Branching-type polling systems with large setups

Citation for published version (APA):
Winands, E. M. M. (2006). Branching-type polling systems with large setups. (SPOR-Report : reports in statistics, probability and operations research; Vol. 200616). Technische Universiteit Eindhoven.

Document status and date:
Published: 01/01/2006

Document Version:
Publisher’s PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:
• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain.
• You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:
www.tue.nl/taverne

Take down policy
If you believe that this document breaches copyright please contact us at:
openaccess@tue.nl
providing details and we will investigate your claim.

Download date: 01. Nov. 2023
Branching-type polling systems with large setups

E.M.M. Winands\textsuperscript{1,2}

\textsuperscript{1}Department of Mathematics and Computer Science
\textsuperscript{2}Department of Technology Management
Technische Universiteit Eindhoven
P.O. Box 513, 5600 MB Eindhoven, The Netherlands
e.m.m.winands@tue.nl

December 6, 2006

Abstract
The present paper considers the class of polling systems that allow a multi-type branching process interpretation. This class contains the classical exhaustive and gated policies as special cases. We present an exact asymptotic analysis of the delay distribution in such systems, when the setup times tend to infinity. The motivation to study these setup time asymptotics in polling systems is based on the specific application area of base-stock policies in inventory control. Our analysis provides new and more general insights into the behavior of polling systems with large setup times.

Keywords: polling systems, multi-type branching processes, setup times, delay distribution, asymptotics.
1 Introduction

A typical polling system consists of a number of queues, attended by a single server in a fixed order. There is a huge body of literature on polling systems that has continued to grow since the late 1950s, when the papers of [12, 13] concerning a patrolling repairman model for the British cotton industry were published. Polling systems have a wide range of applications in communication, production, transportation and maintenance systems. Excellent surveys on polling systems and their applications may be found in [20, 21, 22] and in [9, 24]. One of the most remarkable results in the polling literature is the striking dichotomy in complexity between different polling systems independently illuminated by [6] and [19]. That is, if the service discipline satisfies a certain branching property, the polling system allows for an exact analysis by rather standard methods. If this branching property is, however, violated, the corresponding polling systems cannot be analyzed exactly in the general setting.

Unfortunately, even for branching-type polling systems the interdependence of the queueing processes prohibits an exact explicit analysis, leading to the need of using numerical techniques to determine performance measures of interest. However, such numerical techniques provide only limited insight into the behavior of the system with respect to its input parameters. In these circumstances, one naturally resorts to asymptotic estimates. In particular, the present paper presents an exact asymptotic analysis of the delay distribution in polling systems with general branching-type service discipline - with the classical exhaustive and gated policies as special cases - when the setup times tend to infinity. Since the delay obviously grows without bound in such a case, we focus on the scaled delay, i.e., the delay divided by the total setup time per cycle.

The present study is both relevant from a theoretical and practical point of view. From a theoretical point of view, such an analysis is evidently interesting, since it deepens the understanding of the behavior of systems with large setups. That is, we obtain explicit expressions for the scaled delay distribution, which lead to significant insight into the dependence of the performance measures on the system parameters (e.g., insensitivity and monotonicity properties). From a practical point of view, polling systems with large setup times find a wide variety of applications in production environments. For example, in the stochastic economic lot scheduling problem (SELS), where multiple standardized products have to be produced on a single machine with significant setup times, polling system are frequently encountered as modeling tool for (widely-used) fixed-sequence base-stock policies (details are given in Section 4). We refer to [26] for a survey on the SELSP and for a large number of cases of production environments with large setup times. With respect to the distribution of the setup times, it is important to remark that in production environments setup times are typically deterministic due to the nowadays efficient control of production processes.

Although the number of papers on polling systems is impressive, only a few papers address the problem of large (deterministic) setup times. [15] explores the descendant set approach in combination with the strong law of large numbers for renewal reward processes to analyze polling systems with deterministic setups and mixtures of exhaustive and gated service. [17] presents a somewhat simpler analysis in the case of an exhaustive system with deterministic setups, where the order of service is determined by a polling table. Based on the recently proposed mean value analysis (MVA) for polling systems [27], in [28] we develop an alternative simple approach for cyclic exhaustive polling systems. Since MVA is not limited to exhaustive polling systems, the analysis of the latter can be readily extended to a wide range of polling systems. The main result in all of the above papers [15, 17, 28] is the fact that the scaled delay converges in distribution to a uniform distribution. It is, however, worth remarking that all previous studies deal only with the exhaustive and gated service discipline, whereas the present paper deals with the general class of branching-type service policies.

Before leaving the literature review on polling systems, we should also mention the recent paper [16]. This paper obtains heavy-traffic asymptotics for branching-type polling systems. Although both the asymptotic behavior in such a heavy-traffic regime and the methodology used are fundamentally different from the ones in the present paper, there are some aspects in both asymptotic regimes that bear a resemblance. These similarities are touched upon throughout the present paper. Furthermore, it is remarkable that since the discovery of the strong connection between
polling systems and multi-type branching process by [6, 19], the present paper and [16] are the first studies actually deriving (asymptotic) results for the complete class of branching-type polling systems.

To obtain a unifying theorem for setup time asymptotics in branching-type polling systems, we rely on results from [1]. In this paper, a strong relation between the queue length, as well as delay, distributions in models with and without setup times is exposed. In particular, using results for the transform of the marginal distributions we analytically show that the scaled delay converges in distribution to a uniform distribution for all policies allowing a multi-type branching process interpretation.

The contribution of the present paper is three-fold. First of all, for the large class of polling systems that allow a multi-type branching process interpretation we derive setup time asymptotics, thus exposing the general structure as well as the limitations of the theory. The results of the present paper not only generalize those derived in [15, 17, 28] for the special case of exhaustive and gated service, but are also obtained via a fundamentally different approach. In fact, the present paper focuses on two types of limit theorems for (i) polling systems with increasing deterministic setup times; (ii) polling systems with increasing stochastic setup times under heavy traffic. We stress that the methodology of the present paper has a wide range of applications; it can, for example, also be used for the asymptotic analysis of the delay distribution for cyclic polling systems with the globally-gated service policy, which does not satisfy the branching property, as shown in Section 4.

Secondly, the results obtained in the present paper provide new (and more general) insights into the behavior of polling systems with large setup times. It is shown that the asymptotic scaled delay distribution (i) is independent of the visit order; (ii) depends on the service discipline of the corresponding queue only through a single parameter referred to as the exhaustiveness; (iii) is independent of the service disciplines of the other queues; (iv) depends on the arrival rate and service time distribution of the corresponding queue only through its occupation rate; (v) depends on the arrival rate and service time distribution of the other queues only through the total occupation rate. In this context, it is important to remark that in the heavy-traffic regime studied in [16] the exhaustiveness of the service discipline plays a key - but different - role as well.

Finally, the obtained expressions for the asymptotic scaled delay can be readily used as approximation of the delay distribution in systems with finite setup times. Thanks to the simplicity of the derived expressions, these approximations are easily implementable and allow for back-of-the-envelope calculations. Furthermore, as mentioned above we envision production in general - and fixed-sequence base-stock policies in particular - as the main application of interest for the present paper. Therefore, our results deepen the understanding of the behavior and performance of base-stock policies in production environments with significant setup times.

The structure of the present paper is as follows. In Section 2, we introduce the model and summarize notation. Section 3 analyzes the scaled delay distribution for polling models with general branching-type service policies; firstly, for systems with deterministic setup times and, secondly, for systems with stochastic setup times under heavy traffic. In Section 4, we revisit the results of the present paper and we address a number of challenging topics for further research.

2 Model description and notation

We consider a system with a single server for $N \geq 1$ queues, in which there is infinite buffer capacity for each queue. The server visits and serves the queues in a fixed cyclic order. We index the queues by $i$, $i = 1, 2, \ldots, N$, in the order of the server movement. For compactness of presentation, all references to queue indices greater than $N$ or less than 1 are implicitly assumed to be modulo $N$, e.g., queue $N + 1$ actually refers to queue 1. Throughout the present paper, it is assumed that within a queue customers are served First Come First Served (FCFS). Obviously, the mean waiting times are the same under any work-conserving non-preemptive service discipline that does not account for the actual service requests of the customers.

Customers arrive at all queues according to independent Poisson processes with rates $\lambda_i$, $i =$
1, 2, . . . , N. The service times at queue $i$ are independent, identically distributed random variables with mean $E[B_i]$ and Laplace Stieltjes Transform (LST) $\beta_i(\cdot)$, $i = 1, 2, \ldots, N$. When the server starts service at queue $i$, a deterministic setup time $S_i$ is incurred, $i = 1, 2, \ldots, N$. The total setup time $S$ in a cycle is given by

$$S = \sum_{i=1}^{N} S_i. \tag{1}$$

It is assumed that a setup is incurred even if the subsequent queue is empty. Since we study the system as setup times tend to infinity, this last assumption is irrelevant since queues will never be empty when polled in this situation.

Throughout the present paper, we assume that the service discipline at each queue satisfies the following property [6, 19]:

**Property 2.1.** If the server arrives at queue $i$ to find $k_i$ customers there, then during the course of the server’s visit, each of these $k_i$ 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. $\square$

It is important to remark that we allow different service disciplines at different queues. The present paper focuses on nonidling service disciplines satisfying Property 2.1. The occupation rate $\rho_i$ (excluding setups) at queue $i$ is defined by $\rho_i = \lambda_i E[B_i]$ and the total occupation rate $\rho$ is given by

$$\rho = \sum_{i=1}^{N} \rho_i. \tag{2}$$

Now, $\rho < 1$ and $S < \infty$ constitute necessary and sufficient stability conditions for any nonidling policy that satisfies Property 2.1 with $h_i(z_1, \ldots, z_N) \neq z_i$ (see, e.g., [19]). In the remainder of the present paper, these stability conditions are assumed to hold as we restrict the attention to steady-state behavior.

The performance measure of interest is the delay $W_i$ of a type-$i$ customer, $i = 1, 2, \ldots, N$, in case the setup times tend to infinity. Since the delay grows to infinity in the limiting case, we focus on the asymptotic scaled delay $W_i^S$ as $S \to \infty$, where the ratios of the setup times remain constant, $i = 1, 2, \ldots, N$. Of course, our results for the delay distribution can be readily translated into results for the queue length distribution via the distributional form of Little’s law [8].

## 3 Limit theorems

The present section presents the main results of the paper and is divided into three subsections. After discussing some preliminary results (Subsection 3.1), we focus on polling systems with deterministic setup times (Subsection 3.2) and polling systems with stochastic setup times under heavy traffic (Subsection 3.3).

### 3.1 Preliminaries

Throughout the present subsection we discuss some basic results of branching-type polling systems. The reader is referred to [1] and [19] for more details. First of all, the partial derivative $\frac{\partial}{\partial z_i} h_i(z)|_{z=1}$ of the generating function $h_i(z)$ as introduced in Property 2.1 represents the mean number of type-$i$ children residing in queue $i$ at the end of a visit period generated by a type-$i$ customer present at the start of a visit to queue $i$ (for a formal definition of children, see Subsection 3.2). Subsequently, we define the exhaustiveness $\Phi_i$ of the service discipline at queue $i$ by

$$\Phi_i = 1 - \frac{\partial}{\partial z_i} h_i(z)|_{z=1}, \quad i = 1, 2, \ldots, N. \tag{3}$$
Since we have assumed that \( h_i(z_1, \ldots, z_N) \neq z_i \), we have

\[
0 < \Phi_i \leq 1, \quad i = 1, 2, \ldots, N.
\] (4)

The exhaustiveness \( \Phi_i \) has the following intuitively appealing interpretation: each customer present at the start of a visit to queue \( i \) will be replaced by a number of type-\( i \) customers with mean \( 1 - \Phi_i \).

It is convenient to have an expression for \( \mathbb{E}[T_i] \), i.e., the mean subvisit period generated by a type-\( i \) customer present at the start of a visit to queue \( i \), in terms of \( \Phi_i \). Since \( \frac{d}{dz} h_i(z) \big|_{z=1} \) equals 1 plus the expected number of type-\( i \) arrivals \( \lambda_i \mathbb{E}[T_i] \) minus \( \mathbb{E}[T_i] / \mathbb{E}[B_i] \), which is the expected number of type-\( i \) served during this subvisit period, we can derive, after some rewriting, the following expression

\[
\mathbb{E}[T_i] = \frac{\Phi_i \mathbb{E}[B_i]}{1 - \rho_i}, \quad i = 1, 2, \ldots, N.
\] (5)

Furthermore, (5) also leads to an expression for \( \mathbb{E}[X_i] \), i.e., the number of type-\( i \) customers present at the start of a visit to queue \( i \), by observing that the mean total visit period \( \mathbb{E}[V_i] \) at queue \( i \) which equals the sum of all subvisit periods, is the product of \( \mathbb{E}[X_i] \) and \( \mathbb{E}[T_i] \). That is,

\[
\mathbb{E}[X_i] = \frac{\mathbb{E}[V_i]}{\mathbb{E}[T_i]} = \rho_i \frac{S}{1 - \rho} \frac{\mathbb{E}[T_i]}{\mathbb{E}[T_i]} = \frac{\lambda_i}{\Phi_i} \frac{1 - \rho_i}{1 - \rho} S, \quad i = 1, 2, \ldots, N.
\] (6)

We continue this section with an example.

**Example 3.1.** A variety of service disciplines satisfy the Property 2.1 including a number of classical ones as discussed below.

1. In case of the **exhaustive** discipline, i.e., a queue must be empty before the server moves on, we have

\[
h_i(z_1, \ldots, z_N) = \theta_i \left( \sum_{j \neq i} \lambda_j (1 - z_j) \right),
\] (7)

where \( \theta_i(\cdot) \) denotes the LST of a busy period in an \( M/G/1 \) queue with arrival rate \( \lambda_i \) and LST of the service time distribution \( \beta_i(\cdot) \). The corresponding exhaustiveness reads \( \Phi_i = 1 \).

2. When the **gated** discipline is implemented, i.e., only those customers in the queue at the polling instant are served, the function \( h_i(z_1, \ldots, z_N) \) reads

\[
h_i(z_1, \ldots, z_N) = \beta_i \left( \sum_{j=1}^{N} \lambda_j (1 - z_j) \right),
\] (8)

with exhaustiveness \( \Phi_i = 1 - \rho_i \).

3. Under the **binomial-exhaustive** discipline [10] each of the type-\( i \) customers present at the start of a visit to this queue generates an \( M/G/1 \) busy period with probability \( 0 \leq q_i \leq 1 \),

\[
h_i(z_1, \ldots, z_N) = q_i \theta_i \left( \sum_{j \neq i} \lambda_j (1 - z_j) \right) + (1 - q_i) z_i,
\] (9)

from which the exhaustiveness is found to be \( \Phi_i = q_i \).

4. In case of the **binomial-gated** discipline [11], where each of the type-\( i \) customers present at the start of a visit to this queue is served with probability \( 0 \leq p_i \leq 1 \),

\[
h_i(z_1, \ldots, z_N) = p_i \beta_i \left( \sum_{j=i}^{N} \lambda_j (1 - z_j) \right) + (1 - p_i) z_i,
\] (10)

and, thus, we have for the exhaustiveness \( \Phi_i = p_i (1 - \rho_i) \).
Next, define the offspring generating function as follows

\[ f(z) := (f_1(z), \ldots, f_N(z)) \]

with for \(|z_j| \leq 1, j = 1, 2, \ldots, N,\)

\[ f_i(z) := h_i(z_1, \ldots, z_i, f_{i+1}(z), \ldots, f_N(z)), \quad i = 1, 2, \ldots, N. \] (12)

This offspring generating function represents the generating function of the joint distribution of the numbers of customers at the end of a cycle with respect to queue 1 that are children of a type-\(i\) customer, where a child of a customer is recursively defined as a customer that has arrived during the service time of this customer or of one of his children. Furthermore, define for \(|z_j| \leq 1, j = 1, 2, \ldots, N,\)

\[ f^{(0)}(z) := z, \]

\[ f^{(k)}(z) := f(f^{(k-1)}(z)), \quad k \geq 1, \] (13)

where \(f^{(k)}(\cdot)\) represents the \(k^{th}\) generation offspring.

Since we are interested in the marginal waiting time distribution, we focus - without loss of generality - on \(W_1\) and introduce for \(i = 1, 2, \ldots, N,\)

\[ \hat{h}_i(y) := h_i(y, 1, \ldots, 1), \]

\[ \tilde{f}_i^{(k)}(y) := f_i^{(k)}(y, 1, \ldots, 1), \] (15)

and

\[ \hat{H}(y) := \sum_{k=0}^{\infty} \sum_{i=1}^{N} \lambda_i (1 - \tilde{f}_i^{(k)}(y)). \] (17)

Since the mean number of type-1 customers present at the start of a visit to queue 1 is exactly equal to the average offspring of customers which arrived during a setup time, we have the following relation between \(\hat{H}'(1)\) and \(E[X_1],\)

\[ E[X_1] = -S \hat{H}'(1), \] (18)

and, thus, by applying (6),

\[ \hat{H}'(1) = \frac{\lambda_1}{\Phi_1} \frac{1 - \rho_1}{1 - \rho}. \] (19)

### 3.2 Limit theorems for deterministic setups

In order to derive the limit theorems, we make extensively use of a result of [1] which derives a strong relation between the waiting time distributions in models with and without setup times. This relation is established both by relating the similarities in the offspring generating functions of the underlying branching processes and by expressing the differences between the underlying immigration functions.

In particular, [1] shows that the LST of the waiting time distribution of a type-1 customer is given by

\[ E[e^{-\omega W_{1}^0}] = E[e^{-\omega W_{1}^0}] e^{-S \hat{H}'(h_1(1 - \omega / \lambda_1) - \hat{H}(h_1(1 - \omega / \lambda_1)))} \left[ \hat{H}(1 - \omega / \lambda_1) - \hat{H}(1) \right] \] (20)

where \(W_{1}^0\) is the waiting time in the corresponding polling system with zero setup times. At this point, we feel it is worth reminding the reader that no closed-form expression for \(E[e^{-\omega W_{1}^0}]\) is known. The decomposition as expressed in (20) can be used to derive an explicit expression for the LST of the distribution of the asymptotic scaled delay as presented in the lemma below.
Lemma 3.2. In case of deterministic setup times, the LST of the distribution of the asymptotic scaled delay is given by

$$
E[e^{-\omega \frac{W_1}{S}}] \to \frac{1 - \rho}{(1 - \rho_1)\omega} \left( e^{-\frac{1 - \rho_1}{\Phi_1}} - e^{-\frac{1 - \rho_1}{1 - \rho}} \right), \quad (S \to \infty).
$$

(21)

**Proof.** First of all, the term $E[e^{-\omega W_1^0}]$ in (20) does not depend on $S$ implying that

$$
E[e^{-\omega \frac{W_1}{S}}] \to 1, \quad (S \to \infty).
$$

(22)

Next, we observe that

$$
S \tilde{H}(1 - \omega \lambda_1 S) = -\frac{\omega}{\lambda_1} \tilde{H}(1 - \frac{\omega}{\lambda_1 S}) - \frac{\omega}{\lambda_1 S} \frac{d}{S \to \infty}, -\frac{\omega}{\lambda_1} \tilde{H}'(1) = \frac{1 - \rho_1}{\Phi_1} \frac{1 - \rho}{1 - \rho_1} \omega,
$$

(23)

where the first follows from the fact that $\tilde{H}(1) = 0$ and the last equation from (19). Similarly, we have that

$$
S \tilde{H}(\tilde{h}_1(1 - \omega / \lambda_1)) = -\frac{\omega}{\lambda_1} \tilde{H}(\tilde{h}_1(1 - \frac{\omega}{\lambda_1 S})) - \frac{\omega}{\lambda_1 S} \frac{d}{S \to \infty}, -\frac{\omega}{\lambda_1} \tilde{H}'(1) \tilde{h}_1'(1) = \frac{1 - \Phi_1}{\Phi_1} \frac{1 - \rho_1}{1 - \rho} \omega,
$$

(24)

where the last equality follows from the definition of the exhaustiveness factor and (19). Substituting (23) and (24) into (20) completes the proof (after some rewriting). □

Since the righthand side of (21) is recognized as the LST of the uniform distribution, Lemma 3.2 leads to the following result for the distribution of the asymptotic scaled delay, where we take $\overset{d}{\to}$ to represent convergence in distribution.

**Theorem 3.3.** In case of deterministic setup times, the distribution of the asymptotic scaled delay is given by

$$
\frac{W_1}{S} \overset{d}{\to} W_1^* = \frac{1 - \rho_1}{1 - \rho} U_1, \quad (S \to \infty),
$$

(25)

where $U_1$ is uniformly distributed on $[\frac{1 - \Phi_1}{\Phi_1}, \frac{1}{\Phi_1}]$.

**Proof.** Follows directly from Lemma 3.2 in combination with the convergence theorem of Feller for Laplace-Stieltjes Transforms (see, e.g., page 652 of [3]). □

With the help of Lemma 3.2 we can derive similar results for the PGF of the distribution of the scaled queue length $\frac{L_1}{S}$ of queue 1 at arbitrary moments in time. That is,

$$
E[y^{\frac{L_1}{S}}] = E[e^{-\lambda_1 S(1 - \frac{1}{x}) \frac{W_1}{x}}] \xrightarrow{S \to \infty} E[y^{\lambda_1 \frac{W_1}{x}}],
$$

(26)

where the first equality follows from application of the distributional form of Little’s law [8] and the subsequent limit from the following standard limiting result,

$$
\lim_{x \to \infty} x \left(1 - a^\frac{1}{x}\right) = -\ln(a).
$$

(27)

We immediately observe from (26) that $\frac{L_1}{S}$ equals $\lambda_1 \frac{W_1}{S}$ in distribution as $S \to \infty$ implying that - although the individual service requests are discrete - the scaled queue length converges to a *continuous* uniform distribution in the limit of increasing setup times as well. Intuitively, we
can say that, when the setup times tend to infinity, the system behaves like a fluid model where customers keep trickling in and out like water. We come back to this issue in Section 4.

Besides results for queue lengths at arbitrary moments, our framework also allows us to derive results for the PGF of the distribution of the scaled queue length $X_1$ of queue 1 at a polling instant of this queue. That is, using (29) from [1] yields
\[ E[y^{X_1/S}] = e^{-S\dot{H}(y^{1/S})} \xrightarrow{S \to \infty} y^{E[X_1/S]}, \]  \tag{28}

since
\[ -S\dot{H}(y^{1/S}) = -S(y^{1/S} - 1) \left( \frac{\dot{H}(y^{1/S}) - \dot{H}(1)}{y^{1/S} - 1} \right) \xrightarrow{S \to \infty} -\ln(y)\dot{H}'(1) = \ln(y)E[X_1/S], \]  \tag{29}

where we have again used (27). In words, this means that the scaled number of customers at queue 1 at a polling instant of queue 1 becomes deterministic in the limit.

Theorem 3.3 reveals a number of properties about the dependence of the asymptotic scaled delay with respect to the system parameters, which are discussed below.

**Property 3.4.** We have

1. $W_1^*$ is independent of the visit order;
2. $W_1^*$ depends on the service discipline of queue 1 only through the exhaustiveness $\Phi_1$;
3. $W_1^*$ is independent of the service disciplines of the other queues;
4. $W_1^*$ depends on the arrival rate and service time distribution of queue 1 only through the occupation rate $\rho_1$;
5. $W_1^*$ depends on the arrival rate and service time distribution of the other queues only through the total occupation rate $\rho$.

\[ \square \]

First of all, it is important to stress that the above properties are in general not valid for systems with finite setup times. Property 3.4.2 has the important implication that one may classify and order various policies simply by their exhaustiveness factor without conducting a complete analysis of the policies. Furthermore, it implies that different policies are equivalent - in terms of the asymptotic scaled delay - in the asymptotic regime by proper adaptation of the policy parameters. For example, the binomial-gated policy is equivalent to the binomial-exhaustive policy as $S$ tends to infinity if we set
\[ q_i = p_i(1 - \rho_i), \]  \tag{30}

since in this way the exhaustiveness of both policies is equalized.

From Theorem 3.3 the moments of the scaled delay can be easily computed as done in the corollary below.

**Corollary 3.5.** In case of deterministic setup times, the moments of the asymptotic scaled delay are given by
\[ E[W_1^{*k}] = \left( \frac{1 - \rho_1}{1 - \rho} \right)^k \frac{1 - (1 - \Phi_1)^{k+1}}{(k + 1)\Phi_1^{k+1}}. \]  \tag{31}

\[ \square \]

The closed-form expression for the moments formed by (31) shows the following (monotonicity) properties of the system.
Property 3.6. For $k = 1, 2, \ldots$,
1. $E[W^*_k]$ monotonically decreases in $\Phi_1$;
2. $E[W^*_k]$ is minimized for $\Phi_1 = 1$.

More colloquially, this means the greater the exhaustiveness of the service discipline at a queue, the smaller all moments of the delay experienced by its customers. Recall that Property 3.4.3 has already shown that the delay of the customers at the other queues is independent of the service discipline at the queue under consideration. These properties are again not generally valid for stable systems with finite setup times.

Finally, we take a closer look at the central moments of the asymptotic scaled delay in the corollary below.

Corollary 3.7. In case of deterministic setup times, the central moments of the asymptotic scaled delay are given by

$$E[(W^* - \mu)^k] = \begin{cases} 0, & k = 1, 3, \ldots \\ \frac{1}{(k+1)^2}, & k = 2, 4, \ldots \end{cases}$$

where $\mu = E[W^*_1]$.

It is seen that these central moments are independent of all input parameters of all queues (arrival intensities, service time distribution, setup time distribution, service discipline, etc). Among other things, this means that the variance of the asymptotic scaled delay cannot be influenced by the choice of the service discipline. Once more, this is not generally true for systems with finite setup times. The subsection is closed with an example.

Example 3.8. Let us return to the policies introduced in Example 3.1.

1. In case of the exhaustive discipline, the scaled delay is uniformly distributed on $[0, \frac{1-\rho_1}{1-\rho}]$.
2. When the gated discipline is implemented, the scaled delay follows a uniform distribution on $[\frac{\rho_1}{1-\rho}, \frac{1-\rho}{1-\rho}]$.
3. Under the binomial-exhaustive discipline, the scaled delay is uniformly distributed on the interval $\left[\frac{1-\rho}{1-\rho} - \frac{q_1}{1-\rho} \frac{1}{1-\rho}, \frac{1-\rho}{1-\rho} \frac{1}{1-\rho} \right]$.
4. In case of the binomial-gated discipline, the scaled delay follows a uniform distribution on $\left[\frac{1-\rho}{1-\rho} - \frac{p_1}{1-\rho} \frac{1-\rho}{1-\rho}, \frac{1-\rho}{1-\rho} \frac{1}{1-\rho} \frac{1}{1-\rho} \right]$.

3.3 Limit theorems for general setups under heavy traffic

Very recently, [16] studied the delay distribution for polling systems with general branching-type service policies (and general setup time distributions) under heavy traffic. That is, the delay distribution is considered as function of $\rho$ where the arrival rates are variable, while the service time distributions and the ratios of the arrival rates are fixed. Subsequently, a closed-form expression for the scaled asymptotic delay is obtained, i.e., the limit of $1 - \rho$ times the delay, when $\rho$ tends to 1. Of particular interest for us is the fact that in heavy traffic the impact of higher moments of the setup times on the delay distribution vanishes, i.e., the scaled asymptotic delay depends on the marginal setup time distributions only through the first moment of the total setup time in a cycle.

The aim of the present subsection is to study the asymptotic delay in a polling system with generally distributed setups under heavy traffic when the setup times tend to infinity. The only restriction we make on the setup times is that the first two moments of all the setup times should exist, i.e., they should be finite. We, firstly, let the arrival rates increase in such a way that $\rho$ tends to 1, which allows us to exploit the heavy-traffic results from [16]. Secondly, we let the mean total setup in a cycle $E[S]$ tend to infinity. This step-by-step plan is formalized in the following lemma.
Lemma 3.9. In case of general setup times, the LST of the distribution of the asymptotic scaled delay under heavy traffic is given by

\[
\mathbb{E}[e^{-\omega(1-\rho)W_1}] \to \frac{1}{(1-\rho_1)\omega} \left( e^{- \frac{1-\rho_{i_1}}{\Phi_{i_1}}(1-\rho_i)\omega} - e^{- \frac{1}{\Phi_{i_1}}(1-\rho_i)\omega} \right), \quad (\rho \uparrow 1 \text{ and then } \mathbb{E}[S] \to \infty) \quad (33)
\]

Proof. First of all, we let \( \rho \) tend to 1 in such a way that we can apply the limit theorems of \([16]\), which imply that

\[
\mathbb{E}[e^{-\omega(1-\rho)W_1}] \to \frac{1}{(1-\rho_1)\mathbb{E}[S]\omega} \left( \delta\beta_{i_1}\Phi_{i_1} + (1-\Phi_{i_1})(1-\rho_1)\omega \right)
\]

where

\[
\beta = \frac{\sum_{i=1}^{N} \lambda_i \mathbb{E}[B_i]}{\sum_{i=1}^{N} \lambda_i \mathbb{E}[B^2_i]} \quad \text{and} \quad \delta = \sum_{i=1}^{N} \left( \frac{\rho_i(1-\rho_i)(1-\Phi_i)}{\Phi_i} + \rho_i \sum_{j=i+1}^{N} \rho_j \right). \quad (35)
\]

Applying the following standard limit result,

\[
\lim_{x \to \infty} \left( \frac{a}{a + \frac{b}{x}} \right)^{cx} = e^{-\frac{b}{a}},
\]

(36)

\[
\text{to the scaled delay } (1-\rho)\frac{W_1}{\mathbb{E}[S]} \text{ in (34) completes the proof (after some straightforward manipulations).}
\]

Lemma 3.9 has the following immediate consequence.

Theorem 3.10. In case of general setup times, the distribution of the asymptotic scaled delay under heavy traffic is given by

\[
\frac{(1-\rho)W_1}{\mathbb{E}[S]} \overset{d}{\to} \tilde{W}_1 = (1-\rho_1)U_1, \quad (\rho \uparrow 1 \text{ and then } \mathbb{E}[S] \to \infty),
\]

(37)

where \( U_1 \) is uniformly distributed on \([\frac{1-\Phi_{i_1}}{\Phi_{i_1}}, \frac{1}{\Phi_{i_1}}]\).

Proof. Follows directly from Lemma 3.9 in combination with the convergence theorem of Feller for Laplace-Stieltjes Transforms (see, e.g., page 652 of \([3]\)).

We note that in the case of deterministic setup times Theorem 3.10 with "\( \rho \uparrow 1 \) and then \( \mathbb{E}[S] \to \infty \)" replaced by "\( \mathbb{E}[S] \to \infty \) and then \( \rho \uparrow 1 \)" is implied by Theorem 3.3 and, subsequently, letting \( \rho \) tend to 1. Finally, we close this subsection with an example.

Example 3.11. For a second time, we return to the policies introduced in Example 3.1.

1. In case of the exhaustive discipline, the scaled delay in heavy traffic is uniformly distributed on \([0, 1-\rho_1]\).
2. When the gated discipline is implemented, the scaled delay in heavy traffic follows a uniform distribution on \([\rho_1, 1]\).
3. Under the binomial-exhaustive discipline, the scaled delay in heavy traffic is uniformly distributed on the interval \([(1-\rho_1)\frac{1}{\Phi_{i_1}}, (1-\rho_1)\frac{1}{\Phi_{i_1}}]\)
4. In case of the binomial-gated discipline, the scaled delay in heavy traffic follows a uniform distribution on \([\frac{1-\rho_1}{\rho_1}, \frac{1}{\rho_1}]\).
4 Discussion and extensions

In the present section, we not only elaborate on the applicability of the derived limit theorems - are they of any practical value - but we also discuss possible ways of extending the present study - partly this has already been done, partly this is left as work for further research.

Approximation. Theorem 3.3 suggests the following simple closed-form approximation for the delay distribution in systems with finite setups,

\[ P[W_i < x] \approx P[W_i^* < \frac{x}{S}] \tag{38} \]

and, similarly for the moments,

\[ E[W_i^k] \approx S^k E[W_i^{*k}] \tag{39} \]

where closed-form expressions for \( E[W_i^{*k}] \) are given in Corollary 3.5. Extensive validations of the above approximations fall outside the scope of the present paper, but we refer to [15] and [17] for numerical evaluations validating the above approximation in the special cases of exhaustive and gated service disciplines. Among other things, they show via some cases how "fast" the limiting distribution is approached. Furthermore, [17] shows via numerical testing that similar limit theorems carry over to more general systems with, e.g., dynamic visit orders.

Intuitive interpretation. The closed-form expression of the scaled delay distribution has an intuitively appealing interpretation, certainly worth mentioning. That is, in the case of increasing deterministic setup times the polling system converges to a deterministic cyclic system with continuous deterministic service rates \( \frac{1}{E[B_i]} \) and continuous demand rates \( \lambda_i \), \( i = 1, 2, \ldots, N \), which reveals itself, for example, in the fact that the scaled number of customers at queue \( i \) at a polling instant of queue \( i \) becomes deterministic in the limit as shown in (28). This means that in the limit the customers arrive to the system and are served at constant rates with no statistical fluctuation whatsoever and that the scaled queue lengths can be seen as continuous quantities, see (26). Therefore, the uniform distribution emerging in the limiting theorems can be explained by the fact that it represents the position of the server in the cycle on arrival of a tagged customer.

Furthermore, it is important to note that the lengths of the scaled visit and intervisit times in such a system, which also converge to a constant, are independent of the individual service disciplines. The minimum \( m \) and maximum \( M \) of the scaled amount of work at a certain queue during a cycle, however, do depend on the service discipline. That is, the scaled amount of work at queue 1 ranges between

\[ m = \frac{1 - \Phi_i \rho_i (1 - \rho_i)}{\Phi_i} \frac{1}{1 - \rho} \tag{40} \]

which is obtained at the end of a visit period of queue \( i \) and,

\[ M = \frac{1}{\Phi_i} \frac{\rho_i (1 - \rho_i)}{1 - \rho} \tag{41} \]

which is obtained at a polling instant of queue \( i \). Both the minimum and maximum scaled amount of work in a cycle are minimized by the exhaustive discipline, which is in agreement with Property 3.6.

The above intuitive explanation also clearly indicates the difficulties arising in a system with increasing stochastic setup times, since it is certainly not obvious how such a polling system behaves in the limit.

General arrival process. Throughout the present paper, we have assumed that the arrival processes follow Poisson distributions. If we take a second look at the intuitive interpretation of our results, one would however expect that also in case of general (renewal) arrival processes the
polling systems converge to a deterministic cyclic system when the setup times tend to infinity. Unfortunately, the techniques used throughout the present paper rely heavily on the Poisson assumption, i.e., we have exploited known results for polling systems with finite setup times and, subsequently, we have shown that significant simplifications result as the setup times tend to infinity. However, corresponding polling results for general arrival processes are not known.

To numerically test the above conjecture for general arrival processes, we have performed a couple of simulation experiments of exhaustive polling systems based on the simulation code described in [25]. We first consider a symmetric polling system with 3 queues, where the service times are exponential with mean 0.25. Interarrival times have mean 1 and the corresponding squared coefficient of variation $c^2_{A_i}$ is varied between 0.25, 0.5, 1 and 2. In order to obtain a distribution for these interarrival times, we fit a phase-type distribution on the first two moments as described in Appendix A (cf., e.g., [23]). In case the squared coefficient of variation equals 1 the arrival distribution is approximated by a Poisson distribution and this case is included as benchmark.

Table 1 shows the variance of the scaled number of customers at queue $i$ at a polling instant of queue $i$ for varying values of the marginal setup times $S_i$ in a cycle. From (6) in combination with the value of the exhaustiveness factor for the exhaustive discipline, i.e., $\Phi = 1$, we know that in the case of Poisson arrivals,

$$E\left[\frac{X_i}{S_i}\right] = \lambda_i \frac{1 - \rho_i}{1 - \rho} = 3,$$

(42)

which also holds for the other arrival processes as could be argued from a standard balance argument. From Table 1, we clearly see that the scaled variance approaches zero when the setup times tend to infinity. It goes without saying that a highly variable arrival process has a negative impact on how "fast" the limiting behavior is approached. Via Chebyshev's inequality (see, e.g., [18]) we know that a random variable with zero variance follows a deterministic distribution and, therefore, this observation provides empirical evidence for the fact that the scaled number of customers at queue $i$ at a polling instant of queue $i$ becomes deterministic. Therefore, it confirms the validity of our conjecture that the polling system converges to a deterministic cyclic system as the setup times increase to infinity. Obviously, a more extensive test bed is needed to test our hypothesis more rigorously, but without doubt extending our work to general arrival processes is a very interesting topic for further research.

### Table 1: Variance of the scaled number of customers at queue $i$ at a polling instant of queue $i$.  

| $S_i$ | $c^2_{A_i} = 0.25$ | $c^2_{A_i} = 0.5$ | $c^2_{A_i} = 1$ | $c^2_{A_i} = 2$ |
|-------|--------------------|--------------------|--------------------|--------------------|
| 1     | 0.121              | 0.167              | 0.259              | 0.444              |
| 10    | 0.012              | 0.017              | 0.026              | 0.044              |
| 50    | 0.002              | 0.003              | 0.005              | 0.009              |
| 100   | 0.001              | 0.002              | 0.003              | 0.004              |

Joint distributions. The present paper focusses on the marginal delay and queue length distributions, since these are in most applications the most important performance measures. Among other things, we have observed that the polling system converges to a deterministic system: (28) shows, for example, that the scaled number of customers at queue $i$ at a polling instant of queue $i$ becomes deterministic when the setup times tend to infinity. In the special case of exhaustive and gated service, [15] conjectured that the scaled numbers of customers of all queues become deterministic at such an instant. Within the framework of the present paper this can be easily rigorously proven. That is, [1] proves that the fundamental relationship between polling systems with and without setup times occurs at the level of the joint queue length distributions. By applying the exact same steps we took in (28) of the present paper to (29) of [1], which holds for the joint distributions, the methodology of the present paper can be readily used to prove (and
extend to the complete class of branching-type policies) the conjecture of [15].

**Practical application.** Our motivation to study the setup time asymptotics in polling systems is based on the specific application area of base-stock policies in inventory control. In this section this area of application is discussed in some detail. That is, consider a production-inventory system with a single production capacity for multiple products, in which there is an infinite stock space for each product and raw material is always available. Demands for the various products arrive according to stationary and mutually independent stochastic processes. Demand that cannot be satisfied directly from stock is backlogged until the product becomes available after production. The individual products are produced in a make-to-stock fashion with possibly stochastic production times. A possibly stochastic setup time occurs before the start of the production of a product. Finally, only one product can be produced at a time. This setting is referred to as the stochastic economic lot scheduling problem (SELSP) (see [26], for a survey).

In many firms encountering the SELSP, cyclic base-stock policies are used for the control of the inventory of each product, which work as follows (see, e.g., [4, 5]):

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

Now, the production facility, where the production orders queue up, can be represented as a polling model by identifying each product with a queue and the demand process of a product with the arrival process at the corresponding queue (cf. [4, 5]).

The SELSP is a common problem in process industries, where the setup times are typically extremely large and deterministic. We believe that the results of the present paper give, therefore, new and fundamental insights into the behavior and performance of base-stock policies in process industries. In particular, we have shown that, in the case of increasing deterministic setup times, the polling system converges to a deterministic cyclic system. A reasonable hypothesis is that, in practice, production managers rely more on deterministic production strategies in production environments with significant setups than they do in environments with small setups in which stochastic (dynamic) policies seem to be more appropriate. We hope that this observation stimulates researchers to conduct a large-scale empirical study that investigates the main characteristics of production strategies in environments both with no - or negligible - setup times on the one hand and extremely large setup times on the other hand.

**Heavy traffic.** The stability conditions given at the end of Section 2 indicate that the delays grow without bound not only as $S \to \infty$ but also as $\rho \uparrow 1$. The asymptotic analysis of branching-type polling system in the latter heavy traffic case is the topic of the recent paper [16] (see, also, Section 3.3 of the present paper). The limiting behavior in the present paper turns out to be fundamentally different from that in the heavy-traffic scenario, where the gamma distribution shows up time after time. Surprisingly, in such a heavy-traffic scenario the delay distributions also depend on the service discipline of the queues only through the exhaustiveness factors. Therefore, the results of the present paper and [16] show that the influence of different service policies, but with the same exhaustiveness, becomes the same when the system is in overload either due to large setups or to heavy traffic. We stress that the proper operation of polling systems is particularly critical in such overload situations.

**Globally-gated service policy.** As alternative for the standard gated policy, [2] proposed the so-called globally gated service policy on which we focus now. Under this policy, when the server arrives at queue 1, all customers present in the system are marked and during the coming cycle all and only the marked customers are served. It is important to stress that the globally gated service discipline does not satisfy Property 2.1 for queue $i > 1$. That is, the number of customers present at the start of a visit to queue $i$ can be divided into customers standing before the global
gate, which are only served in the next cycle, and behind the global gate, which are served in the current cycle. These customers are not replaced in an i.i.d. manner, since the former and latter group are replaced by a random population having probability generating function \( z_i \) and \( \beta_i(\sum_{j=1}^{N} \lambda_j (1 - z_j)) \), respectively.

Although the globally gated policy closely resembles the regular gated policy, its analysis is less intricate which in turn allows for the derivation of an explicit expression for the LST of the delay distribution (in contrast to the gated policy). That is, by applying a similar approach as in Section 3 to (2.20) of [2], in case of deterministic setup times, the LST of the distribution of the asymptotic scaled delay can be easily shown to be

\[
\mathbb{E}[e^{-S_i} \frac{W_i}{S}] = \frac{1}{(1 - \rho_i)} \omega e^{-\frac{\sum_{i=1}^{N-1} S_j}{S}} \omega \left( e^{\frac{\sum_{i=1}^{N-1} \rho_j}{1 - \rho}} - e^{\frac{1 + \sum_{i=1}^{N-1} \rho_j}{1 - \rho}} \right), \quad i = 1, 2, \ldots, N, \quad (S \to \infty), \quad (43)
\]

from which again the continuous uniform distribution can be recognized. That is, the distribution of the asymptotic scaled delay for the globally-gated policy reads for \( i = 1, 2, \ldots, N \),

\[
\frac{W_i}{S} \xrightarrow{d} W_i^* = U_i, \quad (S \to \infty),
\]

where \( U_i \) is uniformly distributed on \( [\frac{\sum_{j=1}^{i-1} S_j}{S} + \frac{\sum_{j=1}^{i} \rho_j}{1 - \rho} - \frac{\sum_{j=1}^{i-1} S_j}{S} + \frac{1 + \sum_{j=1}^{i-1} \rho_j}{1 - \rho}] \). Remark that terms like \( \frac{\sum_{j=1}^{i-1} S_j}{S} \) converge since we fix the ratios of the setup times in our analysis. Finally, the following ordering of the mean asymptotic scaled delays is readily observed from (44),

\[
\mathbb{E}[W_1^*] < \mathbb{E}[W_2^*] < \ldots < \mathbb{E}[W_N^*],
\]

which also turns out to hold for the delay in stable polling systems with finite setups as demonstrated by [2].

**Acknowledgment**

The author wishes to thank Sem Borst and Onno Boxma for valuable discussions and for useful comments on earlier drafts of the present paper. Moreover, Sem Borst is thanked for the derivation of (18). Furthermore, the author is indebted to Marcel van Vuuren for his assistance in using the simulation program used in Section 4.

**A Appendix**

To obtain an approximating distribution of a positive random variable \( X \), one may fit a phase-type distribution on the mean \( \mathbb{E}[X] \) and the coefficient of variation \( c_X \) by using the following approach [23]. First of all, a random variable \( X \) is defined to have to a Coxian distribution of order \( k \) if it has to go through up to at most \( k \) exponential phases, where phase \( n \) has rate \( \mu_n \), \( n = 1, 2, \ldots, k \). It starts in phase 1 and after phase \( n \), \( n = 1, 2, \ldots, k - 1 \), it ends with probability \( 1 - p_n \), whereas it enters phase \( n + 1 \) with probability \( p_n \). Finally, \( p_k \) is defined to equal zero.

Now, the distribution of \( X \) is approximated as follows. If \( c_X^2 > 1 \), then the rate and coefficient of variation of the Coxian\(_2\) distribution matches with \( \mathbb{E}[X] \) and \( c_X \), provided the parameters are chosen as (cf. [14]):

\[
\mu_1 = 2/\mathbb{E}[X], \quad p_1 = \frac{1}{2c_X^2}, \quad \text{and} \quad \mu_2 = p_1 \mu_1.
\]

If \( 1/k \leq c_X^2 \leq 1/(k-1) \) for some \( k \geq 2 \), then the rate and coefficient of variation of the Erlang\(_{k-1,k}\) distribution, which is a special case of a Coxian distribution of order \( k \), matches with \( \mathbb{E}[X] \) and
\( c_X \), provided the parameters are chosen as (cf. [23]):

\[
\begin{align*}
  p_n &= 1, & n = 1, 2, \ldots, k - 2, \\
  p_{k-1} &= 1 - \frac{k c_X^2 - \sqrt{k(1 + c_X^2) - k^2 c_X^2}}{1 + c_X^2}, \\
  \mu_1 &= \mu_2 = \ldots = \mu_k = (k - p)E[X].
\end{align*}
\]

Of course, also other phase-type distributions may be fitted on the mean and the coefficient of variation, but numerical experiments suggest that choosing other distributions only has a minor effect on the results, as shown in [7].

References

[1] Borst, S.C., Boxma, O.J., (1997). *Polling models with and without switchover times* (Operations Research, vol. 45, no. 4, pp. 536-543).

[2] Boxma, O.J., Levy, H., Yechiali, U., (1992). *Cyclic reservation schemes for efficient operation of multiple-queue single-server systems* (Annals of Operations Research, vol. 35, pp. 187-208).

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

[4] Federgruen, A., Katalan, Z., (1996). *The stochastic economic lot scheduling problem: cyclical base-stock policies with idle times* (Management Science, vol. 42, pp. 783-796).

[5] Federgruen, A., Katalan, Z., (1998). *Determining production schedules under base-stock policies in single facility multi-item production systems* (Operations Research, vol. 46, pp. 883-898).

[6] Fuhrmann, S.W., (1981). *Performance analysis of a class of cyclic schedules* (Bell Laboratories Technical Memorandum 81-59531-1).

[7] Johnson, M.A., (1993). *An empirical study of queueing approximations based on phase-type distributions* (Stochastic Models, vol. 9, no. 4, pp. 531-561).

[8] Keilson, J., Servi, L.D., (1990). *The distributional form of Little’s law and the Fuhrmann-Cooper decomposition* (Operations Research Letters, vol. 9, no. 4, pp. 239-247).

[9] Levy, H., Sidi, M., (1990). *Polling systems: applications, modeling and optimization* (IEEE Transactions on Communications, vol. COM-38, no. 10, pp. 1750-1760).

[10] Levy, H., (1988). *Optimization of polling systems: The fractional exhaustive service method* (Report, Tel-Aviv University).

[11] Levy, H., (1989). *Analysis of cyclic polling systems with binomial gated service* (In Performance of Distributed and Parallel Systems, (eds.) T. Hasegawa, H. Takagi, Y. Takahashi, North-Holland, Amsterdam, pp. 127-139).

[12] Mack, C., Murphy, T., Webb, N.L., (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, vol. 19, no. 1, pp. 166-172).

[13] Mack, C., (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, vol. 19, no. 1, pp. 173-178).

[14] Marie, R.A., (1980). *Calculating equilibrium probabilities for \( \lambda(n)/C_k/1/N \) queue* (Proceedings Performance ’80, Toronto, pp. 117-125).

[15] Mei, R.D. van der, (1999). *Delay in polling systems with large switch-over times* (Journal of Applied Probability, vol. 36, pp. 232-243).
[16] Mei, R.D. van der, (2006). *Towards a unifying theory on branching-type polling models in heavy traffic* (Report, Vrije Universiteit).

[17] Olsen, T., (2001). *Limit theorems for polling models with increasing setups* (Probability in the Engineering and Informational Sciences, vol. 15, pp. 35-55).

[18] Papoulis, A., (1984). *Probability, Random Variables, and Stochastic Processes, 2nd ed.* (McGraw-Hill, New York).

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

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

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

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

[23] Tijms, H.C., (1994). *Stochastic Models: An Algorithmic Approach* (John Wiley & Sons, Chichester).

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

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

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

[27] Winands, E.M.M., Adan, I.J.B.F., Houtum, G.J. van, (2006). *Mean value analysis for polling systems* (Queueing Systems, vol. 54, no. 1, pp. 45-54).

[28] Winands, E.M.M., (2006). *On polling systems with large setups* (To appear in Operations Research Letters).