EURANDOM PREPRINT SERIES
2011-038 October, 2011

Polling systems with batch service

J.L. Dorsman, R.D. van der Mei, E.M.M. Winands
ISSN 1389-2355

Polling systems with batch service

J.L. Dorsman [*] [†]
j.l.dorsman@tue.nl

R.D. van der Mei [†] [‡]
R.D.van.der.Mei@cwi.nl

E.M.M. Winands [‡]
e.m.m.winands@vu.nl

October 31, 2011

Abstract

Motivated by applications in production and computer-communication systems, we study an N-queue polling system, consisting of an *inner* part and an *outer* part, and where products receive service in batches. Type-i products arrive at the outer system according to a renewal process and accumulate into a type-i batch. As soon as D_i products have accumulated, the batch is forwarded to the inner system where the batch is processed. The service requirement of a type-i batch is independent of its size D_i. For this model we study the problem of determining the combination of batch sizes $\vec{D}^{(opt)}$ that minimizes a weighted sum of the mean waiting times. This model does not allow for an exact analysis. Therefore, we propose a simple closed-form approximation for $\vec{D}^{(opt)}$, and present a numerical approach, based on the recently-proposed mean waiting-time approximation in [1]. Extensive numerical experimentation shows that the numerical approach is slightly more accurate than the closed-form solution, while the latter provides explicit insights into the dependence of the optimal batch sizes on the system parameters and into the behavior of the system. 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.

Keywords: Polling systems, renewal arrivals, batch arrivals, optimal batch sizes

1 Introduction

This paper is motivated by the *stochastic economic lot scheduling problem* (SELSP), where multiple standardized products have to be produced on a single machine (see [23] for a survey on the SELSP). Oftentimes, in this setting a group of products is combined in a batch for production at the machine. The time required to process such a batch depends at most weakly on the size of the batch, because the 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. [25]). Moreover, for a case study on the SELSP with batch service we refer to [24]. Batch-service processing is also widely applicable in the field of computer-communication systems, such as videotex systems and Time Division Multiple Access (TDMA) systems [9]. Management of production facilities often faces a complex trade-off concerning the determination of batch sizes. On the one hand, the smaller the batch size the less product inventory is needed and the shorter the waiting times within a batch. On the other hand, the larger the batch size the fewer batches require processing, which leads to reduction of the workload and the waiting times for the batch as a whole. The goal of this paper is to propose and evaluate methods for properly balancing this trade-off.

In many practical settings the products are produced according to a fixed production sequence, which naturally leads to the modelling via a so-called *polling system*. A typical polling system consists of a number of queues, attended by a single server in a fixed order. There is a body of literature available on

[*]EURANDOM, Department of Mathematics and Computer Science, Eindhoven University of Technology, The Netherlands.
[†]Probability and Stochastic Networks, CWI, Amsterdam, The Netherlands.
[‡]Faculty of Sciences, Department of Mathematics, VU University, Amsterdam, The Netherlands.

non-polling queueing systems with batch service, both from a performance-evaluation perspective [6, 7], and from a design perspective [3, 21, 25]. Despite the fact that the analysis and optimization of polling models have been studied extensively [8, 17, 18], remarkably little attention has been paid to polling models in combination with batch service. As an exception, Boxma et al. [2] study batch-service polling models in which batches are served integrally. Vlasiou and Yechiali [19] 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 [9] and when routing must be done in subsequent Hamiltonian tours in [20]. In [14], the question is studied whether upon arrival the server should poll a station or idle until more products 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 products for service at a time and that products arrive according to Poisson arrival processes. For polling models with renewal arrivals, hardly any exact results are known, except for asymptotic regimes for heavy traffic [13] or large switch-over times [22]. Faced by this, approximations have been developed for the mean waiting time [1], recently extended to the complete waiting-time distribution [4].

In the batch-service models addressed above, there is no limit to the number of products served during one visit of the server to a queue. In this paper, we consider a model that is fundamentally different, and that a type-i batch consists of exactly D_i products. More precisely, we study an N-queue batch-service polling system consisting of an inner part and an outer part. Type-i products arrive at the outer system according to a renewal process and accumulate into a type-i batch until exactly D_i products have accumulated; thus, products have to wait in the outer part until the batch is full. Then, 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 its constituent products, wait until the batch is processed by the server. It is assumed that the service requirement of a type-i batch is independent of its 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 of products in the outer and the inner system. In the absence of exact analysis, we present two approaches to approximate $\vec{D}^{(opt)}$: (1) a numerical approach, and (2) a closed-form approximation. The numerical approach is built upon the mean waiting-time approximation developed in [1], which is accurate over the entire range of parameters. Extensive validation shows that the numerical approach is slightly more accurate but does not scale in the number of queues, while the closed-form approximation works particularly well for large numbers of queues. As a by-product of the closed-form approximation, we observe near-insensitivity properties of the batch sizes with respect to the higher moments of the interarrival and switch-over time distributions. Moreover, the results suggest that the ratio of the optimal batch sizes of two queues is nearly insensitive to the characteristics of other queues.

The structure of this paper is as follows. In Section 2 the model is introduced and the optimization problem is formulated. In Section 3 we analyze the performance of the system. In Section 4 we consider optimization of the system performance, and propose two approximative solution approaches. In Section 5 the accuracy of these two approaches is extensively validated by a large simulation testbed. Finally, in Section 6 we address a number of topics for further research.

2 Model description and notation

We consider a polling systems consisting of an *outer part* and an *inner part* (see Figure 1). The outer system consists of $N > 1$ accumulation stations, where type-i products accumulate at their type-specific station, and the inner system is a classical cyclic-service polling system consisting of N infinite-sized buffers, denoted by $Q_1, \ldots, Q_N$. Type-i products arrive at the outer part according to a renewal arrival process, where the interarrival times are i.i.d. samples from a random variable $A_i^{(out)}$. The arrival rate of type-i products is $\lambda_i^{(out)} = 1/\mathbb{E}[A_i^{(out)}]$, and $\Lambda^{(out)} := \sum_{i=1}^{N} \lambda_i^{(out)}$ is the total arrival rate to the system. Type-i products have to wait in the outer system until exactly $D_i > 0$ type-i products have accumulated in the outer part. As soon as D_i type-i products have accumulated, they form a type-i batch, which is im-

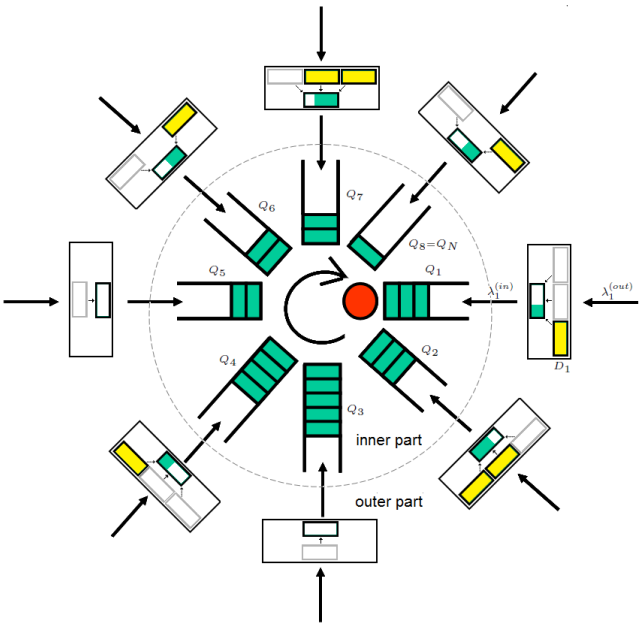

Figure 1: Illustration of the model for $N = 8$.

mediately forwarded to Q_i in the inner system. Note that the arrival of batches at Q_i also forms a renewal process. The interarrival times of these batches are denoted by the random variable $A_i^{(in)}$, whose distribution is the D_i-fold convolution of $A_i^{(out)}$. The type-i batch-arrival rate is denoted by $\lambda_i^{(in)} := 1/\mathbb{E}[A_i^{(in)}]$, and the total arrival rate is denoted by $\Lambda^{(in)} := \sum_{i=1}^{N} \lambda_i^{(in)}$. Note that $\lambda_i^{(in)} = \lambda_i^{(out)}/D_i$.

In the polling system, batches in each queue are served in the order of arrival. The service time of a type-i batch is denoted by the random variable B_i, *independent* of the fixed batch size D_i; thus, the batch size D_i has no impact on the service requirement B_i of the batch itself. Whenever a batch is being processed, it is assumed that the constituent products are all served simultaneously, such that there is no underlying product in the batch that has its service requirement completed before the batch as a whole is served. Let B denote the service time of an *arbitrary* batch, regardless of its type. 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 if Q_i is empty. The server moves along the queues in a cyclic manner, in the order $Q_1, Q_2, \ldots, Q_N, Q_1, \ldots$. The time needed to switch from Q_i to the next 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 from Q_i. Let $S := \sum_{i=1}^{N} S_i$ denote the total switch-over time per cycle. Let $\rho_i = \lambda_i^{(in)}\mathbb{E}[B_i]$ be the load offered to Q_i, and let $\rho := \sum_{i=1}^{N} \rho_i$ be the total load offered to the (inner) system.

A necessary and sufficient condition for the stability of the inner system, and hence of the whole system, reads $\rho < 1$ (cf. [5]). Throughout it is assumed that the stability condition is met. Note that for given service-time distributions and arrival rates to the *outer system*, the load ρ to the inner system generally depends on $\vec{D}$. Let $\mathcal{D}$ be the set of possible values for $\vec{D}$ for which the stability condition $\rho < 1$ holds.

The following notation is useful. For a non-negative random variable X with finite variance, we denote its residual counterpart by $X^{(res)}$, and the squared coefficient of variation (SCV) by c_X^2. Moreover, let $\mathbb{1}_A$ be the indicator function on the event A. Finally, to properly define light- and heavy-traffic limits (i.e., $\rho \downarrow 0$ and $\rho \uparrow 1$, respectively), we scale the system such that the total arrival rate at the outer system $\Lambda^{(out)}$ is varied while the batch-size vector, service-time distributions and the *ratios* between the individual external arrival rates are kept fixed. Moreover, it proves convenient to represent with $\hat{x}$ the value of each variable x

that is a function of ρ, *evaluated at* $\rho = 1$. Finally, let

$$\sigma^2 := \sum_{i=1}^{N} \hat{\lambda}_i^{(in)} \left(\text{Var}[B_i] + c^2_{A_i^{(in)}} \mathbb{E}[B_i]^2 \right), \text{ and } \delta := \sum_{j=1}^{N} \sum_{k=j+1}^{N} \hat{\rho}_j \hat{\rho}_k. \tag{1}$$

Optimization problem

The optimization problem studied in this paper concerns the 'optimal' choice of the vector of batch sizes $\vec{D} := (D_1, \ldots, D_N)$. To this end, note that the waiting time W_i of a type-i product can be decomposed as

$$W_i = W_i^{(out)} + W_i^{(in)}, \tag{2}$$

where $W_i^{(out)}$ is the time a type-i product has to wait in the outer system until its type-i batch is full, and where $W_i^{(in)}$ is the time a type-i product spends waiting in the inner system at Q_i before entering service. We define the cost function as follows: for $\vec{D} \in \mathcal{D}$,

$$C(\vec{D}) := \sum_{i=1}^{N} c_i \mathbb{E}[W_i] = \sum_{i=1}^{N} c_i \left(\mathbb{E}[W_i^{(out)}] + \mathbb{E}[W_i^{(in)}] \right), \tag{3}$$

where $c_i > 0$ denotes the cost per time unit for a type-i product in the system. Then for given weight vector $\vec{c} = (c_1, \ldots, c_N)$ the optimization problem is to find a vector $\vec{D}^{(opt)} = (D_1^{(opt)}, \ldots, D_N^{(opt)})$ that minimizes $C(\vec{D})$ over all $\vec{D} \in \mathcal{D}$.

3 Analysis

To tackle the optimization problem, we observe that a closed-form expression for $\mathbb{E}[W_i^{(out)}]$ can be obtained by conditioning. To this end, let E_{ij} be the event that an arbitrary arriving type-i product is the j-th arriving product in a type-i batch ($j = 1, \ldots, D_i$). Then by conditioning on E_{ij} we have

$$\mathbb{E}[W_i^{(out)}] = \sum_{j=1}^{D_i} \mathbb{E}[W_i^{(out)} | E_{ij}] \mathbb{P}[E_{ij}], \tag{4}$$

where

$$\mathbb{E}[W_i^{(out)} | E_{ij}] = \frac{D_i - j}{\lambda_i^{(out)}}, \text{ and } \mathbb{P}[E_{ij}] = \frac{1}{D_i} \text{ for all } j = 1, \ldots, D_i. \tag{5}$$

Unfortunately, there is no exact expression available for $\mathbb{E}[W_i^{(in)}]$. However, Boon et al. [1] recently proposed the following *approximation* for $\mathbb{E}[W_i^{(in)}]$, considered as a function of the load ρ to the inner system for $i = 1, \ldots, N$, $\rho < 1$:

$$\mathbb{E}[W_i^{(in)}] \approx \mathbb{E}[W_{i,app}^{(in)}] := \frac{K_0 + K_{1,i}\rho + K_{2,i}\rho^2}{1 - \rho}, \tag{6}$$

where the constants K_0, $K_{1,i}$ and $K_{2,i}$ are defined as follows:

$$K_0 = \mathbb{E}[S^{(res)}] = \frac{\mathbb{E}[S^2]}{2\mathbb{E}[S]}, \tag{7}$$

$$K_{1,i} = \hat{\rho}_i \left((c^2_{A_i^{(in)}})^4 \mathbb{1}_{\left\{ c^2_{A_i^{(in)}} \leq 1 \right\}} + 2 \frac{c^2_{A_i^{(in)}}}{c^2_{A_i^{(in)}} + 1} \mathbb{1}_{\left\{ c^2_{A_i^{(in)}} > 1 \right\}} - 1 \right) \mathbb{E}[B_i^{(res)}]$$

$$+ \mathbb{E}[B^{(res)}] + \hat{\rho}_i (\mathbb{E}[S^{(res)}] - \mathbb{E}[S]) - \frac{1}{\mathbb{E}[S]} \sum_{j=0}^{N-1} \sum_{k=0}^{j} \hat{\rho}_{i+k} \text{Var}[S_{i+j}], \tag{8}$$

$$K_{2,i} = \frac{1 - \hat{\rho}_i}{2} \left(\frac{\sigma^2}{2\delta} + \mathbb{E}[S] \right) - K_0 - K_{1,i}. \tag{9}$$

Note that the approximation in (6)-(9) is asymptotically exact both in light-traffic (i.e. when $\rho \downarrow 0$) and in heavy-traffic (i.e. $\rho \uparrow 1$), and highly accurate over a wide range of parameter combinations [1]. Next, combining (3)-(6), the cost function $C(\vec{D})$ can be approximated by the following expression: for $\vec{D} \in \mathcal{D}$,

$$C(\vec{D}) \approx C_{app}(\vec{D}) := \sum_{i=1}^{N} c_i \left(\frac{D_i - 1}{2\lambda_i^{(out)}} + \mathbb{E}[W_{i,app}^{(in)}] \right), \tag{10}$$

where the first term between the brackets follows directly from (4)-(5), noting that $\sum_{j=1}^{D_i} \ j = D_i(D_i + 1)/2$, and where $\mathbb{E}[W_{i,app}^{(in)}]$ is defined in (6)-(9).

4 Optimization

In this section we discuss two approaches to the optimization problem. The first approach, which will throughout be called the *numerical approach*, is simply based on solving the non-linear optimization problem obtained by minimizing $C_{app}(\vec{D})$, defined in (10), numerically with respect to $\vec{D} \in \mathcal{D}$. To this end, standard numerical algorithms for non-linear optimization are available. Numerical results will show that this approach works very well when the number of stations N is not too large, however computation times become prohibitively large when N is large (see Section 5 for details). Furthermore, this approach may be cumbersome to implement and does not reveal explicitly how the optimal batch sizes depend on the system parameters.

To overcome these problems, we will now proceed to develop a *closed-form approximation* for $\vec{D}^{(opt)}$. Unfortunately, we observe that the functional form in the approximation for $C(\vec{D})$ defined in (6)-(10) is too complex to obtain closed-form approximations for $\vec{D}^{(opt)}$. For this reason, we will now further simplify the approximation for $\mathbb{E}[W_i^{(in)}]$ and assume that it has the following form for $i = 1, \ldots, N$, $\rho < 1$:

$$\mathbb{E}[W_{i,app}^{(in,simple)}] = \frac{a + b_i \rho}{1 - \rho}. \tag{11}$$

The coefficients a and b_i will be taken such that the approximation is asymptotically exact in the known limiting cases of light-traffic (LT) and heavy-traffic (HT), respectively. To this end, note that in LT, we have that $\lim_{\rho \downarrow 0} \ \mathbb{E}[W_i^{(in)}] = \mathbb{E}[S^{(res)}]$, so that

$$a = \mathbb{E}[S^{(res)}] = \frac{\mathbb{E}[S^2]}{2\mathbb{E}[S]}. \tag{12}$$

In the case of HT, we have (cf. [12]): for $i = 1, \ldots, N$, $\rho \uparrow 1$,

$$\mathbb{E}[W_i^{(in)}] = \frac{\omega_i}{1 - \rho} + o((1 - \rho)^{-1}), \quad \text{with} \quad \omega_i := \frac{1 - \hat{\rho}_i}{2} \left(\frac{\sigma^2}{2\delta} + \mathbb{E}[S] \right), \tag{13}$$

and where σ^2 and δ are defined in (1). Moreover, if we assume for simplicity (see Remark 4.1) that the arrival process is deterministic (so that $c^2_{A_i^{(out)}} = c^2_{A_i^{(in)}} = 0$), it is readily seen that ω_i defined in (13) is further simplified into

$$\omega_{i,app}^{(simple)} := \frac{1 - \hat{\rho}_i}{2} \left(\frac{\sigma_{app}^2}{2\delta} + \mathbb{E}[S] \right), \ \text{with} \ \sigma_{app}^2 = \sum_{i=1}^{N} \hat{\lambda}_i^{(in)} \text{Var}[B_i], \tag{14}$$

and hence,

$$a + b_i = \lim_{\rho \uparrow 1} \ (1 - \rho)\mathbb{E}[W_i^{(in)}] = \omega_{i,app}^{(simple)}. \tag{15}$$

Finally, the combination of (11), (12), (14) and (15) leads to the following *simplified approximation* for the cost function: for $\vec{D} \in \mathcal{D}$,

$$C(\vec{D}) \approx C_{app}^{(simple)}(\vec{D}) = \sum_{i=1}^{N} c_i \left(\frac{D_i - 1}{2\lambda_i^{(out)}} + \frac{\mathbb{E}[S^2]}{2\mathbb{E}[S]} + \frac{\omega_{i,app}^{(simple)} \rho}{1-\rho} \right). \tag{16}$$

We will now use (16) to derive closed-form approximations for $\vec{D}^{(opt)}$. In Section 4.1 we consider the homogeneous case in which the batch sizes are required to be the same for all queues. Subsequently, in Section 4.2 we consider the heterogeneous case in which the batch sizes are allowed to be different.

Remark 4.1: For later reference, an important observation is that $\omega_{i,app}^{(simple)}$, defined in (14), only depends on the batch sizes $\vec{D} = (D_1, \ldots, D_N)$ through their ratios D_i/D_j. In other words, if the batch-size vector $\vec{D}$ is parameterized as $\vec{D} = \alpha \vec{d}$, with $\vec{d} = (d_1, \ldots, d_N)$ the vector of *relative* batch sizes, normalized such that $\sum_{i=1}^{N} d_i = 1$, and the scalar $\alpha > 0$ a scaling parameter, then $\omega_{i,app}^{(simple)}$ is the same for all α. To this end, note that both $\lambda_i^{(in)} = \lambda_i^{(out)}/D_i = \lambda_i^{(out)}/(\alpha d_i)$ and the per-queue load-values $\rho_i = \lambda_i^{(in)}\mathbb{E}[B_i] = \lambda_i^{(out)}\mathbb{E}[B_i]/(\alpha d_i)$ do depend on α, but that this dependence *cancels out* when taking its heavy-traffic limit. Note that by definition it holds that $1 = \hat{\rho} = \sum_{i=1}^{N} \hat{\rho}_i$, which is readily seen to imply that

$$\hat{\alpha} = \sum_{j=1}^{N} \frac{\hat{\lambda}_j^{(out)}\mathbb{E}[B_j]}{d_j}, \text{ and hence } \hat{\lambda}_i^{(in)} = \frac{\hat{\lambda}_i^{(out)}}{\hat{\alpha} d_i} = \frac{\hat{\lambda}_i^{(out)}/d_i}{\sum_{j=1}^{N} \hat{\lambda}_j^{(out)}\mathbb{E}[B_j]/d_j}, \tag{17}$$

which is independent of the batch-size scaling parameter α, because all parameters $\hat{\lambda}_j^{(out)}$, $\mathbb{E}[B_j]$ and d_j are by definition independent of α. This observation, which provides a strong simplication in the functional form of $C_{app}^{(simple)}(\vec{D})$ considered as a function of $\vec{D}$ in (16), will be useful for later reference.

4.1 Homogeneous case: one-dimensional problem

In this section, we study the optimization problem under the restriction that $D_1 = D_2 = \ldots = D_N =: D$, for D a positive integer. In that case, the approximate cost function, defined in (16), reduces to, for $D = 1, 2, \ldots,$

$$C_{hom}(D) := \sum_{i=1}^{N} \frac{c_i(D-1)}{2\lambda_i^{(out)}} + \sum_{i=1}^{N} c_i \frac{\mathbb{E}[S^2]}{2\mathbb{E}[S]} + \frac{\sum_{i=1}^{N} c_i \omega_{i,app}^{(simple)} \sum_{i=1}^{N} \lambda_i^{(out)} E[B_i]}{D - \sum_{i=1}^{N} \lambda_i^{(out)}\mathbb{E}[B_i]}. \tag{18}$$

Note that Remark 4.1 readily implies that in this case $\omega_{i,app}^{(simple)}$ does not depend on D. Then, if we consider $C_{hom}(x)$ as a function of the *continuous* parameter $x > 0$, optimize with respect to x and then round off appropriately, we immediately obtain the following closed-form approximation for the optimal value of D:

$$D^{(opt)} \approx D_{app}^{(opt)} := \text{round}(x), \tag{19}$$

where

$$x := \sum_{i=1}^{N} \lambda_i^{(out)}\mathbb{E}[B_i] + \sqrt{2 \left(\sum_{i=1}^{N} \frac{c_i}{\lambda_i^{(out)}} \right)^{-1} \left(\sum_{i=1}^{N} c_i \omega_{i,app}^{(simple)} \right) \left(\sum_{i=1}^{N} \lambda_i^{(out)}\mathbb{E}[B_i] \right)} \tag{20}$$

and the round-function rounds its input to the nearest integer if the resulting batch size results in a stable system or rounds to the nearest larger integer otherwise.

Remark 4.2: Note that the approximation in (19)-(20) can easily be generalized to the case where the batch sizes are not necessarily the same, but have fixed proportions (i.e., $D_i = \alpha d_i$ with $\sum_{i=1}^{N} d_i = 1$).

4.2 Heterogeneous case: multi-dimensional problem

When we drop the restriction that the batch sizes need to be the same, the optimization problem is an N-dimensional non-linear optimization problem with integer-valued parameters $\vec{D} \in \mathcal{D}$. This type of problem generally does not scale to large problem instances, in the sense that the computation times then become prohibitively large. Therefore, in this section we propose a further simplification of the approximate cost function defined in (16) based on the following argument. Consider the parameterization of the batch sizes in Remark 4.1 (i.e., $D_i = \alpha d_i$, with $\sum_{i=1}^N d_i$), and scale the batch sizes such that $\alpha > 0$ is varied while the *relative* batch sizes $(d_1, \ldots, d_N)$ are kept fixed. Using this parameterization, the batch sizes, and hence the waiting times in both the inner and the outer system, generally depend on α. As for the outer system, for each i the batch size D_i and hence also the mean waiting times in the outer system $\mathbb{E}[W_i^{(out)}]$, characterized in (4)-(5), will increase only linearly in α. However, as for the inner system, it follows directly from the asymptotic expansion in (13) that when the load on the inner system ρ is high, $\mathbb{E}[W_i^{(in)}]$ is extremely sensitive to ρ, and hence, even a small increase in α will lead to a strong decrease in $\mathbb{E}[W_i^{(in)}]$. Hence, one may suspect that the optimal value of α, and hence of the batch sizes, will be large such that the inner system is only lightly loaded (i.e., ρ close to 0), so that the last term in (16) becomes negligible.

Using these arguments, and omitting the terms that do not depend on $\vec{D}$, the optimization problem is further simplified to the problem of finding the batch-size vector $\vec{D}^{(opt)}$ that

$$\text{minimizes} \quad \sum_{i=1}^N \frac{c_i D_i}{\lambda_i^{(out)}}, \quad \text{subject to the condition} \quad \sum_{i=1}^N \frac{\lambda_i^{(out)} \mathbb{E}[B_i]}{D_i} = \rho < 1. \tag{21}$$

This is equivalent to the problem of finding $\vec{\beta} = (\beta_1, \ldots, \beta_N)$ that

$$\text{minimizes} \quad \sum_{i=1}^N \frac{\gamma_i}{\beta_i}, \quad \text{subject to} \quad \sum_{i=1}^N \beta_i < 1, \quad \text{with } \gamma_i := c_i \mathbb{E}[B_i] \text{ and } \beta_i := \frac{\lambda_i^{(out)} \mathbb{E}[B_i]}{D_i}, \tag{22}$$

where γ_i $(i = 1, \ldots, N)$ are constants, and where β_i $(i = 1, \ldots, N)$, via D_i defined above, are decision variables. If we take $\vec{\beta}$ to be continuous, a standard Lagrangian-multiplier approach leads to the solution $\beta_i^{(opt)} = \sqrt{\gamma_i} \left(\sum_{i=1}^N \sqrt{\gamma_i} \right)^{-1}$ for $i = 1, \ldots, N$. Translating this back to batch-size ratios, we have

$$\frac{D_i^{(opt)}}{D_j^{(opt)}} \approx \frac{\lambda_i^{(out)} \mathbb{E}[B_i] / \beta_i^{(opt)}}{\lambda_j^{(out)} \mathbb{E}[B_j] / \beta_j^{(opt)}} = \frac{\lambda_i^{(out)}}{\lambda_j^{(out)}} \sqrt{\frac{c_j \mathbb{E}[B_i]}{c_i \mathbb{E}[B_j]}}. \tag{23}$$

It should be noted that replacing the constraint in (22) with $\sum_{i=1}^N \beta_i = c$ for any value of $c \in (0, 1]$ would have lead to (23), thus also when demanding that ρ equals a value close to 0 in accordance with the arguments above.

Now that the batch-size ratios have been approximated in (23), it remains to approximate the absolute values of these batch sizes via the scalar α. Therefore, we go back to an arbitrarily loaded system with a cost function as defined in (16) and we write $\vec{D} = \alpha \vec{d}$. As stated in Remark 4.2 the problem of determining the (near-)optimal value of α can be obtained via the same argumentation as in the homogeneous case, which leads to the following result: for $i = 1, \ldots, N$,

$$D_i^{(opt)} \approx D_{i,app}^{(opt)} := \text{round} \left(\alpha_{app}^{(opt)} d_{i,app}^{(opt)} \right), \tag{24}$$

with

$$\alpha_{app}^{(opt)} := \sum_{i=1}^N \frac{\lambda_i^{(out)} \mathbb{E}[B_i]}{d_{i,app}^{(opt)}} + \sqrt{2 \left(\sum_{i=1}^N \frac{c_i d_{i,app}^{(opt)}}{\lambda_i^{(out)}} \right)^{-1} \left(\sum_{i=1}^N c_i \omega_{i,app}^{(simple)} \right) \left(\sum_{i=1}^N \frac{\lambda_i^{(out)} \mathbb{E}[B_i]}{d_{i,app}^{(opt)}} \right)}, \tag{25}$$

$$d_{i,app}^{(opt)} := \frac{\lambda_i^{(out)} \sqrt{\mathbb{E}[B_i]/c_i}}{\sum_{j=1}^N \lambda_j^{(out)} \sqrt{\mathbb{E}[B_j]/c_j}}. \tag{26}$$

Variable	Parameter	Parameter values
-	Symmetry	{Symmetric, Asymmetric}
N	Number of queues	$\{2, 5\}$
$\overline{\lambda}^{(out)}$	Type-averaged arrival rate	$\{1/(2N), 2/N\}$
$\mathbb{E}[\overline{B}]$	Type-averaged mean service time	$\{1\}$
$\mathbb{E}[\overline{S}]$	Type-averaged mean switch-over time	$\{0, 0.2, 1, 10\}$
$c^2_{A_i^{(out)}}$	SCV external 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	$\{(\lambda_1^{(out)}, \lambda_2^{(out)}), (1,1), (2,1)\}$ for $N=2$ or $\{(\lambda_1^{(out)}, \lambda_2^{(out)}, \lambda_3^{(out)}, \lambda_4^{(out)}, \lambda_5^{(out)})$ $(1,1,1,1,1), (5,4,3,2,1)\}$ for $N=5$

Table 1: Parameter values of the simulation testbed.

and the round-function as in the homogeneous case. The obtained approximation in (24)-(26) requires little or no computation time, even for systems with a large number of queues. In the next section, the accuracy of the approximations is evaluated.

5 Validation

In this section, we evaluate the performance of the approximation methods presented in Section 4 using a large simulation testbed.

Description of the simulation testbed
To validate the results for a wide variety of parameter combinations, experiments were done on a testbed containing 1260 parameter combinations, such that the parameter space is well covered. To structure the instances, let

$$\overline{\lambda}^{(out)} := \frac{1}{N}\sum_{i=1}^{N}\lambda_i^{(out)}, \quad \mathbb{E}[\overline{B}] := \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[B_i] \text{ and } \mathbb{E}[\overline{S}] := \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[S_i] \tag{27}$$

be the *type-averaged* arrival rate, mean service time and mean switch-over times, respectively. The parameters of the 1260 instances are then obtained by taking every combination of the parameter values found in Table 1. For every instance, each queue shares the same value for $c^2_{A_i^{(out)}}$, $c^2_{B_i}$ and $c^2_{S_i}$. For symmetric instances, we of course also have that $\lambda_i^{(out)} = \overline{\lambda}^{(out)}, \mathbb{E}[B_i] = \mathbb{E}[\overline{B}]$ and $\mathbb{E}[S_i] = \mathbb{E}[\overline{S}]$ for $i = 1, \ldots, N$. For the asymmetric instances however, the parameters are taken asymmetrically through the formulas

$$\lambda_i^{(out)} := \frac{2i}{N+1}\overline{\lambda}^{(out)}, \quad \mathbb{E}[B_i] := \frac{2i}{N+1}\mathbb{E}[\overline{B}] \text{ and } \mathbb{E}[S_i] := \frac{2(N+1-i)}{N+1}\mathbb{E}[\overline{S}], \tag{28}$$

for $i = 1, \ldots, N$. For the sake of validation of the numerical approach, an implementation of a Newton type-algorithm for unconstrained minimization was used (see [16] for details).

Relative errors
To quantify the accuracy of the approximations, the relative error in the cost function is defined as follows:

$$\Delta\% := \frac{C(\vec{D}_{app}) - C(\vec{D}^{(opt)})}{C(\vec{D}^{(opt)})} \times 100\%, \tag{29}$$

where $C(\vec{D}^{(opt)})$ is the real cost of the optimal batch-size vector $\vec{D}^{(opt)}$ (both obtained by simulation), and $C(\vec{D}_{app})$ is the real cost (obtained by simulation) belonging to the batch-size *approximation* methods.

Experimental results
Tables 2 and 3 show the averages of the relative cost differences $\Delta\%$ of both the *numerical approach* and the results based on the *closed-form approximation*, as well as the distribution of these relative errors over several bins. The results show that both solution techniques perform very well, with average errors up to

	Numerical approach					
	Average	Differences categorized in bins				
Testbed	difference	0%	>0-2%	2-5%	5-20%	>20%
Symmetric	0.30%	82.1%	12.5%	2.98%	2.38%	0.00%
Asymmetric	0.65%	74.5%	19.3%	2.78%	2.91%	0.53%

Table 2: Averages of the the values of $\Delta\%$ and their distribution.

	Closed-form approximation					
	Average	Differences categorized in bins				
Testbed	difference	0%	>0-2%	2-5%	5-20%	>20%
Symmetric	1.62%	73.2%	14.7%	4.17%	5.75%	2.18%
Asymmetric	4.85%	59.2%	25.1%	4.37%	4.63%	6.75%

Table 3: Averages of the the values of $\Delta\%$ and their distribution.

only a few percent. The majority of the differences is concentrated in the bin which represents errors of 0%; that is, in most cases the approximations for $\vec{D}$ lead to the correct optimum $\vec{D}^{(opt)}$.

Table 3 shows that the approximations work well for the symmetric model instances, but that the accuracy of the closed-form approximation tends to degrade for asymmetric model instances.
Next, we evaluate the accuracy of the approximations, where the individual input parameters are varied. Table 4 shows the average relative differences $\Delta\%$ in cost performances categorized in the approximation methods (i.e, *numerical* and *closed-form*), for the type-averaged arrival rate (a), the number of queues (b), the SCV of the switch-over times (c), the type-averaged mean switch-over time (d), the weight vector (e), the SCV of the interarrival times (f), and the SCV of the service times (g).

The results in Table 4 lead to a number of observations. Generally, whenever total mean waiting times become longer (for example, in the case when $\mathbb{E}[\overline{S}] = 10$, or $N = 5$), both the numerical and closed-form approximation become increasingly accurate. In case waiting times are generally short (for example when $\mathbb{E}[\overline{S}] = 0$ or $\overline{\lambda}^{(out)} = 1/(2N)$), the batch-size vector found by the numerical approach and the closed-form approximation usually coincide. However, whenever a difference in the obtained batch-size vector does occur, the difference in terms of cost of that particular system may be considerable. However, it should be noted that although large *relative* differences may occur in these cases, the *absolute* differences are still quite small, due to the fact that the waiting times themselves are small. When considering the particular role of the value of N, a similar effect is observed. We see that when $N = 2$, the numerical approach is slightly more accurate than the closed-form approximation, but when $N = 5$, both approximation methods tend to work well, although the numerical approach still marginally outperforms the closed-form approximation. In this context, recall that when N gets large, the computation times for the numerical approach inevitably become prohibitively large, so that the closed-form approximation is preferred. Table 4(f) and 4(g) show that both methods are quite resistant against variability in the interarrival times and the service times.

Remark 5.1 (Near-insensitivity to the higher moments of the interarrival and switch-over time distributions)
The closed-form approximation for $\vec{D}^{(opt)}$ in (24)-(26) is insensitive to the second and higher moments

(a)

	$\overline{\lambda}^{(out)} = 1/2N$		$\overline{\lambda}^{(out)} = 2/N$	
Testbed	Numerical	Closed-form	Numerical	Closed-form
Symmetric	0.520%	1.013%	0.086%	2.234%
Asymmetric	0.696%	4.543%	0.594%	5.153%

(b)

	$N = 2$		$N = 5$	
Testbed	Numerical	Closed-form	Numerical	Closed-form
Symmetric	0.428%	2.776%	0.178%	0.471%
Asymmetric	0.870%	7.715%	0.419%	1.981%

(c)

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

(d)

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

	$\mathbb{E}[\overline{S}] = 1$		$\mathbb{E}[\overline{S}] = 10$	
Testbed	Numerical	Closed-form	Numerical	Closed-form
Symmetric	0.208%	0.766%	0.011%	1.545%
Asymmetric	0.221%	0.000%	0.000%	0.026%

(e)

	$\vec{c} = (\lambda_1^{(out)}, \lambda_2^{(out)})$ or $(\lambda_1^{(out)}, \ldots, \lambda_5^{(out)})$		$\vec{c}$ = (1,1) or (1,1,1,1,1)		$\vec{c}$ = (2,1) or (5,4,3,2,1)	
Testbed	Numerical	Closed-form	Numerical	Closed-form	Numerical	Closed-form
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%

(f)

	$c^2_{A_i^{(out)}} = 0.25$		$c^2_{A_i^{(out)}} = 1$		$c^2_{A_i^{(out)}} = 2$	
Testbed	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%

(g)

	$c^2_{B_i} = 0$		$c^2_{B_i} = 1$		$c^2_{B_i} = 4$	
Testbed	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 4: Average relative difference of the cost $\Delta\%$ of the approximation methods over categorized subsets of the testbed.

of the interarrival-time and switch-over time distributions. To investigate the sensitivity of the optimal batch sizes, consider a symmetric five-queue system with, for all $i = 1, \ldots, 5$, $\lambda_i^{(out)} = 2$, $c_i = 1$, $\mathbb{E}[B_i] = \mathbb{E}[S_i] = 1$, $\mathbb{E}[B_i^2] = 2$ and $c^2_{A_i^{(out)}} = c^2_{S_i} =: \gamma$. Figure 2 shows the mean optimal 'exact' batch size as a function of γ, where $\gamma \in [0, 40]$. Apart from some noise, which especially for large gamma is due to simulation error, the figure shows a near-constant line, which supports the near-insensitivity property of $\vec{D}^{(opt)}$ with respect to the second and higher moments of the interarrival and switch-over time distributions.

Remark 5.2 (Near-insensitivity of ratios to other types or queues)
Equations (23)-(26) indicate that the ratios between the optimal batch sizes of an arbitrary couple of queues (i.e., D_i/D_j) are (fully) insensitive to the parameters of the *other* queues $k \neq i, j$. This suggests near-

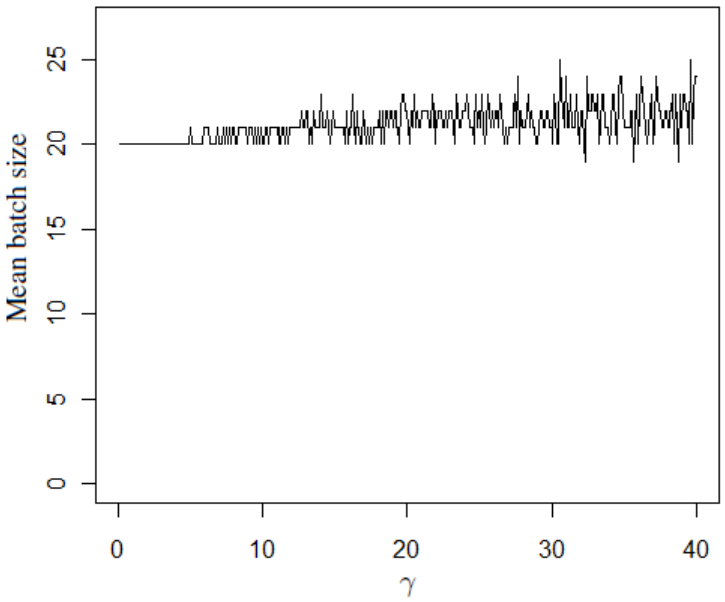

Figure 2: Mean batch size as a function of the SCV of the interarrival and the switch-over time distributions.

insensitivity of the ratios of optimal batch sizes with respect to the parameters of the other queues. To investigate this, consider a three-queue system with Poisson arrivals and exponential switch-over times, and with $(\lambda_1^{(out)}, \lambda_2^{(out)}, \lambda_3^{(out)}) = (4, 2, k)$, $(\mathbb{E}[B_1], \mathbb{E}[B_2], \mathbb{E}[B_3]) = (1, 0.5, k)$, $(\mathbb{E}[B_1^2], \mathbb{E}[B_2^2], \mathbb{E}[B_3^2]) = (2, 0.25, k^2)$, and where all switch-over times have a mean duration 1. Figure 3 shows the ratio of $D_2^{(opt)}/D_1^{(opt)}$, for $k \in [1, 3]$. Note that varying the value of k means that the parameters for queue 3 are varied, while the parameters for queues 1 and 2 are kept fixed. Moreover, it is readily verified from (23) that the approximated ratio D_2/D_1 equals 1/2, independent of k. The results in Figure 3 show that $D_2^{(opt)}/D_1^{(opt)}$ is indeed nearly-insensitive to the value of k, which underlines the fact that the ratio of two optimal batch sizes is nearly-insensitive to the characteristics of other queues.

6 Conclusions and Further research

In this paper we have studied a batch polling system as a model for a multiple product single production capacity with batching constraints. The proper use of the batching mechanisms is not understood fully and rules for implementing optimal batch values are not available in the open literature. This would almost suggest that the scheduling problem does not exist in practice; yet the multiple product single machine system is quite common in industry and batching of demand is done frequently to avoid small production runs and loss of capacity due to set-ups. Our objective has, therefore, been the derivation of these batch sizes so as to minimize a weighted sum of the mean waiting times and, thus, the mean queue lengths. In the absence of exact analysis, we present two novel approaches to approximate the optimal batch sizes, i.e., a numerical approach and a closed-form approximation. The numerical approach is in general the most accurate, but the computation times become prohibitively large when the number of queues in the system increases. Furthermore, it acts as a kind of black box and can, therefore, contribute to the understanding of the system behavior only to a limited extent. It is, for instance, rather difficult to study the impact of parameters like the occupation rates on the optimal batch sizes. The closed-form solution is very simple and, therefore, loses some of the accuracy but still captures the major factors important for efficient operation. It can be concluded that it complements the numerical approach explicitly showing the impact of all system parameters and pointing out near-insensitivity properties.

The research presented in this paper suggest a variety of directions for further research. First, in this paper we focus on minimization of a weighted sum of the mean waiting times. An interesting extension

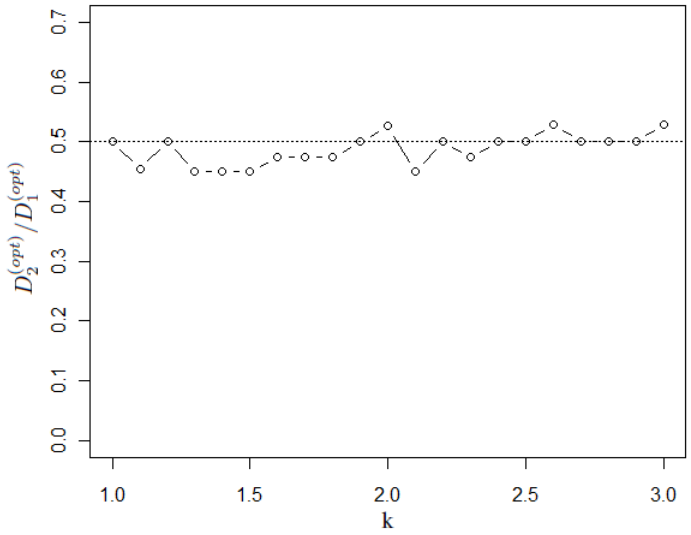

Figure 3: Ratio $\frac{D_2^{(opt)}}{D_1^{(opt)}}$ as a function of k.

from an application point of view would be to add restrictions on for example tail probabilities of the delay figures. A good starting point may be the recently-proposed approximations for the tail probabilities in [4]. Second, the assumptions of exhaustive services at all queues may be relaxed to include more general branching-type policies [15], such as single-phase or multi-phase gated service [11]. Following the lines of the present paper, one would expect that also for these cases simple closed-form approximations can be determined. Third, the results may be extended to the case of periodic, non-cyclic, server routing, taking the results in [13] as the starting point. Fourth, from an application point of view it would be interesting to include some level of dependence between the batch size and its service time. To this end, note that the extreme case of linear dependence corresponds to the classical batch-polling system analyzed in [10]. Finally, an interesting extension is one where the switch-over times depend on the actual state of the system. For example in production systems it may be interesting to let the server skip a queue that is empty, or let the server reside in idle mode when there are no jobs in the system. The approximation techniques in [1] can readily be extended to the case of state-dependent switch-over times, and may be used to extend the current results to this situation as well.

Acknowledgement: The authors wish to thank Marko Boon for placing parts of his polling system simulation program at their disposal, and for useful comments on earlier drafts of this paper.

References

[1] M.A.A. Boon, E.M.M. Winands, I.J.B.F. Adan and A.C.C. van Wijk (2011). Closed-form waiting time approximations for polling systems. Performance Evaluation 68, 290-306.

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

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

[4] J.L. Dorsman, R.D. van der Mei and E.M.M. Winands (2011). A new method for deriving waiting-time approximations in polling systems with renewal arrivals. Stochastic Models 27, 318-332.

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

[6] 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.

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

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

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

[10] R.D. van der Mei (2002). Waiting-time distributions in polling systems with simultaneous batch arrivals. Annals of Operations Research 113, 157-173.

[11] R.D. van der Mei and A. Roubos (2011). Polling models with multi-phase gated service. To appear in Annals of Operations Research.

[12] 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.

[13] 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.

[14] 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.

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

[16] 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.

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

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

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

[20] 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.

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

[22] E.M.M. Winands (2011). Branching-type polling systems with large setups. OR Spectrum 33, 77-97.

[23] E.M.M. Winands, I.J.B.F. Adan and G.J. van Houtum (2011). The stochastic economic lot scheduling problem: a survey. European Journal of Operational Research, vol. 210, 1-9.

[24] E.M.M. Winands, A.G. de Kok and C. Timpe (2009). Case study of a batch-production and inventory system. Interfaces 39, 552-554.

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