N$. This leads to the following equation:

$$D_i \left(1 - \sum_{k=1}^{N-1} \frac{\lambda_k \mathbb{E}[B_k]}{D_k} \right) = \lambda_i \sqrt{\frac{c_N}{c_i} \mathbb{E}[B_N] \mathbb{E}[B_i]}. \tag{3.24}$$

This equation immediately gives a relation between the optimal values for D_i and D_j $(1 \leq i, j < N)$:

$$\frac{D_i}{D_j} = \frac{D_i \left(1 - \sum_{k=1}^{N-1} \nu_k \mathbb{E}[B_k] \right)}{D_j \left(1 - \sum_{k=1}^{N-1} \nu_k \mathbb{E}[B_k] \right)} = \frac{\lambda_i \sqrt{\frac{c_N}{c_i} \mathbb{E}[B_N] \mathbb{E}[B_i]}}{\lambda_j \sqrt{\frac{c_N}{c_j} \mathbb{E}[B_N] \mathbb{E}[B_j]}} = \frac{\lambda_i}{\lambda_j} \sqrt{\frac{c_j \mathbb{E}[B_i]}{c_i \mathbb{E}[B_j]}}. \tag{3.25}$$

By symmetry, we also have that this result holds for $1 \leq i, j \leq N$. This strikingly simple result gives the exact relation between the optimal batch sizes whenever the polling system resides in LT, and a good approximation in case ρ is close to zero. Therefore, we use this result for the determination of $\vec{d}$. Since only the proportionality of these values is relevant, we set $d_1 = 1$. The other elements of $\vec{d}$ immediately follow:

$$\vec{d} = \left(1, \frac{\lambda_2}{\lambda_1} \sqrt{\frac{c_1 \mathbb{E}[B_2]}{c_2 \mathbb{E}[B_1]}}, \ldots, \frac{\lambda_N}{\lambda_1} \sqrt{\frac{c_1 \mathbb{E}[B_N]}{c_N \mathbb{E}[B_1]}} \right). \tag{3.26}$$

Note that the (3.26) is exact whenever $\mathbb{E}[W_i]$ is insensitive to the batch sizes. In practice this may happen in a deterministic production environment, where scheduling is done such that the mean batch waiting time $\mathbb{E}[W_i]$ is zero. Now $\vec{d}$ has been determined, we have reduced the current problem to a one-dimensional optimization problem. Writing $\vec{D} = \vec{d}E$, we are only left with the task of finding the optimal value of E. This problem is solved completely analogously to the solution of the problem in Subsection 3.5.2, leading to

$$E_{app}^{opt} = \sum_{i=1}^{N} \frac{\lambda_i \mathbb{E}[B_i]}{d_i} + \sqrt{2 \left(\sum_{i=1}^{N} \frac{c_i d_i}{\lambda_i} \right)^{-1} \left(\sum_{i=1}^{N} c_i \omega_{i,app} \right) \left(\sum_{i=1}^{N} \frac{\lambda_i \mathbb{E}[B_i]}{d_i} \right)}. \tag{3.27}$$

We end up with the following approximation of the optimal batch sizes:

$$\vec{D_{app}^{opt}} = (D_{1,app}^{opt}, D_{2,app}^{opt}, \ldots, D_{N,app}^{opt}) = (d_1 E_{app}^{opt}, d_2 E_{app}^{opt}, \ldots, d_N E_{app}^{opt}), \tag{3.28}$$

Notation	Parameter	Considered parameter values
N	Number of queues	$\{2, 5\}$
$\overline{\lambda_i} = \frac{1}{N}\sum_{i=1}^{N}\lambda_i$	Type-averaged arrival rate	$\{\frac{0.5}{N}, \frac{2}{N}\}$
$\overline{\mathbb{E}[B_i]} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[B_i]$	Type-averaged mean service time	$\{1\}$
$\overline{\mathbb{E}[S_i]} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[S_i]$	Type-averaged mean switch-over time	$\{0, 0.2, 1, 10\}$
$c^2_{A_i}$	SCV interarrival times	$\{0.25, 1, 2\}$
$c^2_{B_i}$	SCV service times	$\{0, 1, 4\}$
$c^2_{S_i}$	SCV switch-over times	$\{0, 1\}$
$\vec{c}$	Weight vector	$\{(\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5), (1, 1, 1, 1, 1), (5, 4, 3, 2, 1)\}$

Table 3.1: Parameter values of the test beds used in Section 3.6.

where d_i and E^{opt}_{app} are as given above. Since in this formulation the batch sizes can still take fractional values, we again use the same rounding strategy as used in the numerical approach. We round each of the values to the nearest positive integer if this results in a stable system, otherwise we round each of the values upwards to the nearest strictly larger integer. After rounding, $\vec{D^{opt}_{app}}$ gives a straightforward approximation of the solution to the problem formulation in Section 3.3. The obtained approximation requires little or no computation time. Unlike the numerical approach, the computation time required is hardly dependent on the number of queues or any other characteristics of the system at hand.

3.6 Validation

In this section, we assess the performance of the two solution methods presented in Section 3.5 by use of simulation. We compare the performance of the numerical approach and the closed-form approximation in terms of the cost function with respect to the cost of the minimum obtained by simulation. This is done based on two test beds containing 1260 polling system in total, which bring a wide variety of input parameters with them.

We assess the performance of the numerical approach and the closed-form approximation on a test bed of symmetric systems, and a test bed of asymmetric systems. The symmetric test bed consists of symmetric systems corresponding to each of the combinations possible of parameter values found in Table 3.1, which corresponds to a total of 504 systems. Note that in case of $\mathbb{E}[S_i] = 0$ for all i, the realisations of S_i are all zero, irrespective of the value of $c^2_{S_i}$. Moreover, $\vec{c} = (\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5)$ and $\vec{c} = (1, 1, 1, 1, 1)$ result in the same weight vector in case of symmetry. The asymmetric test bed consists of asymmetric systems corresponding to the same combinations found in Table 3.1, with a total of 756. Differences compared to the symmetric systems are found in the mean arrival rates and the mean service and switch-over times, they are not the same for each queue anymore. Let $\overline{\lambda_i}$, $\overline{\mathbb{E}[B_i]}$ and $\overline{\mathbb{E}[S_i]}$ be the type-averaged arrival rate, mean service time and mean switch-over time respectively. Then, the parameters are

taken asymmetrically through the formulas $\lambda_j = \frac{2j}{N+1}\overline{\lambda_i}$, $\mathbb{E}[B_j] = \frac{2j}{N+1}\overline{\mathbb{E}[B_i]}$ and $\mathbb{E}[S_j] = \frac{2(N+1-j)}{N+1}\overline{\mathbb{E}[S_i]}$, $1 \leq j \leq N$.

For the sake of validation of the numerical approach, an implementation of a Newton type-algorithm for unconstrained minimization was used, see [29] for details. To make sure that the polling system remains stable, a penalty term $A\mathbb{1}_{\{\rho \geq 1\}}$ was added to the cost function, where A is an unproportionally large constant. As starting values for the batch sizes, also unproportionally large values were used. For the validation of the derived closed-form approximation, the formula given in (3.28) was implemented.

We express the comparisons in terms of the cost function relative to simulation. Differences are expressed in a percentual, relative way, i.e.,

$$\Delta\% = \frac{a - s}{s} \times 100\%, \tag{3.29}$$

where s is the cost of the optimal batch size vector obtained by simulation, and a is the cost belonging to the batch size approximations of the respective approximation method. It should be noted that the closed-form approach only produces approximations of the optimal batch size vector, not the accompanying value of the cost function. To obtain the latter, the found optimal batch size approximations are used as input for simulation. More specifically, the costs are obtained by averaging over the results of multiple simulation runs of 1.000.000 time units, with the optimal batch sizes as input. Again simulation was used, because it is the only way of obtaining "exact" results. Note that this is also done for the approximations found by the numerical approach. Although $\mathbb{E}[W_{i,Boon}]$ could be used for evaluation as illuminated in Section 3.4, simulation was used in an effort to compare the two approximation methods in an unbiased way.

It should be noted that although both the numerical approach and the closed-form approximation only require at most the first two moments of the interarrival, service and switch-over time distributions as an input, all simulations done in this section need the distributions of them as a whole as input. Therefore, we have deployed the common two-moment distribution fits as described in Appendix B.

These preparations now allow us to compare the two solution techniques. Table 3.2 shows the mean differences in cost performances of the numerical approach and the closed-form approximation relative to simulation, categorized in the test beds used and each of the input parameters. From these tables, one can see that the two solution techniques and simulation are quite close together in terms of performance. Whenever total waiting times become longer, as is the case when $\overline{\mathbb{E}[S_i]} = 10$, $\overline{\lambda_i} = \frac{2}{N}$ or $N = 5$, we see that the numerical and closed-form approximations perform increasingly well with respect to simulation. In some cases they may even work better than simulation due to the fact that simulation errors

become such that they interfere with finding the exact optimum of the cost. In case waiting times are generally short, for example when $\overline{\mathbb{E}[S_i]} = 0$ or $\overline{\lambda_i} = \frac{0.5}{N}$, the batch size vector found by the numerical approach and the closed-form approximation usually coincide. However, whenever a difference in the found batch size vector occurs, the difference in terms of cost of that particular system may be considerable, such that the mean differences grow past 5%. It should be noted that although large relative differences may occur in these cases, the absolute differences are still very small, due to the fact that the waiting times themselves are small. Moreover, the relative difference in terms of the sojourn time would make a far less daunting impression. For example, the worst case of the closed-form approximation performance found in the test beds is the asymmetric system with $N = 2$, $\overline{\lambda_i} = 1$, $\overline{\mathbb{E}[S_i]} = 0.2$, $\overline{\mathbb{E}[S_i^2]} = 0.04$, $c_{B_i}^2 = 0$, $c_{A_i}^2 = 2$ and $\vec{c} = (1, 1)$. Simulation comes up with $D^{opt} = (2, 5)$, while D_{app}^{opt} evaluates to $(1, 4)$. This yields an increase in average waiting time of 189.88%. However, the absolute increase of waiting time is merely 1.76.
When considering the particular role of the value of N, a similar effect is observed. We see that when $N = 2$, the numerical approach works significantly better than the closed-form approximation. However, when $N = 5$, the performance with respect to simulation is better in both cases, and the difference between the methods becomes smaller. Combined with the fact the numerical approach does not scale well with N, this acts as an illustration of the complementary property of the pair of methods as discussed earlier. Also, some interaction effects of the performance with the SCVs of the interarrival times, service times and switch-over times can be observed. Finally, one can see that although often the case, the numerical approach does not always score equally well or better than the closed-form approximation judging by the results of the asymmetric systems with $\overline{\mathbb{E}[S_i]} = 1$ in the test bed. One may have expected this to hold in general, since $\mathbb{E}[W_{i,Boon}]$ approximates $\mathbb{E}[W_i]$ better than $\mathbb{E}[W_{i,app}]$.

To study the differences in the found optimal batch size results, the percentual relative differences in found batch sizes are given in Table 3.3. The results were broken down in the same categories as the results in Table 3.2. Judging from Table 3.3 one may conclude that the three solution techniques not only score comparably well in terms of cost, but also the optimal batch sizes found are similar. The same effects of the input parameters on the accuracy of the approximations can be observed. In particular, the relative differences for $N = 5$ of the two methods are almost equal, while this is not the case for $N = 2$. This again illustrates the complementary effect.

3.7 Influence of input parameters

Whereas simulation methods and numerical methods act as a sort of black box, the closed-form approximation obtained in Subsection 3.5.2 offers suggestions about

(a)

Test bed	Numerical	Closed-form
Symmetric	0.303	1.624
Asymmetric	0.645	4.848

(b)

Test bed	$\overline{\lambda_i} = \frac{0.5}{N}$		$\overline{\lambda_i} = \frac{2}{N}$	
	Numerical	Closed-form	Numerical	Closed-form
Symmetric	0.520	1.013	0.086	2.234
Asymmetric	0.696	4.543	0.594	5.153

(c)

Test bed	N=2		N=5	
	Numerical	Closed-form	Numerical	Closed-form
Symmetric	0.428	2.776	0.178	0.471
Asymmetric	0.870	7.715	0.419	1.981

(d)

Test bed	$c^2_{S_i} = 0$		$c^2_{S_i} = 1$	
	Numerical	Closed-form	Numerical	Closed-form
Symmetric	0.329	1.978	0.279	1.859
Asymmetric	0.763	5.837	0.683	5.649

(e)

Test bed	$\overline{\mathbb{E}[S_i]} = 0$		$\overline{\mathbb{E}[S_i]} = 0.2$	
	Numerical	Closed-form	Numerical	Closed-form
Symmetric	1.066	3.981	0.309	2.976
Asymmetric	1.272	12.009	1.477	9.393

Test bed	$\overline{\mathbb{E}[S_i]} = 1$		$\overline{\mathbb{E}[S_i]} = 10$	
	Numerical	Closed-form	Numerical	Closed-form
Symmetric	0.208	0.766	0.011	1.545
Asymmetric	0.221	-0.050	-0.077	0.026

(f)

Test bed	$\vec{c} = (\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5)$		$\vec{c}$ = (1,1,1,1,1)		$\vec{c}$ = (5,4,3,2,1)	
	Numerical	Closed-form	Numerical	Closed-form	Numerical	Sol. Tech. 3
Symmetric	0.250	0.635	0.250	0.635	0.356	2.613
Asymmetric	0.328	3.400	0.637	7.139	0.969	4.004

(g)

Test bed	$c^2_{A_i} = 0.25$		$c^2_{A_i} = 1$		$c^2_{A_i} = 2$	
	Numerical	Closed-form	Numerical	Closed-form	Numerical	Closed-form
Symmetric	0.231	0.352	0.312	1.420	0.365	3.099
Asymmetric	0.826	0.626	0.643	4.510	0.465	9.408

(h)

Test bed	$c^2_{B_i} = 0$		$c^2_{B_i} = 1$		$c^2_{B_i} = 4$	
	Numerical	Closed-form	Numerical	Closed-form	Numerical	Closed-form
Symmetric	0.082	2.789	0.262	0.687	0.565	1.395
Asymmetric	1.140	10.626	0.627	3.491	0.168	0.427

Table 3.2: Relative difference of the cost obtained by the approximation methods with the minimum cost obtained by simulation. The displayed percentual differences are categorized in the test beds, the approximation methods (a), and the type-averaged arrival rate (b), the number of queues (c), the SCV of the switch-over times (d), the type-averaged mean switch-over time (e), the weight vector (f), the SCV of the interarrival times (g) and the SCV of the service times (h).

(a)

Test bed	Numerical	Closed-form
Symmetric	4.169	5.499
Asymmetric	4.510	6.895

(b)

Test bed	$\overline{\lambda_i} = \frac{0.5}{N}$		$\overline{\lambda_i} = \frac{2}{N}$	
	Numerical	Closed-form	Numerical	Closed-form
Symmetric	4.001	4.127	4.336	6.872
Asymmetric	2.966	5.266	6.054	8.523

(c)

Test bed	N=2		N=5	
	Numerical	Closed-form	Numerical	Closed-form
Symmetric	5.008	7.206	3.330	3.792
Asymmetric	4.290	8.918	4.730	4.872

(d)

Test bed	$c^2_{S_i} = 0$		$c^2_{S_i} = 1$	
	Numerical	Closed-form	Numerical	Closed-form
Symmetric	3.764	5.107	4.527	5.915
Asymmetric	3.985	6.632	5.081	7.287

(e)

Test bed	$\overline{\mathbb{E}[S_i]} = 0$		$\overline{\mathbb{E}[S_i]} = 0.2$	
	Numerical	Closed-form	Numerical	Closed-form
Symmetric	3.981	5.595	3.233	5.376
Asymmetric	4.694	7.414	4.667	7.667

Test bed	$\overline{\mathbb{E}[S_i]} = 1$		$\overline{\mathbb{E}[S_i]} = 10$	
	Numerical	Closed-form	Numerical	Closed-form
Symmetric	3.345	4.712	6.023	5.404
Asymmetric	3.345	6.361	5.425	7.353

(f)

Test bed	$\vec{c} = (\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5)$		$\vec{c}$ = (1,1,1,1,1)		$\vec{c}$ = (5,4,3,2,1)	
	Numerical	Closed-form	Numerical	Closed-form	Numerical	Closed-form
Symmetric	3.406	4.193	0.250	0.635	4.932	6.805
Asymmetric	3.199	6.993	4.655	7.188	5.675	6.503

(g)

Test bed	$c^2_{A_i} = 0.25$		$c^2_{A_i} = 1$		$c^2_{A_i} = 2$	
	Numerical	Closed-form	Numerical	Closed-form	Numerical	Closed-form
Symmetric	3.800	4.010	3.890	5.414	4.816	7.075
Asymmetric	4.779	4.621	4.408	7.326	4.342	8.737

(h)

Test bed	$c^2_{B_i} = 0$		$c^2_{B_i} = 1$		$c^2_{B_i} = 4$	
	Numerical	Closed-form	Numerical	Closed-form	Numerical	Closed-form
Symmetric	3.181	5.388	3.763	5.044	5.562	6.066
Asymmetric	5.006	8.548	4.010	5.979	4.514	6.157

Table 3.3: Relative differences of the batch sizes obtained by the approximation methods with the simulated optimal batch sizes. The displayed percentual differences are categorized in the test beds, the solution techniques (a), and the type-averaged arrival rate (b), the number of queues (c), the SCV of the switch-over times (d), the type-averaged mean switch-over time (e), the weight vector (f), the SCV of the interarrival times (g) and the SCV of the service times (h).

the influence of the input parameters on the optimal batch size vector and implies some near-insensitivity properties.

It is suggested by (3.27) that the average optimal batch size increases directly in λ_i and $\mathbb{E}[B_i]$, and indirectly in $\mathbb{E}[B_i^2]$ and $\mathbb{E}[S]$ through the $\omega_{i,app}$-term, $1 \leq i \leq N$. This is natural behavior, since an increase in any of these terms naturally translates into an increase of $\mathbb{E}[W_i]$, which is remedied by taking larger batch sizes. Concerning the ratios of the optimal batch sizes, it is suggested by (3.26) that the optimal value of D_i relative to the optimal batch sizes corresponding to other queues is increasing in λ_i and $\mathbb{E}[B_i]$ and decreasing in c_i. Intuitively, this is also natural behavior. When λ_i and $\mathbb{E}[B_i]$ increase, ρ_i and thus $\mathbb{E}[W_i]$ become larger, which is remedied by taking D_i larger. A larger value of c_i implies a larger penalty for the value of $\mathbb{E}[V_i] + \mathbb{E}[W_i]$. Therefore, one will want to take D_i smaller and all other batch sizes larger, such that the value of $\mathbb{E}[V_i] + \mathbb{E}[W_i]$ is decreased.

Next to these observations, (3.26) and (3.27) also suggest the following near-insensitivity properties:

- The optimal batch sizes are near-insensitive to higher moments of the interarrival and switch-over time distributions,
- The ratios of the optimal batch sizes is not as sensitive to the weights and mean service times as it is to the arrival rates,
- The ratio of optimal batch sizes belonging to a pair of types/queues is nearly insensitive to characteristics of any other type/queue.

As these near-insensitivity properties give valuable insights into the behavior of the optimal batches, these insensitivity properties are further discussed below. Simulation results are presented to illustrate the validity of the results.

Near-insensitivity to higher moments of the interarrival and switch-over time distributions. Near-insensitivity to the higher moments of the interarrival and switch-over time distributions is suggested because the value of E_{app}^{opt} as well as $\vec{d}$ are not dependent on them. Insensitivity to the higher moments of the interarrival time distribution is a result of the simplification in $\omega_{i,app}$. This makes the approximation useful for practical purposes, because in reality information about more than the first two moments is often hard to get. The assumption of the average size being (near) independent of the higher moments was justified by the fact that the internal batch arrival processes to the polling system converge to deterministic arrival processes quickly as batch sizes increase. Independence of higher moments of the switch-over time distribution can be justified by the fact that these moments have no impact on the mean waiting time of customers that is due to the server being in a switch-over period and therefore have negligible impact on the batch sizes as well. To investigate the plausibility of near-insensitivity, consider

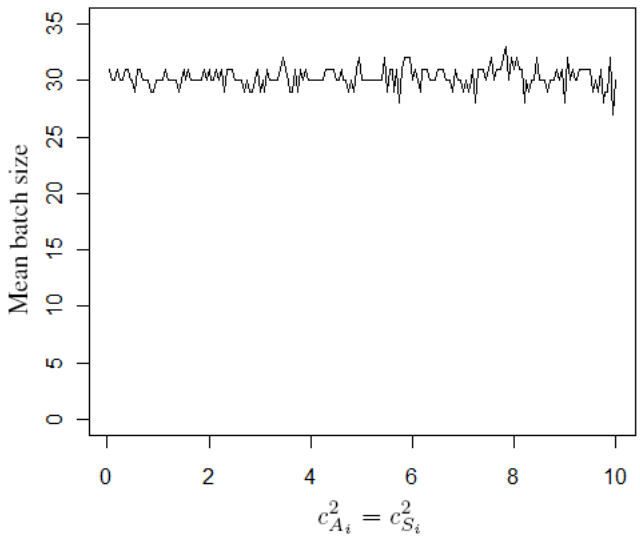


Figure 3.3: Mean batch size as a function of the SCV of the interarrival and the switch-over time distributions.

a symmetric system with 5 queues. The mean arrival rate of each queue equals $\lambda_i = 3$. Furthermore we have that for each queue $c_i = \mathbb{E}[B_i] = \mathbb{E}[S_i] = 1$, $\mathbb{E}[B_i^2] = 2$ and $c^2_{A_i} = c^2_{S_i}$ for $1 \leq i \leq N$. A small simulation study results in Figure 3.3. In this figure the behavior of the mean optimal exact batch size is given as a function of $c^2_{A_i} = c^2_{S_i}$. Apart from some noise which can be contributed to simulation error, the figure shows a near constant line, which again implies plausibility that optimal batch sizes are at least nearly insensitive to the higher moments of the interarrival and switch-over time distributions.

Lesser sensitivity of ratios to weights and mean service times compared to arrival rates. It may not be intuitive to see why the impact of the weights and the mean service times on the ratios is less than the mean arrival rates, as suggested by the placement of the square root in (3.26). The ratios as given in (3.26) are known not to hold exactly in general, however it is used for the general case in the closed-form approximation of (3.28). Therefore, we again undertake a small simulation study to study the validity of this assumption. Consider a two-type system where customers arrive according to Poisson processes — i.e., $c^2_{EA_i} = 1$ —, exponential service times and deterministic switch-over times with a duration of one time unit, and a variable k ranging from one to three. More specifically, we consider the three scenarios as given in Table 3.4. It is easily seen that the ratio as given in (3.26) evaluates to 1 in all three of the scenarios given in the table. Figure 3.4 shows the ratio of the exact optimal batch size values, $\frac{D_1^{opt}}{D_2^{opt}}$ obtained by simulation as a function of k for all scenarios as a function of k. Although these lines may seem to have a downward drift, this drift is quite faint compared

Scenario	Q_i	λ_i	c_i	$\mathbb{E}[B_i]$
Arrival rate vs. mean service time	Q_1	4	1	1
	Q_2	$5-k$	1	$\left(\frac{4}{5-k}\right)^2$
Arrival rate vs. weight	Q_1	4	1	1
	Q_2	$5-k$	$\left(\frac{5-k}{4}\right)^2$	1
Mean service time vs. weight	Q_1	4	1	1
	Q_2	4	k	k

Table 3.4: Parameter input used for the results found in Figure 3.4.

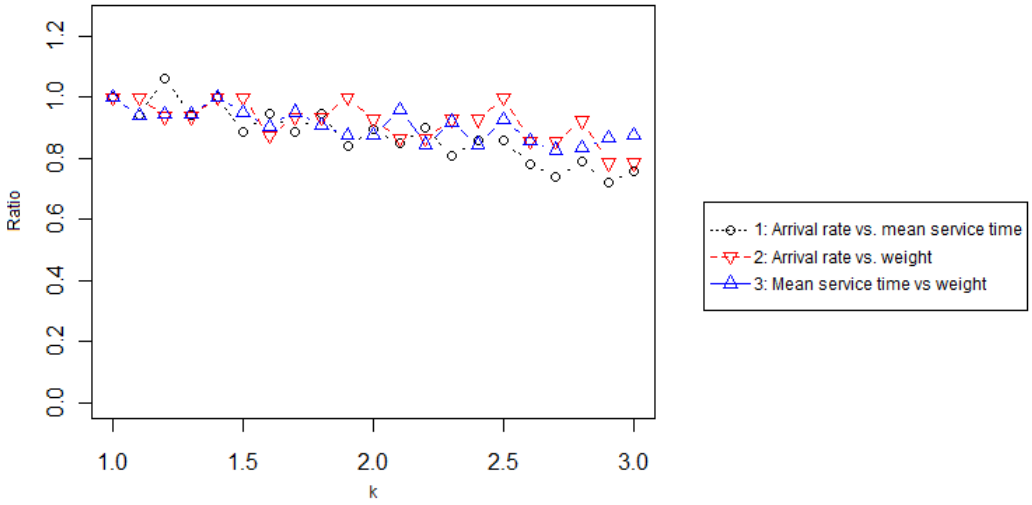


Figure 3.4: Ratio $\frac{D_1^{opt}}{D_2^{opt}}$ as a function of k.

to the observed integer effects and simulation errors. On top of that, the lines are close together. Taking into account the accompanying scenarios as given in Table 3.4, this reinforces the plausibility of lesser sensitivity to weights and mean service times when compared to arrival rates.

Near-insensitivity of ratios to other types or queues. As mentioned earlier, (3.26) suggests that the ratio of optimal batch sizes of two types is at least nearly independent of the characteristics of a third type or queue in the system. To investigate the plausibility of this suggestion, we again deploy a simulation study. Consider a three-type system with Poisson arrivals and exponential switch-over times, and again a variable k ranging from one to three. In addition, we have that $(\lambda_1, \lambda_2, \lambda_3) = (4, 2, k)$, $(\mathbb{E}[B_1], \mathbb{E}[B_2], \mathbb{E}[B_3]) = (1, 0.5, k)$, $(\mathbb{E}[B_1^2], \mathbb{E}[B_2^2], \mathbb{E}[B_3^2]) = (2, 0.25, k^2)$ and all switch-over times have a mean duration of one. Figure 3.5 shows the ratio of $\frac{D_2^{opt}}{D_1^{opt}}$ - which according to (3.26) should

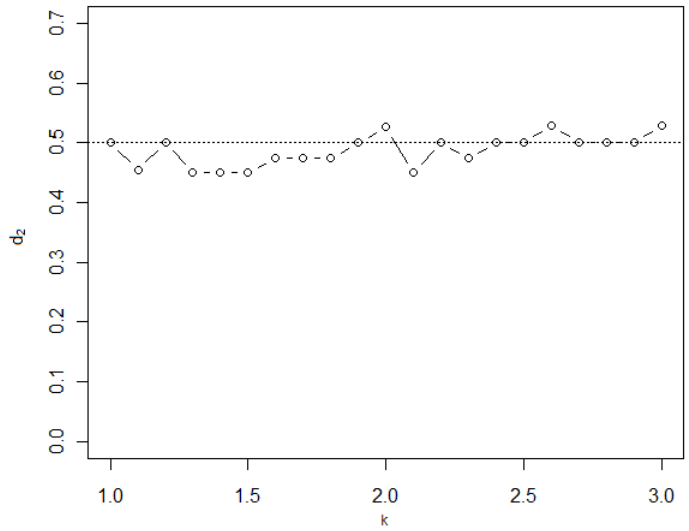


Figure 3.5: Ratio $\frac{D_2^{opt}}{D_1^{opt}}$ as a function of k.

be equal to 0.5 for all values of k - as a function of k by means of simulation. A more or less constant line can be observed, which indeed gives plausibility to the suggestion that the ratio of two optimal batch sizes is at least nearly independent of the characteristics of a third customer-type or queue.

We end this section with several other practical remarks concerning the obtained closed-form approximation.

Remark 1 (Ratio constraints). The derived closed-form approximative solution allows for constraints concerning the relation between the different batch sizes, by adjusting the d_i-values. For example, if one wants to take the batch sizes equal for each of the queues, we have that $D_1^{opt} = D_2^{opt} = \ldots = D_N^{opt}$ and therefore one could set $d_1 = d_2 = \ldots = d_N = 1$. In that case, (3.27) and (3.28) combined reduce to (3.20).

Remark 2 ($c\mu$-rule). The values taken for d_i bear resemblance to the well-known $c\mu$-rule (see for example [4] or [24]). This rule prescribes that in a multi-queue single server system with Poisson arrivals, zero switch-over times, free server routing and non-preemptive service, the server should always prioritize the service of customers belonging to queues with the highest value of $\frac{c_i}{\mathbb{E}[B_i]}$ in order to minimize the weighted mean number of customers in the queues.
To compare this with our model, note that taking smaller type-i batch sizes is equivalent to an increase in priority of Q_i. Then, V_i will generally become smaller in such a way that the mean total waiting time of type-i customers reduces — even though W_i will become larger —, and the load of the polling system will increase

at the expense of extra waiting time for other type-j customers. In order to be able to compare our model with models where the $c\mu$-rule is applicable, regard a system in which all the $V_i \to 0$, giving rise to $\lambda_i \to \infty$ for all the queues. In that case (3.25) suggests that

$$\frac{D_i}{D_j} = \sqrt{\frac{c_j}{\mathbb{E}[B_j]} \left(\frac{c_i}{\mathbb{E}[B_i]} \right)^{-1}}. \tag{3.30}$$

This equation suggests that if $\frac{c_i}{\mathbb{E}[B_i]} > \frac{c_j}{\mathbb{E}[B_j]}$, D_i should be taken smaller than D_j. This in accordance with the $c\mu$-rule, which states that if $\frac{c_i}{\mathbb{E}[B_i]} > \frac{c_j}{\mathbb{E}[B_j]}$, Q_i should get priority over Q_j.

3.8 Further research

The research done in the present chapter gives birth to a variety of directions for further research. In this section, we discuss some of the possibilities.

Higher moments and tail probabilities. In this chapter, the evaluation and optimization done encompassed the minimization of the weighted mean waiting time of customers. This could be extended to the minimization of higher moments of the waiting time or even tail probabilities of the waiting time. This may prove very interesting in practical situations, where the variance of the waiting times must be kept small, or where waiting thresholds are set, which are not to be exceeded. For the evaluation of weighted higher moments, the intuitive cost function $\sum_{i=1}^{N} c_i \mathbb{E}[(V_i + W_i)^n]$ may be replaced by $\sum_{i=1}^{N} c_i (\mathbb{E}[V_i^n] + \mathbb{E}[W_i^n])$, where $\mathbb{E}[W_i^n]$ may be estimated by using an approximation proposed by [8], the main result of the previous chapter. Especially in case of $\mathbb{E}[S] = 0$ this may prove to be very tractable. Then, optimization may be performed based on the obtained expressions.
The result of the previous chapter, an approximation of the complete waiting time distribution in polling systems with renewal arrivals, may also act as a basis for evaluation and optimization of tail probabilities.

Other service disciplines. We have only discussed the case where the server operates through an exhaustive service discipline to attend batches. The present study could be adjusted for models with other branching-type service policies [28] or even be generalized to them. The question remains whether a simple closed-form approximation can still be determined in the latter case.

Non-cyclic routing. Throughout this chapter, only systems where the server assumed a cyclic routing mechanism have been looked at. This could be extended to other routing mechanisms, or even generalized to routing according to general

polling tables. Since the behavior of the waiting time in LT and HT [26] is known under custom polling table routing, generalization may be feasible.

Model variations. In the model considered the service requirement of a batch is independent of the size of the batch, as commonly observed in practice. This could be extended to models with size-dependent service requirements. Also, in the model at hand the load of the polling system can become very small when batch sizes increase, which may result in the polling system having no batches waiting in its queues. Currently, the server will keep switching over in this case, while for example in production systems it may be interesting to let the server reside in idle mode. The HT characteristics of the waiting time will not change, since there will always be batches to serve. The LT characteristics of the waiting time is generally also easily identified, which allows for identification of $\mathbb{E}[W_{i,Boon}]$ and $\mathbb{E}[W_{i,app}]$ for idling policies. Following the approach of this paper, both numerical and closed-form approximations may then be found for this model variation.

Rounding strategy. Both the numerical approach and the closed-form approximation produce fractional batch sizes. To obtain integer batch sizes, we round each of the batch sizes to the nearest integer if that results in a stable system, otherwise we round each of them upwards to the nearest larger integer. One could attempt to improve the approximations by adopting more sophisticated rounding strategies. However, this may introduce extra complexity in the approximation, while the performance cannot greatly increase, judging by the already very nice results of Table 3.2.

Bibliography

[1] J.P.C. Blanc (1992). Performance evaluation of polling systems by means of the power-series algorithm. Annals of Operations Research 35, 155-186.

[2] M.A.A. Boon, E.M.M. Winands, I.J.B.F. Adan and A.C.C. van Wijk (2009). Closed-form waiting time approximations for polling systems. Eurandom Report No. 2009-030, Eindhoven.

[3] O.J. Boxma, J. van der Wal and U. Yechiali (2008). Polling with batch service. Stochastic Models 24, 604-625.

[4] C. Buyukkoc, P. Varaiya, and J. Walrand (1985). The $c\mu$-rule revisited. Advances in Applied Probability 17, 237-238.

[5] G. Choudhury and W. Whitt (1994). Computing transient and steady state distributions in polling models by numerical transform inversion. Performance Evaluation 25, 267-292.

[6] K.L. Chung (1974). A Course in Probability, 2nd ed. Academic Press, New York.

[7] R.K. Deb and R.F. Serfozo (1973). Optimal control of batch service queues. Advances in Applied Probability 5, 340-361.

[8] J.L. Dorsman, R.D. van der Mei and E.M.M. Winands (2010). A new method for deriving waiting-time approximations in polling systems with renewal arrivals. Manuscript submitted for publication.

[9] D. Down (1998). On the stability of polling models with multiple servers. Journal of Applied Probability 35, 925-935.

[10] D. Everitt (1986). Simple approximations for token rings. IEEE Transactions on Communications 34, 719-721.

[11] M.J. Ferguson (1986). Computation of the variance of the waiting time for token rings. IEEE Journal on Selected Areas in Communications 4, 775-782.

[12] C. Fricker and R. Jaibi (1981). Monotonicity and stability of periodic polling models. Queueing Systems 15, 211-238.

[13] S.W. Fuhrmann (1981). Performance analysis of a class of cyclic schedules. Bell laboratories technical memorandom 81-59531-1.

[14] S.W. Fuhrmann (1992). A decomposition result for a class of polling models. Queueing Systems 11, 109-120.

[15] S.W. Fuhrmann and R.B. Cooper (1985). Stochastic decompositions in the M/G/1 queue with generalized vacations. Operations Research Letters 33, 1117-1129.

[16] H. Gold and P. Tran-Gia (1993). Performance analysis of a batch service queue arising out of manufacturing system modelling. Queueing Systems 14, 413-426.

[17] W. Henderson and P.G. Taylor (1990). Product form in networks of queues with batch arrivals and batch service. Queueing systems 6, 71-88.

[18] K.K. Leung (1991). Cyclic-service systems with probabilistically-limited service. IEEE Journal on Selected Areas in Communications 9, 185-193.

[19] H. Levy and M. Sidi (1990). Polling systems: applications, modeling, and optimization. IEEE Transactions on Communications 38, 1750-1760.

[20] Z. Liu and P. Nain (1992). Optimal scheduling in some multiqueue single-server systems. IEEE Transactions on Automatic Control 37, 247-252.

[21] C. Mack (1957). The efficiency of N machines uni-directionally patrolled by one operative when walking time is constant and repair times are variable. Journal of the Royal Statistical Society Series B 19, 173-178.

[22] C. Mack, T. Murphy and N.L. Webb (1957). The efficiency of N machines uni-directionally patrolled by one operative when walking time and repair times are constants. Journal of the Royal Statistical Society Series B 19, 166-172.

[23] R.D. van der Mei and E.M.M. Winands (2008). A note on polling models with renewal arrivals and nonzero switch-over times. Operation Research Letters 36, 500-505.

[24] P. Nain and D. Towsley (1994). Optimal scheduling in a machine with stochastic varying processing rate. IEEE Transactions on Automatic Control 39, 1853-1855.

[25] T.L. Olsen (2001). Approximations for the waiting time distribution in polling models with and without state-dependent setups. Operations Research Letters 28, 113-123.

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

[27] M.P. Van Oyen and D. Teneketzis (1996). Optimal batch service of a polling system under partial information. Mathematical Methods of Operations Research 44, 401-419.

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

[29] R.B. Schnabel, J.E. Koontz and B.E. Weiss (1985). A modular system of algorithms for unconstrained minimization. ACM Transactions on Mathematical Software 11, 419-440.

[30] H. Takagi (1985). Analysis of polling systems. MIT Press, Cambridge.

[31] H.C. Tijms (1994). Stochastic models: an algorithmic approach. Wiley, Chichester.

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

[33] M. Vlasiou and U. Yechiali (2008). M/G/∞ polling systems with random visit times. Probability in the Engineering and Informational Sciences 22, 81-105.

[34] J. van der Wal and U. Yechiali (2003). Dynamic visit-order rules for batch-service polling. Probability in the Engineering and Informational Sciences 17, 351-367.

[35] H.J. Weiss (1979). The computation of optimal control limits for a queue with batch services. Management Science 25, 320-328.

[36] W. Whitt (1982). Refining diffusion approximations for queues. Operations Research Letters 1, 165-169.

[37] W. Whitt (1982). The Marshall and Stoyan bounds for IMRL/G/1 queues are tight. Operations Research Letters 1, 209-213.

[38] W. Whitt (1989). An interpolation approximation for the mean workload in a GI/G/1 queue. Operations Research 37, 936-952.

[39] E.M.M. Winands (2007). Polling, production & priorities. PhD thesis, Eindhoven University of Technology.

[40] E.M.M. Winands (2009). Branching-type polling systems with large setups. To appear in OR Spectrum.

[41] E.M.M. Winands, R.D. van der Mei and M.A.A. Boon (2010). Applications of polling systems. Working paper.

[42] P.H. Zipkin (1985). Models for design and control of stochastic multi-item batch production systems. Operations Research 34, 91-104.

Appendix A

Boon's approximation

There are no closed-form expressions or numerical algorithms available in literature for the exact computation of the mean waiting time in polling systems with renewal arrivals. A well-performing closed-form approximation however has been derived by Boon et al. [2]. This approximation is not only shown to perform very well, it also is easily implementable in software and gives insights in the impact of system parameters on the mean waiting time. Adopting the notation as given in Section 3.2, Boon's approximation $\mathbb{E}[W_{i,Boon}]$ can be expressed as follows:

$$\mathbb{E}[W_{i,Boon}] = \frac{K_0 + K_{1,i}\rho + K_{2,i}\rho^2}{1-\rho}, \tag{A.1}$$

where the constants $K_0, K_{1,i}$ and $K_{2,i}$ depend on several parameters of the polling system at hand. In case all queues receive exhaustive service, we have

$$\begin{aligned}
K_0 &= \mathbb{E}[S^{res}], \\
K_{1,i} &= \hat{\rho}_i((c_{A_i}^2)^4 \mathbb{1}_{\{c_{A_i}^2 \leq 1\}} + 2\frac{c_{A_i}^2}{c_{A_i}^2+1}\mathbb{1}_{\{c_{A_i}^2>1\}} - 1)\mathbb{E}[B_i^{res}] + \mathbb{E}[B^{res}] \\
&\quad + \hat{\rho}_i(\mathbb{E}[S^{res}] - \mathbb{E}[S]) - \frac{1}{\mathbb{E}[S]}\sum_{j=0}^{N-1}\sum_{k=0}^{j}\hat{\rho}_{i+k}\mathrm{Var}[S_{i+j}], \\
K_{2,i} &= \frac{1-\hat{\rho}_i}{2}\Big(\frac{\sum_{j=1}^{N}\hat{\lambda}_j(\mathrm{Var}[B_j] + \hat{\rho}_j^2\mathrm{Var}[\hat{A}_j])}{\sum_{j=1}^{N}\hat{\rho}_j(1-\hat{\rho}_j)} + \mathbb{E}[S]\Big) - K_0 - K_{1,i}.
\end{aligned}$$

If all queues receive gated service, we have

$$
\begin{aligned}
K_0 &= \mathbb{E}[S^{res}], \\
K_{1,i} &= \hat{\rho}_i((c_{A_i}^2)^4 \mathbb{1}_{\{c_{A_i}^2 \leq 1\}} + 2\frac{c_{A_i}^2}{c_{A_i}^2+1}\mathbb{1}_{\{c_{A_i}^2>1\}} - 1)\mathbb{E}[B_i^{res}] + \mathbb{E}[B^{res}] \\
&\quad + \hat{\rho}_i\mathbb{E}[S^{res}] - \frac{1}{\mathbb{E}[S]}\sum_{j=0}^{N-1}\sum_{k=0}^{j}\hat{\rho}_{i+k}\mathrm{Var}[S_{i+j}], \\
K_{2,i} &= \frac{1+\hat{\rho}_i}{2}\Big(\frac{\sum_{j=1}^{N}\hat{\lambda}_j(\mathrm{Var}[B_j] + \hat{\rho}_j^2\mathrm{Var}[\hat{A}_j])}{\sum_{j=1}^{N}\hat{\rho}_j(1+\hat{\rho}_j)} + \mathbb{E}[S]\Big) - K_0 - K_{1,i}.
\end{aligned}
$$

The idea behind this approximation can be explained as follows. The restrictions imposed on the approximation are firstly that the formula should be closed-form and allow easy implementation, since these are necessities for optimization purposes and the development of software tools. Secondly, the approximation should capture the light traffic limit and the heavy traffic limit behavior in an exact way. To comply to these restrictions, the form above with $(1-\rho)$ in the denominator and a second order polynomial in the numerator was chosen.
It is proved in [23] that capturing the HT behavior requires having the term $(1-\rho)$ in the denominator. Furthermore, this term is generally a common thing to have in exact expressions for the mean waiting time in queueing systems. Perhaps the most trivial example of this is the Pollaczek-Khintchine formula for the $M/G/1$ queue.
Having a second order polynomial fulfills the need for simplicity and is sufficient to obtain exact results for several limiting cases.

The parameters K_0, $K_{1,i}$ and $K_{2,i}$ were chosen such that the approximation satisfies the following three light traffic (LT) and heavy traffic (HT) properties:

1. LT requirement: $\mathbb{E}[W_{i,Boon}]|_{\rho=0} = \mathbb{E}[W_i]|_{\rho=0}$,
2. LT requirement: $\frac{d}{d\rho}\mathbb{E}[W_{i,app}]|_{\rho=0} = \frac{d}{d\rho}\mathbb{E}[W_i]|_{\rho=0}$,
3. HT requirement: $(1-\rho)\mathbb{E}[W_{i,app}]|_{\rho=1} = (1-\rho)\mathbb{E}[W_i]|_{\rho=1}$.

When the system is in the state of LT, the server will be switching constantly. Therefore, if a certain type-i customer would arrive, its waiting time would equal a residual switch-over time needed by the server to get to Q_i, giving rise to $\mathbb{E}[W_i]|_{\rho=0} = \mathbb{E}[S^{res}]$. Since the system is completely empty, this is completely insensitive to the service discipline.
An expression for $\frac{d}{d\rho}\mathbb{E}[W_i]|_{\rho=0}$ is derived in [2], both for the exhaustive and gated discipline. This is mainly done using the well-known Fuhrmann-Cooper decomposition [14] in combination with a LT theorem found by Whitt [38].

The Fuhrmann-Cooper decomposition 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. Combining this with Whitt's result, the LT limit of the mean waiting time in a GI/G/1 queue, ultimately results in an expression for $\frac{d}{d\rho}\mathbb{E}[W_i]|_{\rho=0}$.
For the exhaustive and gated service disciplines, an expression for $(1-\rho)\mathbb{E}[W_i]|_{\rho=1}$ is derived in [23]. Using these expressions, the three requirements above together completely determine the parameters K_0, $K_{1,i}$ and $K_{2,i}$ as given above, which completely identifies Boon's approximation of the mean waiting time.

Appendix B

Two-moment fits

In Chapters 2 and 3, at most the first two moments of the interarrival, service and switch-over time distributions are involved in the analysis performed. Simulations were used as a tool to measure the performance of the approximations obtained. Simulation methods however need complete distributions as an input rather than just the first two moments of each of them. Therefore, we have used various two-moment distribution fits throughout.
Let X be a random variable with first moment $\mathbb{E}[X]$, second moment $\mathbb{E}[X^2]$ and squared coefficient of variation (SCV) $c_X^2 = \frac{\mathbb{E}[X^2]}{\mathbb{E}[X]^2} - 1$. When $\mathbb{E}[X]$ and $\mathbb{E}[X^2]$ are known, the simulation tools will fit one of the following common distributions to X, characteristic for the SCV at hand (cf. [31]).

Case $c_X^2 = 0$. In case of a SCV of zero, X is taken to be deterministic with value $\mathbb{E}[X]$.

Case $0 < c_X^2 < 1$. Whenever the SCV of X takes a value between zero and one, a mixed Erlang distribution is fitted. The density function of this distribution is

$$f_X(x) = \left(p\frac{k-1}{lx} + (1-p) \right) f_Y(x), \tag{B.1}$$

where

$$k = \lceil \frac{1}{c_X^2} \rceil, \quad p = \frac{kc_X^2 - \sqrt{k(1+c_X^2) - k^2 c_X^2}}{1+c_X^2}, \quad l = \frac{k-p}{\mathbb{E}[X]},$$

and Y is a Erlang distributed random variable with shape parameter k, rate parameter l and density function

$$f_Y(y) = \frac{e^{-ly} l^k y^{k-1} \mathbb{1}_{\{y \geq 0\}}}{(k-1)!}.$$

The distribution fitted to X can be interpreted as a mixture of two Erlang distributions with shape parameters $k-1$ and k respectively. Both of these Erlang distributions can thus each be interpreted as a sum of $k-1$ and k exponential distributions respectively.

Case $c_X^2 = 1$. When the SCV equals one, X is taken to be exponential with rate parameter $\frac{1}{\mathbb{E}[X]}$.

Case $c_X^2 > 1$. For a SCV larger than one, a hyperexponential distribution with two phases (H_2) is fitted, having density function

$$f_X(x) = p\omega_1 e^{-\omega_1 x} + (1-p)\omega_2 e^{-\omega_2 x}, \tag{B.2}$$

where

$$p = \frac{1}{2}\big(1 + \sqrt{\frac{c_X^2 - 1}{c_X^2 + 1}}\big), \quad \omega_1 = \frac{2p}{\mathbb{E}[X]} \quad \text{and} \quad \omega_2 = \frac{2(1-p)}{\mathbb{E}[X]},$$

under the assumption of balanced means, i.e., $\frac{p}{\omega_1} = \frac{1-p}{\omega_2}$.