ELSEVIER

Performance Evaluation 38 (1999) 133–148

PERFORMANCE EVALUATION
An International Journal

www.elsevier.com/locate/peva

Distribution of the delay in polling systems in heavy traffic

R.D. van der Mei *

AT&T Labs, Department of Network Design and Performance Analysis, Middletown, NJ 07748, USA

Received 22 October 1998; received in revised form 5 August 1999

Abstract

We consider asymmetric cyclic polling systems with an arbitrary number of queues, with general mixtures of exhaustive and gated service and with generally distributed service-times and switch-over times, in heavy traffic. We derive closed-form expressions for the Laplace–Stieltjes transform (LST) of the steady-state delay incurred at each of the queues, under standard heavy-traffic scalings. The expressions give an explicit characterization of the complete (scaled) waiting-time distributions at each of the queues. The results are strikingly simple and provide a variety of new insights into the behavior of heavily loaded polling systems. In addition, the results lead to simple and fast-to-evaluate approximations for the waiting-time distributions in stable polling systems that are close to saturation. Numerical results demonstrate that the approximations are highly accurate in many practical heavy-traffic scenarios. ©1999 Elsevier Science B.V. All rights reserved.

Keywords: Polling systems; Delay; Heavy traffic; Approximations

1. Introduction

The basic polling system consists of a number of queues attended by a single server that visits the queues in cyclic order to render service to the customers waiting at the queues. Polling models occur naturally in the modeling systems in which service capacity (CPU, bandwidth, processing power) is shared by different types of users, each having specific traffic characteristics and Quality of Service (QoS) requirements. Polling models find many applications in areas like computer-communication networks, production systems and maintenance and manufacturing. We refer to [21] for an extensive overview of the applicability of polling models. Because of their wide applicability, polling models have received a lot of attention in the literature since the late 1960s (cf. [25,26] for overviews). An exact analysis of the delay in polling models is generally difficult, and hopes for explicit solutions are often abandoned in favor of numerical methods. However, the usefulness of numerical techniques is limited in the sense that they do not reveal explicitly how the system performance depends on the system parameters, and therefore, can

* Current work address: KPN Research, Department of Planning, Performance & Reliability, Room LC139, P.O. Box 421, 2260 AK Leidschendam, The Netherlands. Tel.: +31-70-3326452.
E-mail address: r.d.vandermei@research.kpn.com (R.D. van der Mei)

PII: S0166-5316(99)00048-6

only provide limited insight into the behavior of the system. Exact solutions provide much more insight into the dependence of the performance measures on the system parameters. Moreover, the efficiency of the numerical algorithms may degrade significantly for heavily loaded, highly asymmetric systems with a large number of queues, while the proper operation of the system is particularly critical for those systems. These observations raise the importance of an exact asymptotic analysis of the performance of polling models in heavy traffic.

In the literature, exact results on polling models are scarce. The most general results are the formulations of pseudo-conservation laws, giving exact expressions for a specific weighted sum of the expected waiting times [4]. Exact results on the complete probability distribution of the delay are mainly restricted to two-queue models (cf., e.g., [5,12,14,34]), but even in those cases non-trivial numerical techniques need to be used to obtain the tail probabilities of the delay. In the absence of exact distributional results, numerical techniques have been proposed to calculate the waiting-time and queue-length distributions, like Blanc's power-series algorithm [1], Leung's technique based on discrete Fourier transforms [20] or the numerical transform inversion technique [8]. Federgruen and Katalan [15] propose a method to approximate the queue-length and waiting-time distributions in a class of polling models. Recently, several papers have focused on the heavy-traffic behavior of polling models. For a two-queue model with exhaustive service at both queues and with zero switch-over times, Coffman et al. [10] show that the total amount of unfinished work in the system tends to a Reflected Brownian Motion, under standard heavy-traffic scalings. In [11], the results in [10] are extended to the case of nonzero switch-over times. Using the results in [10], Reiman and Wein [23] study set-up scheduling problems for two-class single-server queues. Van der Mei and Levy [27,28] and Van der Mei [29–31] use the concept of descendant sets to obtain expressions for the moments of the delay in heavy traffic. Kroese [19] analyzes the heavy-traffic behavior of continuous polling systems, and shows that the total number of customers has approximately a gamma distribution.

We consider asymmetric cyclic polling models with general mixtures of exhaustive and gated service, and with general service times and switch-over times. We study the distribution of the delay incurred at each of the queues in heavy traffic, i.e., in which the load (denoted by ρ) tends to unity. Since all queues become instable in the limiting case, we focus on the limiting distribution of the random variable $(1-\rho)W_i$, referred to as the scaled delay at queue i. We derive closed-form expressions for the Laplace–Stieltjes transform (LST) of the limiting distribution of the scaled delay at queue i, in a general parameter setting. The key observation underlying these results of the fact that both the (scaled) cycle times and intervisit times can be shown to converge to gamma-distributions with known parameters. This leads to an explicit and complete characterization of the complete waiting-time distributions in heavy traffic. The results are remarkably simple and provide a variety of insights into the heavy-traffic behavior of the system that have not been observed before. In addition, the results suggest simple and fast-to-evaluate approximations for the waiting-time distributions in stable polling systems that are close to saturation. Numerical results show that the approximations are highly accurate in many practical heavy-traffic scenarios, where the load is 80–90% or more.

This paper generalizes, and explicitly uses, the results obtained in [30], where we obtained expressions for the moments of the delay incurred at the queues, in heavy traffic. The motivation for extending the results in [30] to the complete distributions of the delay is threefold. First, in many applications (e.g., in telecommunication networks) the main performance measure of interest is the probability that the delay exceeds a certain threshold, rather than more aggregated performance measures like the moments of the delay. In view of those applications, the importance of extending the results in [30] to the complete probability distribution of the delay is evident. Second, the computation times of the existing numerical techniques for evaluating the tail probabilities of the delay may degrade dramatically for heavily-loaded

systems. This raises the need for simple and fast approximations of the tail probabilities of the delay in heavy-loaded polling systems. Such an approximation is directly obtained by the results presented in this paper (see Section 5 for more details). Third, we have a theoretical interest in obtaining an exact characterization of the (asymptotic) waiting-time distributions, showing explicitly how they depend on the system parameters.

The remainder of this paper is organized as follows. In Section 2 the model is described. In Section 3 we derive closed-form expressions for the LST of the limiting distribution of the scaled delay, in heavy traffic. In Section 4 we discuss a number of properties of the delay distribution with respect to specific system parameters. In Section 5 we address the practicality of the results by assessing the accuracy of the approximations suggested by the expressions obtained in Section 3. In Section 6 we address a number of topics for further research.

2. Model description

We consider a system consisting of $N \geq 2$ infinite-buffer queues, $Q_1, \ldots, Q_N$, and a single server that visits and serves the queues in cyclic order. Customers arrive at Q_i according to a Poisson arrival process with rate λ_i, and are referred to as type-i customers. The total arrival rate is denoted by $\Lambda = \sum_{i=1}^{N} \lambda_i$. The service time of a type-i customer is a random variable B_i, with LST $B_i^*(\cdot)$ and with finite kth moment $b_i^{(k)}$, $k = 1, 2, \ldots$. The kth moment of the service time of an arbitrary customer is denoted by $b^{(k)} = \sum_{i=1}^{N} \lambda_i b_i^{(k)}/\Lambda$, $k = 1, 2, \ldots$. The load offered to Q_i is $\rho_i = \lambda_i b_i^{(1)}$, and the total offered load is equal to $\rho = \sum_{i=1}^{N} \rho_i$. Define a polling instant at Q_i as an epoch at which the server arrives at Q_i. Similarly, a departure instant at Q_i is defined as an epoch at which the server departs from Q_i. Denote by I_i the intervisit time of Q_i, i.e., the duration of the time between a departure of the server from Q_i and its successive visit to Q_i, and denote the corresponding LST by $I_i^*(\cdot)$. Define the cycle time C_i at Q_i to be the time between two successive polling instants at Q_i, and denote the corresponding LST by $C_i^*(\cdot)$. The service at each queue is either according to the gated policy or the exhaustive policy. Under the gated policy only the customers that were present at the polling instant at Q_i are served; customers that arrive at Q_i while it is being served are served during the next visit of Q_i. Under the exhaustive policy the server visits Q_i until it is empty. The service policy at each queue remains the same for all visits. Define $E := \{i\colon Q_i \text{ is served exhaustively}\}$ and $G := \{i\colon Q_i \text{ receives gated service}\}$. At each queue the customers are served on a FIFO basis. The switch-over time required by the server to proceed from Q_i to Q_{i+1} is a random variable R_i with finite moments and with mean r_i. Denote by $r = \sum_{i=1}^{N} r_i > 0$ the expected total switch-over time per cycle. All interarrival times and service times are assumed to be mutually independent and independent of the state of the system. A necessary and sufficient condition for the stability of the system is $\rho < 1$ (cf. [17]).

Let W_i be the delay incurred by an arbitrary customer at Q_i. Throughout, W_i will be considered as a function of ρ, where the arrival rates are variable, while the service-time distributions and the ratios of the arrival rates are kept fixed. It is known that when $\rho \uparrow 1$, all queues become unstable. Therefore, we focus on the random variable $(1-\rho)W_i$ (referred to as the *scaled* delay at Q_i), and derive its limiting distribution when ρ tends to unity; thus, the analysis is focused on the distribution of the random variables

$$\tilde{W}_i := \lim_{\rho \uparrow 1} (1-\rho) W_i \quad (i = 1, \ldots N). \tag{1}$$

The main result of the paper is the derivation of a closed-form expression for the LST of $\tilde{W}_i$.

The following notation will be convenient. For an event F, denote by I_F the indicator function on F. Denote by $\underline{e}_i$ the ith unit vector ($i = 1, \ldots, N$). Finally, for each variable x that is a function of ρ, $\hat{x}$ denotes its value evaluated at $\rho = 1$.

3. Analysis

The waiting-time distribution at Q_i is related to the intervisit-time and cycle-time distributions according to the following relations (cf., e.g. [25]): For $\mathrm{Re}\, s > 0$,

$$W_i^*(s) = \frac{(1-\rho_i)s}{s - \lambda_i(1 - B_i^*(s))} \cdot \frac{1 - I_i^*(s)}{sE[I_i]} \quad (i \in E), \tag{2}$$

$$W_i^*(s) = \frac{(1-\rho_i)s}{s - \lambda_i(1 - B_i^*(s))} \cdot \frac{C_i^*(\lambda_i(1 - B_i^*(s))) - C_i^*(s)}{(1-\rho_i)sE[C_i]} \quad (i \in G). \tag{3}$$

Thus, the waiting-time distributions are completely determined by the distributions of the intervisit times and cycle times. The following result gives an expression for the limiting moments of the (scaled) intervisit times and cycle times when ρ tends to 1.

Theorem 1. *For $k = 1, 2, \ldots,$*

$$\lim_{\rho \uparrow 1} (1-\rho)^k E[I_i^k] = (1-\hat{\rho}_i)^k \prod_{j=0}^{k-1} \left[r + j \left(\frac{b^{(2)}/b^{(1)}}{\delta} \right) \right] \quad (i \in E), \tag{4}$$

$$\lim_{\rho \uparrow 1} (1-\rho)^k E[C_i^k] = \prod_{j=0}^{k-1} \left[r + j \left(\frac{b^{(2)}/b^{(1)}}{\delta} \right) \right] \quad (i \in G), \quad \textit{with } \delta := 1 - \sum_{m \in E} \hat{\rho}_m^2 + \sum_{m \in G} \hat{\rho}_m^2. \tag{5}$$

Proof. Without loss of generality, assume $i = 1$. Denote by X_1 the number of customers at Q_1 at an arbitrary polling instant at Q_1, and denote its corresponding probability generating function (PGF) by $X_1^*(\cdot)$. The moments of X_1 can be obtained numerically via a set of recursive relations [18]. The heavy-traffic behavior of these recursive relations is discussed extensively in [29]. These results, in turn, can be used to obtain the following expressions for the moments of X_1 (cf. [30] for details): For $k = 1, 2, \ldots,$

$$\lim_{\rho \uparrow 1} (1-\rho)^k E[X_1^k] = \hat{\lambda}_1^k (1 - \hat{\rho}_1 I_{\{1 \in E\}})^k \prod_{j=0}^{k-1} \left[r + j \left(\frac{b^{(2)}/b^{(1)}}{\delta} \right) \right]. \tag{6}$$

For the case of exhaustive service at Q_1, the customers present at a polling instant at Q_1 are exactly those who arrived during the preceding intervisit period. This implies, for $|z| \leq 1$, $X_1^*(z) = I_1^*(\lambda_1(1-z))$, or equivalently $I_1^*(s) = X_1^*(1 - s/\lambda_1)$, for $\mathrm{Re}\, s > 0$. Then it is readily verified by differentiating Eq. (6) k times that, for $1 \in E$, $k = 1, 2, \ldots,$

$$\lim_{\rho\uparrow 1}(1-\rho)^k E[I_1^k] = \lim_{\rho\uparrow 1}(1-\rho)^k \frac{E[X_1(X_1-1)\cdots(X_1-k+1)]}{\lambda_1^k}$$
$$= (1-\hat{\rho}_1)^k \prod_{j=0}^{k-1}\left[r + j\left(\frac{b^{(2)}/b^{(1)}}{\delta}\right)\right]. \tag{7}$$

Similarly, for the case of gated service at Q_1, the customers present at Q_1 at a polling instant at Q_1 are exactly those which arrived during the preceding cycle, which implies $X_1^*(z) = C_1^*(\lambda_1(1-z))$, for $|z| \le 1$, or equivalently, $C_1^*(s) = X_1^*(1-s/\lambda_1)$. Then using it is readily verified from Eq. (6) that, for $1 \in G, k = 1, 2, \ldots, N$,

$$\lim_{\rho\uparrow 1}(1-\rho)^k E[C_1^k] = \lim_{\rho\uparrow 1}(1-\rho)^k \frac{E[X_1(X_1-1)\cdots(X_1-k+1)]}{\lambda_1^k} = \prod_{j=0}^{k-1}\left[r + j\left(\frac{b^{(2)}/b^{(1)}}{\delta}\right)\right]. \tag{8}$$

This completes the proof of Theorem 1. □

A random variable Γ with a gamma-distribution with scale parameter $\alpha > 0$ and rate parameter $\mu > 0$ has the following probability density function:

$$f_\Gamma(t) := \frac{1}{\Gamma(\alpha)} e^{-\mu t}\mu^\alpha t^{\alpha-1}, \quad t \ge 0, \quad \text{where } \Gamma(\alpha) := \int_0^\infty e^{-t}t^{\alpha-1}dt. \tag{9}$$

It is readily verified that the LST and the moments of Γ are given by

$$\Gamma^*(s) = \left(\frac{\mu}{\mu+s}\right)^\alpha \quad (\mathrm{Re}\, s > 0), \quad \text{and} \quad E[\Gamma^k] = \frac{\prod_{j=0}^{k-1}(\alpha+j)}{\mu^k} \quad (k = 1, 2, \ldots), \tag{10}$$

respectively.

A sequence of real-valued random variables $\{X_n,\ n = 1, 2, \ldots\}$ is said to converge in distribution to a random variable X, denoted by $X_n \to_{\mathrm{d}} X$, if there exists a dense subset A of $\mathcal{R}$ (i.e., the set of real numbers) such that $\lim_{n\to\infty} P(X_n < a) \to P(X < a)$, for all $a \in A$. Similarly, two random variables X and Y are said to have the same distribution (almost surely), denoted by $X =_{\mathrm{d}} Y$, if there exists a dense subset A of $\mathcal{R}$ such that $P(X < a) = P(Y < a)$ for all $a \in A$.

An important observation is that the moments expressed in Eqs. (4),(5) have the same functional form as the moments of a gamma distribution (see (10)). More precisely, it may be shown that the moments of the (scaled) intervisit times and cycle times converge to the moments of gamma-distributions with properly chosen parameters (when ρ tends to 1). This convergence *in moments* will be used to show that both the (scaled) intervisit times and the cycle time also converge *in distribution* (which is much stronger than convergence in moments) to gamma-distributions with known parameters. In general, however, a probability distribution is not uniquely determined by its moments; in fact, the moments of a probability distribution may not even exist at all. Nonetheless, a special class (say $\mathcal{P}$) of probability distributions *is* uniquely determined by its (finite) moments. For this class $\mathcal{P}$ of probability distributions, the so-called *Method of Moments* states that convergence in moments *implies* convergence in distribution. Below we show that the gamma-distribution belongs to $\mathcal{P}$, so that the Method of Moments applies.

Lemma 1. *Let Γ be a gamma-distributed random variable with parameters α and μ. Let Y be a random variable with finite moments, such that*

$$E[Y^k] = E[\Gamma^k], \quad k = 1, 2, \ldots . \tag{11}$$

Then $Y =_{\mathrm{d}} \Gamma$.

Proof. A sufficient condition for the (almost sure) uniqueness of the gamma-distribution is the following (cf. [16, p. 514, Eq. (4.15)]):

$$\limsup_{k\to\infty} \frac{1}{k} E[\Gamma^k]^{1/k} < \infty. \tag{12}$$

Using (10), the validity of this requirement follows from the following relations:

$$\lim_{k\to\infty} \frac{1}{k} E[\Gamma^k]^{1/k} = \lim_{k\to\infty} \frac{1}{k} \left[\prod_{j=0}^{k-1} \frac{\alpha + j}{\mu} \right]^{1/k} \leq \lim_{k\to\infty} \frac{1}{k} \left[\frac{\alpha + k}{\mu} \right]^{k/k} = \lim_{k\to\infty} \frac{\alpha + k}{k\mu} = \frac{1}{\mu} < \infty. \tag{13}$$

This completes the proof of Lemma 1. □

The following result shows that convergence in moments implies convergence in distribution if the limiting distribution is uniquely determined by its moments.

Lemma 2. *Let Y be a random variable whose distribution is uniquely determined by its finite moments $E[Y^k], k = 1, 2, \ldots$. Suppose $\{Y_n\}$ is a sequence of random variables with finite moments $E[Y_n^k], k = 1, 2, \ldots$, and that*

$$\lim_{n\to\infty} E[Y_n^k] = E[Y^k], \quad k = 1, 2, \ldots . \tag{14}$$

Then $Y_n \to_{\mathrm{d}} Y$.

Proof. See [9, Theorem 4.5.5]. Combining Lemmas 1 and 2 leads to the following result, which will play a key role in the derivation of the results. □

Theorem 2 (Method of Moments). *Let Γ be a gamma-distributed random variable with parameters α and μ. Let $\{Y_n\}$ be a sequence of random variables with finite moments, satisfying*

$$\lim_{n\to\infty} E[Y_n^k] = E[\Gamma^k], \quad k = 1, 2, \ldots . \tag{15}$$

Then $Y_n \to_{\mathrm{d}} \Gamma$.

Proof. Follows directly from Lemmas 1 and 2. □

In words, Theorem 2 states that if the moments of a sequence of random variables $\{Y_n\}$ converge to the corresponding moments of a gamma-distribution, then $\{Y_n\}$ converges in distribution to that gamma-distribution. Based on Theorem 2, we are now ready to formulate the distributional form of Theorem 1.

Theorem 3 (Convergence in distribution of the intervisit times). *If $i \in E$, then*

$$(1 - \rho) I_i \to_{\mathrm{d}} \tilde{I}_i \quad (\rho \uparrow 1), \tag{16}$$

where $\widetilde{I}_i$ has a gamma-distribution with parameters

$$\alpha := r\delta \frac{b^{(1)}}{b^{(2)}}, \quad \mu_i := \frac{\delta}{1 - \hat{\rho}_i} \frac{b^{(1)}}{b^{(2)}} \quad with \;\; \delta := 1 - \sum_{m\in E} \hat{\rho}_m^2 + \sum_{m\in G} \hat{\rho}_m^2. \tag{17}$$

Proof. From Eq. (10) it is readily seen that the moments in (4) converge to the moments of a gamma-distribution with parameters α and μ_i, defined in (17). Then we apply Theorem 2 to show that this convergence in moments implies convergence in distribution. To this end, let $\{\rho^{(n)}, n = 1, 2, \dots\}$ be an arbitrary sequence of ρ-values with $\rho^{(n)} \uparrow 1$. The results follow then from Theorem 2 by taking $Y_n^{(i)} := (1 - \rho^{(n)})W_i$ (note that the distribution of W_i is also a function of $\rho^{(n)}$). This completes the proof of the result. □

Theorem 4 (Convergence in distribution of the cycle times). *If $i \in G$, then*

$$(1 - \rho)C_i \to_{\mathrm{d}} \tilde{C}_i \quad (\rho \uparrow 1), \tag{18}$$

where $\tilde{C}_i$ has a gamma-distribution with parameters

$$\alpha := r\delta\frac{b^{(1)}}{b^{(2)}}, \quad \mu_i := \delta\frac{b^{(1)}}{b^{(2)}} \quad with\ \ \delta := 1 - \sum_{m\in E}\hat{\rho}_m^2 + \sum_{m\in G}\hat{\rho}_m^2. \tag{19}$$

Proof. Similar to the proof of Theorem 3, by using (5) instead of (4). □

We are now ready the present the main results of the paper.

Theorem 5 (Main result). *For $i = 1, \dots, N$,*

$$(1 - \rho)W_i \to_{\mathrm{d}} \tilde{W}_i \quad (\rho \uparrow 1), \tag{20}$$

where the Laplace–Stieltjes transform of $\tilde{W}_i$ is given by the following expressions: For $\mathrm{Re}\, s > 0$,

$$\tilde{W}_i^*(s) = \frac{1}{(1 - \hat{\rho}_i)rs}\left\{1 - \left(\frac{\mu_i}{\mu_i + s}\right)^{\alpha}\right\} \quad (i \in E), \tag{21}$$

$$\tilde{W}_i^*(s) = \frac{1}{(1 - \hat{\rho}_i)rs}\left\{\left(\frac{\mu_i}{\mu_i + s\hat{\rho}_i}\right)^{\alpha} - \left(\frac{\mu_i}{\mu_i + s}\right)^{\alpha}\right\} \quad (i \in G), \tag{22}$$

where

$$\alpha := r\delta\frac{b^{(1)}}{b^{(2)}}, \quad \mu_i := \frac{\delta}{1 - \hat{\rho}_i I_{\{i\in E\}}}\frac{b^{(1)}}{b^{(2)}}, \quad \delta := 1 - \sum_{m\in E}\hat{\rho}_m^2 + \sum_{m\in G}\hat{\rho}_m^2. \tag{23}$$

Proof. For $i \in E$, relation (21) follows from the following sequence of equalities: For $\mathrm{Re}\, s > 0$,

$$\tilde{W}_i^*(s) := \lim_{\rho\uparrow 1} W_i^*(s(1 - \rho)) = \lim_{\rho\uparrow 1}\frac{(1 - \rho_i)s(1 - \rho)}{s(1 - \rho) - \lambda_i(1 - B_i^*(s(1 - \rho)))} \cdot \lim_{\rho\uparrow 1}\frac{1 - I_i^*(s(1 - \rho))}{s(1 - \rho)E[I_i]} \tag{24}$$

$$= \frac{1}{(1 - \hat{\rho}_i)rs}\left\{1 - \left(\frac{\mu_i}{\mu_i + s}\right)^{\alpha}\right\} \quad (i \in E). \tag{25}$$

The second equation follows from (2), and the third equality follows from Theorem 3, the observation that $E[I_i] = r(1 - \rho_i)/(1 - \rho)$ and several straightforward manipulations. Similarly, for $i \in G$, relation (22) follows from the following equalities: For $\mathrm{Re}\, s > 0$,

$$\tilde{W}_i^*(s) = \lim_{\rho\uparrow 1}\frac{s(1 - \rho)}{s(1 - \rho) - \lambda_i(1 - B_i^*(s(1 - \rho)))} \cdot \lim_{\rho\uparrow 1}\frac{C_i^*(\lambda_i(1 - B_i^*(s(1 - \rho)))) - C_i^*(s(1 - \rho))}{s(1 - \rho)E[C_i]} \tag{26}$$

$$= \frac{1}{(1-\hat{\rho}_i)rs}\left\{\left(\frac{\mu_i}{\mu_i+s\hat{\rho}_i}\right)^{\alpha}-\left(\frac{\mu_i}{\mu_i+s}\right)^{\alpha}\right\} \quad (i \in G). \tag{27}$$

The first equality follows from (3) and the definition of $\tilde{W}_i(s)$ in (24). The second equality follows from Theorem 4, the fact that $E[C_i] = r/(1-\rho)$, and several straightforward manipulations. This completes the proof of Theorem 5. □

Theorem 5 gives a closed-form expression for the LST of the steady-state waiting-time distributions at each of the queues, providing an explicit characterization of the complete distribution of the delay incurred at each of the queues (under heavy-traffic scalings). To the best of the author's knowledge, similar expressions have not been observed in the literature in the general parameter setting of the model.

Remark 3.1. The results in this paper are in line with a number of known results in special cases. For the case $N = 2$ and $E = \{1, 2\}$, the results correspond to those in [10] for the case of zero switch-over times and to [11] for non-zero switch-over times. Kroese [19] considers continuous polling systems in heavy traffic, which corresponds to the present model for the case of a fully symmetric system with $N \to \infty$. Note that in that case, we have $\rho_i \to 0$ $(i = 1, \ldots, N)$ and $\delta \to 1$ (regardless of the service policies). According to Theorems 3 and 4, both the (scaled) cycle times and intervisit times (which coincide for continuous polling) converge, for $\rho \uparrow 1$, to a gamma-distribution with shape parameter $\alpha = rb^{(1)}/b^{(2)}$ and rate parameter $\mu = b^{(1)}/b^{(2)}$, which is in line with the results in [19]. For systems with large (deterministic) switch-over times, it is shown in [32] that the C_i/r and I_i/r converge to a deterministic distribution when r tends to infinity (even for stable polling systems). It is readily verified from Theorems 3 and 4 that both $\tilde{I}_i/r$ and $\tilde{C}_i/r$ converge to a deterministic distribution when r tends to infinity, which is readily shown (by using relations (2) and (3)) to imply that $\tilde{W}_i/r$ congerves to a uniform distribution when r grows without bound, which is in line with the results in [32].

Remark 3.2. In many applications, the switch-over times are negligible. In this context, Theorem 5 implies that for the case of zero switch-over times, the LST of the waiting-time distribution at Q_i is given by the following expressions: For $\mathrm{Re}\, s > 0$,

$$\lim_{r\downarrow 0}\tilde{W}_i^*(s) = \frac{\delta}{(1-\hat{\rho}_i)s}\frac{b^{(1)}}{b^{(2)}}\log\left(\frac{\mu_i+s}{\mu_i}\right) \quad (i \in E), \tag{28}$$

$$\lim_{r\downarrow 0}\tilde{W}_i^*(s) = \frac{\delta}{(1-\hat{\rho}_i)s}\frac{b^{(1)}}{b^{(2)}}\log\left(\frac{\mu_i+s}{\mu_i+s\hat{\rho}_i}\right) \quad (i \in G), \tag{29}$$

where $\log(\cdot)$ is an inverse function of the (complex) function $f(z) := \exp(z)$. These results follows directly from Eqs. (21)–(23) and several straightforward manipulations. We refer to [2,24] for a detailed discussion of the relation between the delay in polling systems with and without switch-over times.

Remark 3.3. Federgruen and Katalan [15] propose a numerical method to approximate the queue-length and waiting-time distributions in polling models with exhaustive and gated service. The method is based on fitting the first few moments of the intervisit times and cycle times to their exact values, which can be determined numerically by means of the Descendant Set Approach (DSA) [18]. Interestingly, numerical

experiments in [15] show that highly accurate approximations are obtained by fitting the first two moments of the intervisit times and cycle times only. In this context, recall that Theorems 3 and 4 show that both the (scaled) cycle times and intervisit times converge to a gamma-distribution, and hence, are completely determined by their first two moments. In other words, the two-moment fitting of the cycle times and intervisit times are asymptotically exact when the load tends to unity. In this way, the results in this paper form a theoretical basis for the observed accuracy of the two-moment fitting of the cycle times and intervisit times in [15].

Remark 3.4. For stable systems (i.e., with $\rho < 1$), the tail probabilities of the waiting-times distribution can be computed by means of the numerical transform inversion technique (NTI) discussed in [8]. This approach is highly effective for lightly and medium-loaded systems. However, the computation times may increase dramatically when ρ is close to 1. Alternatively, in the limiting case $\rho \uparrow 1$, the tail probabilities can be obtained almost instantaneously by applying NTI directly to the exact expressions in Theorem 5. In this way, the limiting distribution may be used as a fast-to-evaluate approximation of the waiting-time distribution in stable systems with ρ close to 1 (see Section 5 for a discussion of the accuracy of the approximation). In this perspective, the applicability of the approach in [8] (for light and medium load) and the approach of combining Theorem 5 with NTI (for heavy load) is complementary.

4. Asymptotic properties

The results obtained in Section 4 reveal a number of properties of the heavy-traffic behavior of polling systems, in a general parameter setting. In Section 4.1 we discuss a number of insensitivity properties of the asymptotic delay distributions with respect to specific system parameters. In Section 4.2 we discuss properties of the asymptotic tail behavior of the limiting waiting-time distributions.

4.1. Insensitivity

Theorem 5 reveals a variety of properties about the dependence of the limiting delay distribution with respect to the system parameters.

Property 1. *For $i = 1, \ldots, N$, the distribution of $\tilde{W}_i$*
- *is independent of the visit order,*
- *depends on the switch-over time distributions only through r, i.e., the total expected switch-over time per cycle,*
- *is independent of the lth moment of the service-time distributions at the queues for $l > 2$,*
- *depends on the second moments of the service-time distributions only through $b^{(2)}$, i.e., the second moment of the service time of an "arbitrary" customer.*

Property 1 is known to be not generally valid for stable systems (i.e., for $\rho < 1$), where the visit order, the individual switch-over time distributions and the higher moments of the service-time distributions *do* have an impact on the distribution of the waiting-times. Hence, Property 1 shows that the influence of these parameters on the waiting-time distributions *vanishes* when the load tends to unity, and as such can be viewed as *lower-order* effects in heavy traffic.

4.2. Asymptotic decay rate

Define the *asymptotic decay rate* of the probability distribution of a real-valued random variable Y by

$$\eta(Y) := \lim_{x\to\infty} -\frac{\ln(\Pr\{Y > x\})}{x}, \tag{30}$$

where the $\ln(\cdot)$ is the inverse of the (real-valued) function $f(x) := e^x$. The following result gives the asymptotic decay rate of the (scaled) waiting times in the limiting case.

Property 2. *For $i = 1, \ldots, N$, the asymptotic decay rate of $\tilde{W}_i$ is given by*

$$\eta(\tilde{W}_i) = \frac{\delta}{1-\hat{\rho}_i}\frac{b^{(1)}}{b^{(2)}} \quad (i \in E), \qquad \eta(\tilde{W}_i) = \delta\frac{b^{(1)}}{b^{(2)}} \quad (i \in G). \tag{31}$$

To show the validity of Property 2, note that for $i \in E$, Theorem 3 implies that $\eta(\tilde{I}_i) = \mu_i$ (defined in (17)), i.e., the asymptotic decay rate of the scaled intervisit times $\tilde{I}_i$ equals to μ_i. From Eq. (2) it follows that the distribution of $\tilde{W}_i$ is the forward recurrence-time distribution of $\tilde{I}_i$. The latter is readily verified to have the same asymptotic decay rate as $\tilde{I}_i$, which shows the validity of the result. Similar arguments may be used to show the validity of the result for $i \in G$.

Property 2 implies the following properties of the asymptotic decay rate of the scaled waiting times when the load tends to unity.

Property 3. *For $i = 1, \ldots, N$, the asymptotic decay rate of the distribution of $\tilde{W}_i$*
- *decreases as the variability in the service times (i.e., $b^{(2)}$) increases,*
- *decreases as $E \to E + \{j\}$ for some $j \in G$, $j \neq i$.*
- *increases as $E \to E - \{j\}$ for some $j \in E$, $j \neq i$.*

In other words, part 1 implies that increasing the variability of the service times implies that the tails of the waiting times tend to become heavier. Part 2 implies that if Q_j receives exhaustive instead of gated service for some $j \neq i$, then the tails of the delay at Q_i become heavier. Part 3 implies that if Q_j receives gated instead of exhaustive service for some $j \neq i$, then the tails of the delay at Q_i become thinner.

Remark 4.1. A monotonicity property similar to Property 3 is not necessarily true when the service discipline at Q_i itself is modified. To this end, consider for example a three-queue model with $\hat{\rho}_1 = \hat{\rho}_2 = 1/10$, $\hat{\rho}_3 = 8/10$. It is readily verified by using Property 2 that if $G = \{1, 2, 3\}$, $E = \emptyset$, then replacing the service discipline at Q_1 by the exhaustive service leads to an increase of $\eta(\tilde{W}_1)$. On the other hand, if $G = \{1, 3\}$, $E = \{2\}$, then giving Q_1 exhaustive instead of gated service leads is readily verified to lead to a decrease of $\eta(\tilde{W}_1)$. Thus, a monotonicity property similar to Property 3 is not necessarily true when the service policy at Q_i itself is modified.

Remark 4.2. Choudhury and Whitt [8] study the asymptotic tail bahavior of the waiting-time distribution for (stable) polling systems with gated and exhaustive service at the queues. Based on the use of numerical transform inversion (NTI), they numerically calculate the asymptotic decay parameter $\eta(W_i)$ (among other parameters), which is the dominant singularity (i.e., the singularity closest to the origin) of the LST of the waiting-time distribution. The discussion of the numerical results in [8] leads to a number of interesting observations, which are found empirically but are unproven: (a) comparing fully symmetric systems with either gated service at all queues and systems with exhaustive at all

queues, the asymptotic decay rate is smaller for the case of exhaustive service, (b) the dominant singularity of $W_i^*(\cdot)$ is typically located "very near" $-\eta(W_i)$, (c) for systems with zero switch-over times, logarithmic singularities are observed in several cases, and (d) the asymptotic decay rate is the same for systems with and without switch-over times (provided that the switch-over time contribution does not dominate the zero-switch-over time contribution). Each of these observations can be shown to be asymptotically exact in the limiting case $\rho \uparrow 1$. (Note that in our analysis we consider the scaled delay $(1-\rho)W_i$, rather than W_i. It is readily verified from the definition in (30) that the corresponding asymptotic decay rates are related as $\eta((1-\rho)W_i) = \eta(W_i)/(1-\rho)$.) More precisely, in the limiting case $\rho \uparrow 1$, observation (a) follows from Property 3. Similarly, the asymptotic correctness of (b) follows from Property 2. Observation (c) follows directly from Eq. (22), and finally, (d) follows directly from Property 2. Thus, the results in this paper show that the empirical observations (a)–(d) are asymptotically exact when $\rho \uparrow 1$, which provides a theoretical basis for the validity of the empirical observations in [8].

Remark 4.3. An interesting observation is made by Duffield [13], who studies fully symmetric stable polling systems with gated service, and shows that "the dominant effect on the tail of the waiting-time distribution is from the service-time distribution or the switch-over times distribution, whichever has the heavier tail". Property 1, however, shows that the asymptotic decay rate of the scaled delay (i.e., $(1-\rho)W_i$) does not depend on the switch-over time distributions at all (assuming the moments are finite), when the system is close to saturation. Apparently, the impact of the tail distribution of the switch-over times on the asymptotic decay rate of the waiting-time distribution are of "lower order" when the system tends to saturate.

Remark 4.4. Despite the fact that monotonicity results, like those in Property 1, are often quite intuitive, they may not be trivial to prove. In the literature only a few monotonicity properties have been proven. Levy et al. [22] show pathwise monotonicity properties of the total amount of unfinished work in the system. Van der Mei and Levy [27,28] show monotonicity of the expected delay at the individual queues with respect to the so-called exhaustiveness of the service policies (in heavy traffic). The exhaustiveness of the service policy at Q_i, denoted by f_i, is defined as one minus the ratio between the average number of customers at Q_i at a departure epoch at Q_i and the average number of customers at a polling instant at Q_i. For example, it is readily verified that for $i \in E$ we have $f_i = 1$, and that for $i \in G$ we have $f_i = 1-\rho_i$. The reader is referred to [28] for more details on the notion of exhaustiveness. Borst et al. [3] obtain semi-conjectured monotonicity properties regarding the expected delay in polling systems with K-limited service.

To the best of the author's knowledge, Properties 1–3, and the corresponding remarks discussed above, have not been observed before in the literature, and provide a variety new and useful insights into the behavior of heavily-loaded polling systems.

5. Approximation

Theorem 5 suggests the following approximation for the waiting-time distributions in stable polling systems: For $i = 1, \ldots, N$, $\rho < 1$,

$$\Pr\{W_i < x\} \approx \Pr\{\tilde{W}_i < x(1-\rho)\} \quad (x > 0). \tag{32}$$

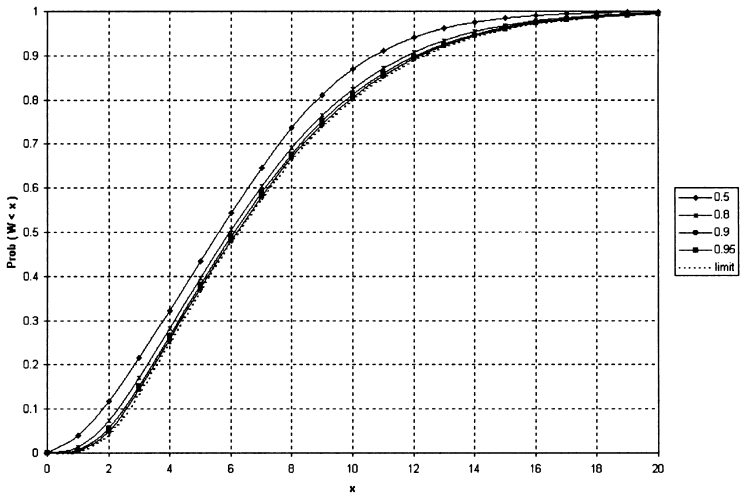

Fig. 1. Cumulative distribution function of $(1-\rho)W_1$ for different values of the load in a fully symmetric 5-queue model.

The right-hand side of (32) can be calculated almost instantaneously by applying the NTI directly to the Eqs. (21) and (22) and therefore, can be used as a fast-to-evaluate approximation of the waiting-time distributions for stable polling systems with ρ close to 1 (see also Remark 3.4). In this context, Theorem 5 implies that the approximation is asymptotically exact for $\rho \uparrow 1$.

To assess the accuracy of the approximation in (28), in terms of "How high should the load be for the approximation to be accurate?", we consider a fully symmetric 5-queue model with gated service at all queues, with exponential service times with mean 1 and with deterministic switch-over times with mean 2 between all queues. We used the approach in [8] to compute the tail probabilities of the delay for different values of the load. To compute the tail probabilities in the limiting case, we applied NTI directly to expression (22) in Theorem 5. Fig. 1 shows the cumulative probability distribution of $(1-\rho)W_1$ for different values of ρ, and for the limiting case $\rho \uparrow 1$ (indicated as "limit"). The results in Fig. 1 demonstrate that the waiting-time distribution of $(1-\rho)W_1$ indeed converges to the limiting distribution when the load tends to unity, as expected on the basis of Theorem 5. Moreover, Fig. 1 shows that the distribution of $(1-\rho)W_1$ converges to its limiting distribution rather quickly when $\rho \uparrow 1$, and is (visually) "close" to the limiting distribution for load-values, say, 80% or more. In other words, when the load exceeds 80%, the approximation (32) is fairly accurate. This observation demonstrates the applicability of the asymptotic results for practical heavy-traffic scenarios. (The assessment of the accuracy of the approximation for extremely small tail probabilities, which is important in some applications but not covered by Fig. 1, is beyond the scope of this paper and is left as a topic for further research.)

To investigate the accuracy of the approximation (32) for highly asymmetric systems, we also consider a 10-queue model with the following parameters: The ratios between the arrival rates are 1:1:10:1:1:1:10:1:10:1. The service times at Q_3 and Q_{10} are exponentially distributed with means 1 and 10, respectively, whereas the service times at Q_4 and Q_8 are deterministically distributed with respective means 10 and 25 and all other service time are deterministic with mean 1. The switch-over times from Q_3 to Q_4 are deterministic with mean 1, and all other switch-over times are 0. $E = \{3, 4, 5, 10\}$ and $G = \{1, 2, 6, 7, 8, 9\}$. Clearly, the model is highly asymmetric in the arrival rates, service times and service policies. Fig. 2

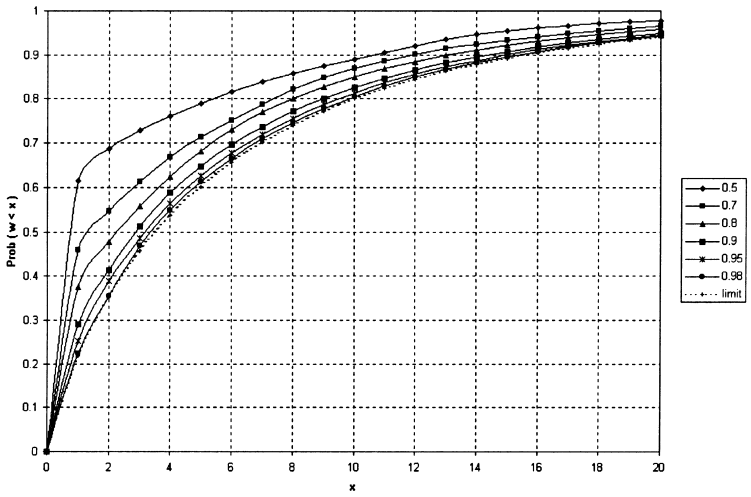

Fig. 2. Cumulative distribution function of $(1-\rho)W_1$ for different values of the load in a highly asymmetric 10-queue model.

shows the cumulative probability distribution of $(1-\rho)W_1$, for different values of ρ and for the limiting case $\rho \uparrow 1$ ("limit"). shows that the waiting-time distributions indeed converge to their limiting values when the load tends to unity, which confirms the validity of Theorem 5. To check the accuracy of the approximation (32) for this model, Fig. 2 shows that the convergence to the limiting distribution is somewhat slower than in the model in Fig. 1. This was to be expected because of the strong asymmetry in the model parameters. Nonetheless, the limiting distribution still "closely" resembles the waiting-time distribution when the load is, say, 90% or more. We emphasize that the computation times required to obtain accurate calculations of the tail probabilities of the delay for the higher values of the load (e.g., $\rho = 0.98$ and higher) are on the order of minutes each (on a modern work station), whereas the approximations in (32) can be obtained almost instantaneously by applying NTI directly to Theorem 5. These observations demonstrate the usefulness of the asymptotic results in practical heavy-traffic scenarios.

6. Topics for further research

The results in this paper suggest simple and fast approximations for the waiting-time distributions in stable polling systems, which are (visually) "accurate" when the load is about 80–90% or more. However, in some applications (e.g., in telecommunication systems) the most important performance measures are very small tail probabilities, which are not covered in the numerical examples discussed in Section 5. In Section 4.2 we obtained expressions for the asymptotic decay rate of the delay under heavy-traffic assumptions. A more detailed analysis of the asymptotic tail behavior of the delay and an assessment of the accuracy of the approximation in (32) for very small tail probabilities are left as a topic for further research.

Theorem 5 shows that the limiting waiting-time distribution depends on the system parameters only through a few aggregated parameters (namely, $r, b^{(1)}, b^{(2)}, \delta$ and $\hat{\rho}_i$ $(i = 1, \ldots, N)$). This observation greatly simplifies optimization of the system performance with respect to the configurable parameters (e.g., service policies, visit order). Optimization of the system performance in heavy traffic is a topic

for further research. Feasibility problems of the form "Can the system be operated such that $\Pr\{W_i > x_i\} < \alpha_i$ $(i = 1, \ldots, N)$?" are of main practical importance. Here, the values of x_i and α_i are typically specified by the users, and typical decision variables are the choice of the service policies and the visit order. To the best of the author's knowledge, this type of problems has not been studied before in the polling literature. The results in this paper open possibilities for obtaining (approximative) solutions for solving feasibility problems under heavy-traffic assumptions.

Recent studies have revealed that in many applications the arrival processes are non-Poisson. It is a topic for further research to obtain expressions for the waiting-time distribution in heavy-traffic under non-Poisson arrival processes. In this perspective, encouraging results are obtained by Coffman et al. [10,11], who obtain simple expressions for the heavy-traffic limit of the waiting-time distribution for a class of non-Poisson arrival processes.

In this paper it is assumed that all moment of the service times and switch-over times, and hence of the cycle times and intervisit times, are finite. However, the limiting waiting-time distributions depend on the first and first two moments of the switch-over times and service times, respectively (see Property 1). This observation suggests that the heavy-traffic results in this paper may be obtained under weaker assumptions about the finiteness of the moments of the service times and switch-over times. Derivation of such results is left as a topic for further research.

Another interesting topic for further research is to analyze the impact of heavy-tailed service-time and switch-over time distributions on the distributions of the delay in heavy traffic. Boxma et al. [7] use the theory of so-called regularly varying functions and show that in a model with gated and exhaustive service at each queue the waiting-time distribution is regularly varying of an index one plus the index of the heaviest service-time or switch-over time distribution. Boxma and Cohen [6] obtain the heavy-traffic limiting distribution for the single-server queue with regularly varying service-time distributions. An interesting observation in [6] is that the scaling factor needed to obtain a proper limiting distribution is generally not equal to $1 - \rho$, as is the case for the polling systems (with finite moments) discussed in this paper. The identification of the proper scaling factor and the limiting distribution for polling models with heavy-tailed service-time and switch-over time distributions is left as a topic for further research.

Acknowledgements

The author wishes to thank G.L. Choudhury for making the computer program corresponding to [8] available for performing the numerical experiments, and D.P. Heyman for useful discussions about the Method of Moments used in the derivation of the results. The results presented in this paper generalize the results obtained in joint work with H. Levy. The author is indebted to him for many interesting discussions on the subject. A partial and preliminary version of this paper appeared in [33].

References

[1] J.P.C. Blanc, Performance analysis and optimization with the power-series algorithm, in: L. Donatiello, R. Nelson (Eds.), Performance Evaluation of Computer and Communication Systems, Springer, Berlin, 1993, pp. 53–80.
[2] S.C. Borst, O.J. Boxma, Polling models with and without switchover times, Oper. Res. 45 (1997) 536–543.

[3] S.C. Borst, O.J. Boxma, H. Levy, The use of service limits for efficient operation of multistation single-medium communication systems, IEEE Trans. Networks 3 (1995) 602–612.
[4] O.J. Boxma, W.P. Groenendijk, Pseudo-conservation laws in cyclic-service systems, J. Appl. Probab. 24 (1987) 949–964.
[5] O.J. Boxma, W.P. Groenendijk, Two queues with alternating service and switching times, in: O.J. Boxma, R. Syski (Eds.), Queueing Theory and its Applications —Liber Amicorum for J.W. Cohen, North-Holland, Amsterdam, 1988, pp. 261–282.
[6] O.J. Boxma, J.W. Cohen, The single server queue: heavy tails and heavy traffic, in: K. Park, W. Willinger (Eds.), Self-Similar Network Traffic and Performance Evaluation, Wiley, New York, in press.
[7] O.J. Boxma, Q. Deng, J.A.C. Resing, Polling systems with regularly varying service and/or switch-over times, Technical Report, COSOR 99–10.
[8] G. Choudhury, W. Whitt, Computing transient and steady state distributions in polling models by numerical transform inversion, Perf. Eval. 25 (1996) 267–292.
[9] K.L. Chung, A Course in Probability, 2nd ed., Academic Press, New York, 1974.
[10] E.G. Coffman, A.A. Puhalskii, M.I. Reiman, Polling systems with zero switch-over times: a heavy-traffic principle, Ann. Appl. Probab. 5 (1995) 681–719.
[11] E.G. Coffman, A.A. Puhalskii, M.I. Reiman, Polling systems in heavy-traffic: a Bessel process limit, Math. Oper. Res. 23 (1998) 257–304.
[12] J.W. Cohen, O.J. Boxma, The M/G/1 queue with alternating service formulated as a Rieman–Hilbert boundary value problem, in: F.J. Kylstra (Ed.), Performance 1981, North-Holland, Amsterdam, 1981, pp. 181–199.
[13] N.G. Duffield, Exponents for the tails of distributions in some polling models, Queueing Systems 26 (1997) 105–119.
[14] M. Eisenberg, Two queues with alternating service, SIAM J. Appl. Math. 36 (1979) 287–303.
[15] A. Federgruen, Z. Katalan, Approximating queue size and waiting-time distributions in general polling systems, Queueing Systems 18 (1994) 353–386.
[16] W. Feller, An Introduction to Probability Theory and its Applications, 2nd ed.,Wiley, New York, 1970.
[17] C. Fricker, M.R. Jaïbi, Monotonicity and stability of periodic polling models, Queueing Systems 15 (1994) 211–238.
[18] A.G. Konheim, H. Levy, M.M. Srinivasan, Descendant set: an efficient approach for the analysis of polling systems, IEEE Trans. Commun. 42 (1994) 1245–1253.
[19] D.P. Kroese, Heavy traffic analysis for continuous polling models, J. Appl. Probab. 34 (1997) 720–732.
[20] K.K. Leung, Cyclic service systems with probabilistically-limited service, IEEE J. Sel. Areas Commun. 9 (1991) 185–193.
[21] H. Levy, M. Sidi, Polling models: applications, modeling and optimization, IEEE Trans. Commun. 38 (1991) 1750–1760.
[22] H. Levy, M. Sidi, O.J. Boxma, Dominance relations in polling systems, Queueing Systems 6 (1990) 155–171.
[23] M.I. Reiman, L.M. Wein, Dynamic scheduling of a two-class queue with setups, Oper. Res. 46 (1998) 532–547.
[24] M.M. Srinivasan, S.C. Niu, R.B. Cooper, Relating polling models with zero and nonzero switchover times, Queueing Systems 19 (1995) 149–168.
[25] H. Takagi, Queueing analysis of polling models: an update, in: H. Takagi (Ed.), Stochastic Analysis of Computer and Communication Systems, North-Holland, Amsterdam, 1990, pp. 267–318.
[26] H. Takagi, Queueing analysis of polling models: progress in 1990–1994, in: J.H. Dshalalow (Ed.), Frontiers in Queueing: Models, Methods and Problems, CRC Press, Boca Raton, 1997, pp. 119–146.
[27] R.D. Van der Mei, H. Levy, Expected delay in polling systems in heavy traffic, Adv. Appl. Probab. 30 (1998) 586–602.
[28] R.D. Van der Mei, H. Levy, Polling systems in heavy traffic: exhaustiveness of the service policies, Queueing Systems 27 (1997) 227–250.
[29] R.D. Van der Mei, Polling systems in heavy traffic: higher moments of the delay, Queueing Systems 31 (1999) 265–294.
[30] R.D. Van der Mei, Polling systems with switch-over times under heavy load: moments of delay, Queueing Systems, to appear.
[31] R.D. Van der Mei, Polling systems with periodic server routing in heavy traffic, Stochast. Models 15 (1999) 273–297.
[32] R.D. Van der Mei, Delay in polling systems with large switch-over times, J. Appl. Probab. 36 (1999) 232–243.
[33] R.D. Van der Mei, Waiting-time distributions in polling systems in heavy traffic, in: P. Key, D. Smith (Eds.), Teletraffic Engineering in a Competitive World, Elsevier, Amsterdam, 1999, pp. 325–334.
[34] J.A. Weststrate, R.D. Van der Mei, Waiting times in a two-queue model with exhaustive and Bernoulli service, Z.O.R. Math. Meth. Oper. Res. 40 (1994) 289–303.

Robert D. van der Mei (born 1966) received his M.Sc. degrees in Mathematics and in Decision Sciences from the Free University of Amsterdam, Netherlands, and his Ph.D. degree in Operations Research from Tilburg University, Netherlands. In the meantime, he worked for several years in industry for the Center for Quantitative Methods (formerly Philips), the Royal Dutch Shell Laboratories, and KPN Research, Netherlands. After working as a visiting Scholar at Rutgers University, USA, and Columbia University, USA, in 1996 he joined AT&T Labs, Department of Network Design and Performance Analysis. In 1999, he returned to his home country to rejoin KPN Research, Department of Planning, Performance and Reliability, where he now leads the Quality of Service group. Dr. van der Mei is the technical chair of the SPIE Conference of Performance and Control of Network Systems in 1999. His research interests include performance analysis of computer and communication networks, Internet performance, queueing analysis, and stochastic models.