Annals of Operations Research 113, 155–173, 2002

Waiting-Time Distributions in Polling Systems with Simultaneous Batch Arrivals

R.D. VAN DER MEI
KPN Research, Quality of Service Control, Leidschendam, Netherlands and
Vrije Universiteit, Mathematics and Computer Science, Amsterdam, Netherlands

Abstract. We study the delay in asymmetric cyclic polling models with general mixtures of gated and exhaustive service, with generally distributed service times and switch-over times, and in which batches of customers may arrive simultaneously at the different queues. We show that $(1-\rho)X_i$ converges to a gamma distribution with known parameters as the offered load ρ tends to unity, where X_i is the steady-state length of queue i at an arbitrary polling instant at that queue. The result is shown to lead to closed-form expressions for the Laplace–Stieltjes transform (LST) of the waiting-time distributions at each of the queues (under proper scalings), in a general parameter setting. The results show explicitly how the distribution of the delay depends on the system parameters, and in particular, on the simultaneity of the arrivals. The results also suggest simple and fast approximations for the tail probabilities and the moments of the delay in stable polling systems, explicitly capturing the impact of the correlation structure in the arrival processes. Numerical experiments indicate that the approximations are accurate for medium and heavily loaded systems.

Keywords: polling, simultaneous arrivals, batch processes, heavy traffic, waiting-time distribution

1. Introduction

Polling systems are multi-queue systems in which a single server visits the queues in some order to serve the customers waiting at the queues, typically incurring some amount of switch-over time to proceed from one queue to the next. Polling models find a wide variety of applications in which processing power (e.g., CPU, bandwidth) is shared among different types of users. During the past few decades polling models have received much attention in the literature. In the vast majority of the papers on polling models it is assumed that the arrival processes at the different queues are independent unit Poisson processes, i.e., where exactly one customer arrives at a particular queue at a time. However, in many applications customers arrive in batches and batches of customers may arrive at different queues simultaneously. Neglecting the correlation structure in the arrival processes may lead to strongly erroneous performance predictions, and consequently, to improper decisions about the operation of the system, in particular when the system load is significant. These observations raise the need to get a better understanding of the impact of correlations between the arrival processes on the delay incurred at each of the queues. This paper, in which we analyze the delay in

polling models with a specific class of correlation structures in the arrival processes, is a significant first step in that direction.

The possibility of simultaneous batch arrivals strongly enhances the modeling and analysis capabilities of polling models. Many examples are found in computer–communication systems. Consider, for example, a Web server that needs to respond to numerous document-retrieval requests initiated by the end users. A Web document generally consists of a number of files (e.g., pieces of text, in-line images, audio files), each of which generates a separate file-retrieval request to the Web server. The server typically implements a scheduling mechanism to determine the order in which the incoming file-retrieval requests are handled. In many implementations the incoming file-retrieval requests are buffered in separate queues, e.g., depending on the size of the requested document. The Web server continuously polls the different queues to check for pending file-retrieval request to be executed. In this application, the server represents the Web server, the customers represent individual file-retrieval requests, and each document-retrieval request is represented by a joint batch of customers. Another example is a Local Area Network (LAN) in which the right for transmission, represented by a so-called token, is circulated among the users. If a user wants to initiate a transaction over the network, a transaction request is placed into an output buffer. However, the amount of data that can be sent over the network at a time is limited, so that fragmentation may be needed. When the user gets the right for transmission, a number of (possibly all) fragments are transmitted over the network. In this application, the server represents the right for transmission (token) and the queues represent the output buffers. The customers represent the fragments, and as such, a transmission initiation request placed by a user is represented by a (not necessarily joint) batch arrival of customers. Applications are also found in the area of flexible manufacturing and production systems. Consider, for example, a production facility that can produce different types of products, but only one product type at a time. Wholesalers from time to time place joint replenishment orders for different products. Incoming orders for a given product that can not be processed immediately are placed into a buffer of pending orders for that product. After a number of (possibly all) outstanding orders for a specific product have been processed, the production facility is installed to process the next type of products. In this way, the facility "visits" the buffers of pending orders for the different product types in a round-robin fashion. In this example, the server represents the production facility, the customers represent replenishment orders of one unit of a product and joint replenishment orders are represented by simultaneous batch arrivals.

In the literature, polling systems with correlated arrivals have received only little attention. Levy and Sidi [15] study polling models with simultaneous batch arrivals. For models with gated or exhaustive service, they derive a set of linear equations for the expected delay at each of the queues. They also provide a pseudo-conservation law for the system, i.e., an exact expression for a specific weighted sum of the expected waiting times at the different queues. The moments of the delay in the model studied in [15] can also be obtained by means of the Descendant Set Approach (DSA), an iterative technique based on the concept of descendant sets [9] (see section 3 for more details). Boxma and

Groenendijk [3] derive a pseudo-conservation law for discrete-time polling models with batch arrivals. Langaris [12–14] studies several variants of polling models with correlated batch arrivals and with so-called retrial customers. For this type of models, the expected delay can be obtained by solving sets of linear equations. Recently, polling models in heavy traffic have received attention in the literature. For a two-queue model with exhaustive service and independent renewal arrival processes, Coffman et al. [6,7] use the theory of diffusion processes to derive expressions for the joint workload distribution and the waiting-time distributions under heavy traffic assumptions. For models with independent unit Poisson arrivals, Kudoh et al. [11] give explicit expressions for the second moment of the waiting time in fully symmetric systems with gated or exhaustive service at each queue for models with two, three and four queues. They also give conjectures for the heavy-traffic limits of the first two moments of the waiting times for systems with an arbitrary number of queues. Kroese [10] studies continuous polling systems in heavy traffic with unit Poisson arrivals on a ring and shows that the steady-state number of customers at each queue has approximately a gamma distribution.

In this paper we consider an asymmetric cyclic polling model with general mixtures of gated and exhaustive service and general service-time and switch-over time distributions. The correlation structure in the arrival processes is modeled as follows. Arrival points are generated according to a Poisson process. At each arrival point, batches of customers may arrive *simultaneously* at the different at the queues, according to a general joint batch-size distribution. We study the heavy-traffic behavior of the distribution of the delay at each of the queues. We derive closed-form expressions for the Laplace–Stieltjes transform (LST), and hence also the moments, of the limiting distribution of the delay when the load tends to unity (under proper scalings), in a general parameter setting. The expressions explicitly reveal how the distribution of the delay depends on the system parameters, and in particular, how the correlation structure between the arrival processes impacts the delay incurred at each of the queues. In addition, the results reveal a variety of asymptotic insensitivity properties of the delay with respect to specific system parameters. These observations give new fundamental insights into the heavy-traffic behavior of polling systems that have not been observed before. Finally, the results suggest simple and fast approximations for the tail probabilities and the moments of the delay in stable polling systems. Numerical results indicate that the approximations are accurate for medium and heavily loaded systems, which demonstrates the practical usefulness of the results.

The motivation for this paper is two-fold. First, we have a queueing-theoretical interest in explicitly quantifying the impact of correlations between the arrival processes at the different queues on the delay incurred at each of the queues. The results presented in this paper, where we consider a specific class of correlation structures, form a significant step in that direction. Second, in many applications of polling models the arrival processes at the different queues are correlated (see the examples above). In view of those applications it is important to be able to predict the queueing behavior accurately, in particular when the system load is significant. However, the effectiveness of the existing numerical techniques (e.g., [4,9]) tends to degrade strongly when the system is

heavily loaded. This raises the need for the development of simple and fast approximations for the delay incurred at each of the queues, explicitly capturing the impact of correlated arrivals.

The results presented in this paper generalize the results in [19–22]. In [22] we derived closed-form expressions for the first moments of the delay in the model with simultaneous batch arrivals considered in the present paper (see section 2 for more details). In [19–21] we derived closed-form expressions for LST of the waiting-time distribution for the special case of independent unit Poisson arrival processes. The methodology developed in those papers, based on exploring the so-called Descendant Set Approach (DSA), has been shown to be highly effective for deriving heavy-traffic results. Therefore, in this paper we adopt the same methodological steps to derive heavy-traffic results for the model under consideration. We emphaszie that the contribution of the present paper is fairly limited from a purely methodological point of view. However, the results presented in this paper contain a significant number of new elements. First, we obtain closed-form expressions for the LST of the (scaled) waiting-time distribution at each of the queues in heavy traffic, in a general parameter setting. As a by-product, we also obtain closed-form expressions for the moments of the delay. The results *explicitly* show how the delay figures depend on the system parameters, and particular, on the correlation structure between the arrival processes. In this context, we emphasize that exact results on the individual delay figures in polling models are scarce, making the results particularly interesting from a queueing-theoretical perspective. Second, the results lead to a number of asymptotic properties of the waiting-time distributions with respect to the system parameters, and in particular the simultaneity of the arrivals. These results provide new fundamental insights in the behavior of polling models. Third, the results lead to simple and fast approximations for the tail probabilities and the higher moments of the delay in stable polling systems, explicitly capturing the impact of the simultaneity of the arrivals. The approximations are particularly accurate when the system is heavily loaded (see section 6), whereas the efficiency of the existing numerical techniques to obtain the higher moments and the tail probabilities of the delay tends to degrade significantly when the system is heavily loaded. In this way, the approximations proposed in the present paper *complement* the applicability of the existing numerical techniques. To summarize, although the methodological contribution is rather limited, the added value of the present paper is evident.

The remainder of this paper is organized as follows. In section 2 the model is described. In section 3 we give some preliminary results and discuss the use of the Descendant Set Approach (DSA) for the present model. In section 4 we derive closed-form expressions for the LST of the distribution of the delay in heavy traffic, based on the use of the DSA. In section 5 a variety of asymptotic insensitivity properties is discussed. In section 6 we propose and test new, simple and fast approximations for the tail probabilities and the moments of the delay in stable polling systems. Finally, in section 7 we address a number of topics for further research.

2. Model

Consider a system consisting of $N \geqslant 2$ infinite-buffer stations $Q_1, \ldots, Q_N$ that are served by a single server that visits and serves the queues in cyclic order. Arrival points are generated according a Poisson process with rate λ. At each arrival point, a random batch of size $\underline{K} = (K_1, \ldots, K_N)$ arrives at the queues, where K_i stands for the number of customers arriving at Q_i at an arrival point. The random vector $\underline{K}$ is assumed to be independent of previous or future arrival points. Denote the joint batch-size distribution by $\pi(k_1, \ldots, k_N) := \text{Prob}\{K_1 = k_1, \ldots, K_N = k_N\}$, and denote the corresponding probability generating function (PGF) of $\underline{K}$ by $K^*(\underline{z})$. The PGF of the marginal batch-size distribution at Q_i is denoted by $K_i^*(z) := \underline{K}^*(1, \ldots, 1, z, 1, \ldots, 1)$, $|z| \leqslant 1$, where the z occurs at the ith entry. Denote the arrival rate at Q_i by $\lambda_i := \lambda E[K_i]$, and let $K_{i,i} := E[K_i(K_i - 1)]$ for $i = 1, \ldots, N$ and $K_{i,j} := E[K_i K_j]$ for $i \neq j$. Denote the total arrival rate by $\Lambda := \sum_{i=1}^N \lambda_i$. The service time of a customer at Q_i is a random variable B_i with Laplace–Stieltjes transform (LST) $B_i^*(\cdot)$, and finite kth moment $b_i^{(k)}$, $k = 1, 2, \ldots$. The load offered to Q_i is defined by $\rho_i = \lambda_i b_i^{(1)}$, and the total offered load is equal to $\rho = \sum_{i=1}^N \rho_i$. Denote the kth moment of the service time of an arbitrary customer by $b^{(k)} := (1/\Lambda) \sum_{i=1}^N \lambda_i b_i^{(k)}$, $k = 1, 2$. A polling instant at Q_i is defined as the epochs at which the server arrives at a queue to serve customers waiting 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 an 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. The corresponding LST is denoted by $I_i^*(\cdot)$. Similarly, define the cycle time C_i to be the time between two successive polling instants at Q_i, and denote the corresponding LST by $C_i^*(\cdot)$. We consider two types of service disciplines: gated and exhaustive. 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 to Q_i. Under the exhaustive policy the server visits Q_i until it is empty. We allow general mixtures of exhaustive and gated service, but the service policy at each queue remains the same for all visits. Define $E := \{i\colon\ Q_i \text{ is served exhaustively}\}$ and $G := \{i\colon\ Q_i \text{ is served according to the gated policy}\}$. At each queue the customers are served on a FIFO basis. After completing service at Q_i the server immediately proceeds to Q_{i+1}, incurring a switch-over period whose duration is an independent random variable R_i, with LST $R_i^*(\cdot)$ and with mean r_i. Denote the first moment of the total switch-over time per cycle by $r = \sum_{i=1}^N r_i$. All interarrival times and service times are assumed to be mutually independent and independent of the state of the system. It is assumed that the system is stable (i.e., $\rho < 1$) and that the system is in steady state.

Let W_i be the steady-state delay incurred by an arbitrary customer at Q_i. We focus on the probability distribution of the delay in heavy traffic, under proper scalings. More precisely, we are interested in the probability distribution of

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

Here, the random variable $(1-\rho)W_i$ is considered as a function of ρ, where λ (i.e., the rate at which the arrival points occur) is variable, whereas the service-time and batch-size distributions remain fixed. The main result of this paper is the derivation of a closed-form expression for $\widetilde{W}_i^*(s) := E[\mathrm{e}^{-s\widetilde{W}_i}](\mathrm{Re}\, s > 0)$, the LST of $\widetilde{W}_i$.

For each variable that is a function of ρ, we use the hat-notation to indicate its value at $\rho = 1$. Let I_E denote the indicator function on the event E. The notation $\to_{\mathrm{d}}$ stands for convergence in distribution.

3. Preliminaries

Denote by X_i the number of customers present at Q_i at a polling instant at Q_i when the systen is in steady-state, and denote the corresponding PGF by $X_i^*(z) := E[z^{X_i}]$, $|z| \leqslant 1$, $i = 1, \ldots, N$. The waiting-time distribution at Q_i can be expressed in terms of the distribution of X_i as follows.

Lemma 1. For $i \in G$, $\mathrm{Re}\, s > 0$,

$$W_i^*(s) \tag{2}$$

$$= \frac{1-\rho}{r} \cdot \frac{X_i^*(K_i^*(B_i^*(s))) - X_i^*(1-s/\lambda)}{s - \lambda(1 - K_i^*(B_i^*(s)))} \cdot \frac{1 - K_i^*(B_i^*(s))}{E[K_i](1 - B_i^*(s))}. \tag{3}$$

For $i \in E$, $\mathrm{Re}\, s > 0$,

$$W_i^*(s) \tag{4}$$

$$= \frac{1-\rho}{r} \cdot \frac{1 - X_i^*(1-s/\lambda)}{s - \lambda(1 - K_i^*(B_i^*(s)))} \cdot \frac{1 - K_i^*(B_i^*(s))}{E[K_i](1 - B_i^*(s))}. \tag{5}$$

Proof. Consider a tagged customer T_i that arrives at Q_i in a joint batch B_{T_i} of customers, and let U_i be the number of customers in B_{T_i} that are served at Q_i before T_i. Then the waiting time of T_i can be expressed as

$$W_i = W_i^{(s)} + W_i^{(b)}, \tag{6}$$

where $W_i^{(s)}$ is the waiting time of the first customer in B_{T_i} and where $W_i^{(b)}$ is the total service time of all U_i customers in B_{T_i} that are served at Q_i before T_i. First, it is readily seen that $W_i^{(s)}$ is stochastically identical to the waiting time at Q_i in a system with where each batch of customers is considered as a single super-customer, where the LST of the (marginal) service-time distribution of a super-customer at Q_j is $K_j^*(B_j^*(s))$, $j = 1, \ldots, N$. The first two factors in (2)–(5) then follow directly from equations (4.32) and (5.45) in [17]. Second, the PGF of the number of customers in U_i is given by $(1 - K_i(z))/(E[K_i](1-z))$ (cf., e.g., [18, equation (3.8a)]), which implies that the LST of the total amount of waiting time caused by those U_i customers is $(1 - K_i^*(B_i^*(s)))/(E[K_i](1 - B_i^*(s)))$. The proof is completed by observing that $W_i^{(s)}$ and $W_i^{(b)}$ are mutually independent. □

Lemma 1 implies that the distribution of W_i is completely characterized by the distribution of X_i. The distribution of X_i, in turn, can be characterized by means of the Descendant Set Approach (DSA), exploring the branching structure of the evolution of the system (cf. [9]). The DSA is focused on the determination of the distribution of X_1 (without loss of generality), the number of customers at Q_1 present at an arbitrary polling instant P^* at Q_1, referred to as the reference point. Define a cycle as the elapsed time between two successive polling instants at Q_1. The key observation is that we can evaluate $X_1(P^*)$ by considering, recursively, contributions to X_1 from waiting customers at all queues of the past polling epochs, working backward from the reference point. Let $T_{i,c}$ be a customer served at Q_i during the cth cycle. Define the children set of $T_{i,c}$ to be the set of customers arriving during the service of $T_{i,c}$; the descendant set of $T_{i,c}$ is recursively defined to consist of $T_{i,c}$, its children and the descendants of its children. Let $A_{i,c}$ be the number of customers at Q_1 at the reference point (at Q_1) that are descendants of $T_{i,c}$, and let $A^*_{i,c}(\cdot)$ denote its PGF. In this way, $A_{i,c}$ can be viewed as the *contribution* of $T_{i,c}$ to $X_1(P^*)$. Denote by $R_{i,c}$ the switch-over time (from Q_i to the next) immediately after the visit starting at $P_{i,c}$. Let $S_{i,c}$ be the total contribution to $X_1(P^*)$ of all (original) customers that arrive in the system during $R_{i,c}$, and denote the corresponding PGF by $S^*_{i,c}(\cdot)$. Then X_1 can be expressed as the independent sum $X_1 = \sum_{i=1}^{N} \sum_{c=0}^{\infty} S_{i,c}$, or equivalently, for $|z| \leqslant 1$,

$$X_1^*(z) = \prod_{i=1}^{N} \prod_{c=0}^{\infty} S^*_{i,c}(z), \tag{7}$$

where for $i = 1, \ldots, N$, $c = 0, 1, \ldots$, $|z| \leqslant 1$,

$$S^*_{i,c}(z) \tag{8}$$

$$= R_i^*\big(\lambda - \lambda K^*\big(A^*_{1,c-1}(z), \ldots, A^*_{i,c-1}(z), A^*_{i+1,c}(z), \ldots, A^*_{N,c}(z)\big)\big). \tag{9}$$

The descendant set (DS) variables satisfy the following set of relations, based on the observation that the contribution to $X_1(P^*)$ of a tagged customer $T_{i,c}$ is equal to the total contribution to $X_1(P^*)$ of the children of $T_{i,c}$: for $i \in G$, $c = 0, 1, \ldots$, $|z| \leqslant 1$,

$$A^*_{i,c}(z) \tag{10}$$

$$= B_i^*\big(\lambda - \lambda K^*\big(A^*_{1,c-1}(z), \ldots, A^*_{i,c-1}(z), A^*_{i+1,c}(z), \ldots, A^*_{N,c}(z)\big)\big), \tag{11}$$

and for $i \in E$,

$$A^*_{i,c}(z) \tag{12}$$

$$= B_i^*\big(\lambda - \lambda K^*\big(A^*_{1,c-1}(z), \ldots, A^*_{i-1,c-1}(z), A^*_{i,c}(z), \ldots, A^*_{N,c}(z)\big)\big). \tag{13}$$

Since we focus on the number of customers at Q_1 at the reference point, the initial conditions are $A^*_{1,-1}(z) = z$ and $A^*_{i,-1}(z) = 1$, for $i = 2, \ldots, N$, $|z| \leqslant 1$. Notice that relations (7)–(13) give a complete, but not explicit, characterization of the probability distribution of X_1. This characterization is useful to obtain properties of the limiting

distribution of X_1. Using equations (2)–(5) these properties, in turn, will be used to derive properties of the distribution of $\widetilde{W}_i$ $(i = 1, \ldots, N)$, defined in (1).

4. Analysis

In this section we derive closed-form expressions for the LST of the waiting-time distributions at each of the queues. The derivation of the results proceeds along a number of steps, according to the DSA-based methodology developed in [19–21] for the case of independent unit Poisson arrivals. Define the kth factorial moment of the DSA-variables $A_{i,c}$ as follows: for $i = 1, \ldots, N$, $c = 0, 1, \ldots, k = 1, 2, \ldots,$

$$\alpha_{i,c}^{(k)} := E\big[A_{i,c}(A_{i,c} - 1) \cdots (A_{i,c} - k + 1)\big]. \tag{14}$$

To start the derivation of the results, we decompose the sequences of DSA-variables $\{\alpha_{i,c}^{(k)}, c = 0, 1, \ldots\}$ in a dominant and a recessive part, in a sense to be specified (theorem 1). The decomposition is then used to obtain an expression for the quantities $\tilde{h}_1^{(k)} := \lim_{\rho\uparrow 1} (1-\rho)^k \sum_{i=1}^N \sum_{c=0}^\infty \lambda_i \alpha_{i,c}^{(k)}$ (theorem 2). This result, in turn, is used to obtain explicit expressions for the limiting moments $\tilde{x}_i^{(k)} := \lim_{\rho\uparrow 1} (1-\rho)^k E[X_i^k]$, which are shown to match the moments of a gamma distribution with known parameters (theorem 3). Then, the so-called method of moments is applied to show that $(1-\rho)X_i \to_{\mathrm{d}} \Gamma_i$, where Γ_i is a gamma distribution with known parameters (theorem 4). Finally, lemma 1 then leads to closed-form expressions for the LST of $\widetilde{W}_i^*(s)$ (see (1)), which is the main result of the paper (theorem 5). The proofs of the various intermediate results proceed along similar lines as the proofs for the case of independent unit Poisson arrivals, elaborated in [19–21]. To avoid duplication of derivations, only a brief sketch of the proofs is given.

It is convenient to define the following quantities:

$$\delta := \frac{1}{2}\left(1 - \sum_{m\in E} \hat{\rho}_m^2 + \sum_{m\in G} \hat{\rho}_m^2\right), \tag{15}$$

and

$$\varphi := \frac{b^{(2)}}{b^{(1)}} + \hat{\lambda} \sum_{l=1}^N \sum_{m=1}^N b_l^{(1)} b_m^{(1)} K_{l,m}. \tag{16}$$

We are now ready to focus on the heavy-traffic behavior of the sequences $\{\alpha_{i,c}^{(k)}, c = 0, 1, \ldots\}$. To start, we obtain the following result for the case $k = 1$.

Lemma 2. For $i = 1, \ldots, N$, $c = 0, 1, \ldots,$ we can write

$$\alpha_{i,c}^{(1)} = \xi^{c+1} v_i w + s_{i,c}, \tag{17}$$

where

(1) $\xi < 1$ if and only if $\rho < 1$; $\xi = 1$ if and only if $\rho = 1$,

(2) $\lim_{\rho\uparrow 1}\xi = 1$,

(3) $\hat{v}_i := b_i^{(1)}/\delta \quad (i = 1, \ldots, N)$,

(4) $\hat{w} := \lambda_1(1 - \hat{\rho}_1 I_{\{1\in E\}})$,

(5) $\lim_{\rho\uparrow 1}(1-\xi)/(1-\rho) = \delta^{-1}$,

(6) $|s_{i,c}| < C\xi_*^c$ for some C $(0 < C < \infty)$ and $\xi_*(0 < \xi_* < \xi)$.

Proof. For $k = 1$, it is readily verified by differentiating relations (10)–(13) once that the recursive relations for $\alpha_{i,c}^{(1)}$ are identical to those in the case of cyclic polling systems with independent unit Poisson arrivals. Hence, the proof of lemma 2 follows directly from [20]. □

The following result extends lemma 2 to the higher moments of $A_{i,c}$.

Theorem 1. For $i = 1, \ldots, N$, $c = 0, 1, \ldots$, $k = 1, 2, \ldots$,

$$\alpha_{i,c}^{(k)} = f_{i,c}^{(k)} + g_{i,c}^{(k)}, \quad \text{with } f_{i,c}^{(k)} = \xi^c \sum_{j=0}^{k-1} \pi_{i,j}^{(k)} \xi^{jc}, \tag{18}$$

where for $j = 0, 1, \ldots, k-1$,

$$\text{(a)} \quad \lim_{\rho\uparrow 1}(1-\rho)^{k-1}\pi_{i,j}^{(k)} \tag{19}$$

$$= (-1)^j \binom{k-1}{j} \frac{k!}{2^{k-1}} \frac{\hat{\lambda}_1^k (1 - \hat{\rho}_1 I_{\{1\in E\}})^k}{\delta^k} \varphi^{k-1} b_i^{(1)}, \tag{20}$$

$$\text{(b)} \quad \sum_{c=0}^{\infty} g_{i,c}^{(k)} = O\big((1-\rho)^{-(k-1)}\big), \quad \rho \uparrow 1. \tag{21}$$

Proof. Consider the case $i \in G$. For notational convenience, define for $i = 1, \ldots, N$, $c = 0, 1, \ldots, k = 1, 2, \ldots$,

$$\alpha_{j,c}^{(k)}\{i\} := \alpha_{j,c}^{(k)} \quad (j > i), \qquad \alpha_{j,c}^{(k)}\{i\} := \alpha_{j,c-1}^{(k)} \quad (j \leqslant i) \tag{22}$$

and

$$\gamma_{i,c}^{(k)} := \sum_{j=i+1}^{N} \lambda_j \alpha_{j,c}^{(k)} + \sum_{j=1}^{i} \lambda_j \alpha_{j,c-1}^{(k)} = \sum_{j=1}^{N} \lambda_j \alpha_{j,c}^{(k)}\{i\}. \tag{23}$$

Then by repeatedly differentiating (10)–(11) we obtain the following recursive relations for the DS variables $\alpha_{i,c}^{(k)}$ (defined in (14)): for $c = 0, 1, \ldots$,

$$\alpha_{i,c}^{(1)} = b_i^{(1)} \gamma_{i,c}^{(1)}, \qquad \alpha_{i,c}^{(2)} = b_i^{(1)} \gamma_{i,c}^{(2)} + b_i^{(2)} \big(\gamma_{i,c}^{(1)}\big)^2. \tag{24}$$

Similarly, for $k = 3, 4, \ldots$ and $c = 0, 1, \ldots$,

$$\alpha_{i,c}^{(k)} = b_i^{(1)} \gamma_{i,c}^{(k)} + \big(b_i^{(2)} \Gamma_{i,c}^{(k)} + b_i^{(1)} \Delta_{i,c}^{(k)}\big) + \Theta_{i,c}^{(k)}, \tag{25}$$

with

$$\Gamma_{i,c}^{(k)} := \sum_{l=1}^{k-2} \zeta_{l,k} \gamma_{i,c}^{(l)} \gamma_{i,c}^{(k-l)}, \tag{26}$$

and

$$\Delta_{i,c}^{(k)} := \lambda \sum_{l=1}^{k-2} \zeta_{l,k} \sum_{m=1}^{N} \sum_{n=1}^{N} K_{m,n} \alpha_{m,c}^{(l)}\{i\} \alpha_{m,c}^{(k-l)}\{i\}. \tag{27}$$

$\Theta_{i,c}^{(k)}$ in (25) is a linear combination of terms of the form $\prod_{l=1}^{k} (\alpha_{j,c}^{(p_l)}\{i\})^{q_l}$, where p_l and q_l are non-negative integers satisfying $\sum_{l=1}^{k} p_l q_l = k$ and $\sum_{l=1}^{k} q_l \geqslant 2$. The coefficients $\zeta_{k,l}$ in (26) are defined by the following recursive relations: for $k = 3, 4, \ldots$ and $l = 0, 1, \ldots, k-3$,

$$\zeta_{0,k} := 1, \qquad \zeta_{k-2,k} := 3, \qquad \zeta_{j,k} := \zeta_{j,k-1} + \zeta_{j-1,k-1}. \tag{28}$$

We make the following three observations. First, one may show by induction in k that all terms in $\Theta_{i,c}^{(k)}$ are lower order in the sense that $\sum_{c=0}^{\infty} \Theta_{i,c}^{(k)}$ has a pole of order $\sum_{l=1}^{k} (p_l - 1) q_l + 1 < k$ at $\rho = 1$. Second, in the special case of independent unit Poisson arrivals we have $K_{m,n} = 0$ for $m, n = 1, \ldots, N$, so that $\Delta_{i,c}^{(k)}$ vanishes in (25). Third, one may prove by induction that the terms $\Gamma_{i,c}^{(k)}$ and $\Delta_{i,c}^{(k)}$ are of the same order, in the sense that $\sum_{c=0}^{\infty} \Gamma_{i,c}^{(k)}$ and $\sum_{c=0}^{\infty} \Delta_{i,c}^{(k)}$ have a pole of the order k at $\rho = 1$. With these observations, the derivation of the result can be obtained along the lines as discussed in [20]. The details are omitted for compactness of the paper. The results for $i \in E$ can be obtained along similar lines. □

Theorem 1 is a fundamental and powerful result, decomposing the sequences $\{\alpha_{i,c}^{(k)}, c = 0, 1, \ldots\}$ in a dominant part and a recessive part. The dominant part can be exactly analyzed, whereas the impact of recessive part becomes negligible for $\rho \uparrow 1$, and hence does not play a role in the limiting case. A key role will be played by the quantities $h_1^{(k)}$ and $\tilde{h}_1^{(k)}$, defined as follows: for $k = 1, 2, \ldots$,

$$h_1^{(k)} := \sum_{i=1}^{N} \sum_{c=0}^{\infty} \lambda_i \alpha_{i,c}^{(k)}, \qquad \tilde{h}_1^{(k)} := \lim_{\rho \uparrow 1} (1-\rho)^k h_1^{(k)}. \tag{29}$$

The following result gives an explicit expression for $\tilde{h}_1^{(k)}$.

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

$$\tilde{h}_1^{(k)} = \frac{(k-1)!}{2^{k-1}} \frac{\hat{\lambda}_1^k (1 - \hat{\rho}_1 I_{\{1 \in E\}})^k}{\delta^{k-1}} \varphi^{k-1}. \tag{30}$$

Proof. The result can be obtained by combining theorem 1 and the properties listed in lemma 2. We refer to [22] for the details of the proof for the case $k = 2$, and to [20]

for details on the proof of the result for $k \geqslant 1$ for the special case independent Poisson arrival processes (i.e., $K_{i,j} = 0$ for all i, j). □

Next, we define the kth moment of X_i, and its heavy-traffic residue, as follows: for $i = 1, \ldots, N$, $k = 1, 2, \ldots$,

$$x_i^{(k)} := E\big[X_i^k\big], \qquad \tilde{x}_i^{(k)} := \lim_{\rho \uparrow 1} (1-\rho)^k x_i^{(k)}. \tag{31}$$

Then the following result gives an explicit expression for $\tilde{x}_i^{(k)}$.

Theorem 3. For $i = 1, \ldots, N$, $k = 1, 2, \ldots$,

$$\tilde{x}_i^{(k)} = \hat{\lambda}_i^k \big(1 - \hat{\rho}_i I_{\{i \in E\}}\big)^k \prod_{j=0}^{k-1} \left(r + \frac{j\varphi}{2\delta} \right). \tag{32}$$

Proof. The result can be obtained from by repeatedly differentiating (7), using lemma 1 and theorems 1 and 2 along the lines similar to the proof for the case of independent Poisson arrival processes (cf. [21]), and the observation that we assumed $i = 1$ without loss of generality. □

A random variable $\Gamma_{\alpha,\mu}$ with a gamma distribution with scale parameter $\alpha > 0$ and rate parameter $\mu > 0$ has the following probability density function: for $t > 0$,

$$f_{\Gamma_{\alpha,\mu}}(t) := \frac{1}{\Gamma(\alpha)} \mathrm{e}^{-\mu t} \mu^\alpha t^{\alpha-1}, \quad \text{where } \Gamma(\alpha) := \int_0^\infty \mathrm{e}^{-t} t^{\alpha-1} \, \mathrm{d}t. \tag{33}$$

It is readily verified that the LST and the moments of $\Gamma_{\alpha,\mu}$ are given by

$$\Gamma^*_{\alpha,\mu}(s) = \left(\frac{\mu}{\mu + s} \right)^\alpha \quad (\mathrm{Re}\, s > 0), \tag{34}$$

and

$$E[\Gamma^k_{\alpha,\mu}] = \frac{\prod_{j=0}^{k-1} (\alpha + j)}{\mu^k} \quad (k = 1, 2, \ldots), \tag{35}$$

respectively.

Theorem 4. For $i = 1, \ldots, N$, $k = 1, 2, \ldots$,

$$(1-\rho) X_i \to_d \Gamma_{\alpha, \mu_i/\hat{\lambda}_i} \quad (\rho \uparrow 1), \tag{36}$$

where

$$\alpha := \frac{2r\delta}{\varphi}, \qquad \mu_i := \frac{2\delta}{(1 - \hat{\rho}_i I_{\{i \in E\}})\varphi}, \tag{37}$$

where δ and φ are defined in (15) and (16), respectively.

Proof. Notice that the expression at the right-hand side of (32) corresponds to the kth moment of the gamma distribution with scale parameter α and rate parameter $\mu_i/\hat{\lambda}_i$. The result then follows directly from theorem 3 and application of the method of moments (cf., e.g., [5]), which provides sufficient conditions under which convergence in moments implies convergence in distribution. □

We are now ready to present the main result of the paper.

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

$$(1-\rho)W_i \to_d \widetilde{W}_i \quad (\rho \uparrow 1), \tag{38}$$

where the Laplace–Stieltjes transform of $\widetilde{W}_i$ is given by the following expressions: for $i \in G$, $\operatorname{Re} s > 0$,

$$\widetilde{W}_i^*(s) = \frac{1}{(1-\hat{\rho}_i)rs}\left\{\left(\frac{\mu_i}{\mu_i + s\hat{\rho}_i}\right)^{\alpha} - \left(\frac{\mu_i}{\mu_i + s}\right)^{\alpha}\right\}, \tag{39}$$

and for $i \in E$, $\operatorname{Re} s > 0$,

$$\widetilde{W}_i^*(s) = \frac{1}{(1-\hat{\rho}_i)rs}\left\{1 - \left(\frac{\mu_i}{\mu_i + s}\right)^{\alpha}\right\}, \tag{40}$$

with

$$\alpha := \frac{r(1 - \sum_{m\in E}\hat{\rho}_m^2 + \sum_{m\in G}\hat{\rho}_m^2)}{b^{(2)}/b^{(1)} + \hat{\lambda}\sum_{l=1}^{N}\sum_{m=1}^{N} b_l^{(1)} b_m^{(1)} K_{l,m}}, \tag{41}$$

and

$$\mu_i := \frac{1 - \sum_{m\in E}\hat{\rho}_m^2 + \sum_{m\in G}\hat{\rho}_m^2}{(1-\hat{\rho}_i I_{\{i\in E\}})(b^{(2)}/b^{(1)} + \hat{\lambda}\sum_{l=1}^{N}\sum_{m=1}^{N} b_l^{(1)} b_m^{(1)} K_{l,m})}. \tag{42}$$

Proof. The result is obtained directly by combining lemma 1 and theorem 4 and several straightforward manipulations. □

Remark 4.1. By repeatedly differentiating relations (39) and (40), and substituting $s = 0$, it is readily seen that the moments of the limiting delay distribution, defined as $\omega_i^{(k)} := \lim_{\rho\uparrow 1}(1-\rho)^k E[W_i^k]$, are given by the following closed-form expressions: for $k = 1, 2, \ldots$,

$$\omega_i^{(k)} = \frac{1 + \hat{\rho}_i + \cdots + \hat{\rho}_i^k}{k+1}\prod_{j=1}^{k}\left(r + \frac{j\varphi}{2\delta}\right) \quad (i \in G), \tag{43}$$

$$\omega_i^{(k)} = \frac{(1-\hat{\rho}_i)^k}{k+1}\prod_{j=1}^{k}\left(r + \frac{j\varphi}{2\delta}\right) \quad (i \in E), \tag{44}$$

where φ and δ are defined in (15) and (16), respectively. The results in (43)–(44) generalize those in [19], where we considered the special case of independent unit Poisson arrival processes corresponds to the case $K_{l,m} = 0$ for *all* $l, m = 1, \ldots, N$. This implies that the simultaneity of the arrivals leads to an *increase* of the kth moment of the delay (in the limiting case) at each of the queues, for all $k = 1, 2, \ldots$. These asymptotic results can be used to obtain simple approximations for the moments of the waiting times in stable polling systems (i.e., with $\rho < 1$). See section 6 for more details.

Remark 4.2. Relation (29) indicates that both the (scaled) intervisit times and the cycle times converge to a gamma distribution with known parameters. More precisely, for $i \in E$, the customers present at a polling instant at Q_i are exactly those who arrived during the preceding intervisit period. Hence, for $|z| \leqslant 1$, $X_i^*(z) = I_i^*(\lambda(1 - K_i^*(z)))$, which is readily seen to imply that, for $i \in E$,

$$(1 - \rho)I_i \to_{\mathrm{d}} \Gamma_{\alpha,\mu_i} \quad (\rho \uparrow 1), \tag{45}$$

where α and μ_i are defined in (37). Similarly, for $i \in G$, we have $X_i^*(z) = C_i^*(\lambda(1 - K_i^*(z)))$, which implies that

$$(1 - \rho)C_i \to_{\mathrm{d}} \Gamma_{\alpha,\mu_i} \quad (\rho \uparrow 1). \tag{46}$$

Thus, relations (45), (46) and (36) show that the (scaled) intervisit times, cycle times and queue lengths at polling instants converge to gamma distributions with known parameters when the system tends to saturate.

5. Asymptotic properties

The results discussed in the previous section reveal several asymptotic properties of the distribution of the delay in heavy traffic.

Property 1 (Insensitivity). For $i = 1, \ldots, N$, the distribution of $\widetilde{W}_i$:

(1) depends on the second-order moments of the joint batch-size distribution $K_{j,k}$ $(j, k = 1, \ldots, N)$ only through $\sum_{j=1}^{N} \sum_{k=1}^{N} b_j^{(1)} b_k^{(1)} K_{j,k}$;

(2) is independent of the third and higher-order moments of the joint batch-size distribution;

(3) 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;

(4) is independent of the third and higher moments of the service-time distributions;

(5) depends on the switch-over time distributions only through r, the total expected switch-over time per cycle;

(6) is independent of the visit order.

Remark 5.1. Theorem 5 implies that the asymptotic waiting-time distribution depends on the correlation structure in the arrival process (represented by $\sum_{l=1}^{N}\sum_{l=1}^{N} b_l^{(1)} b_m^{(1)} K_{l,m}$) and the variability in the service times (represented by $b^{(2)}$) only through the *linear combination*, φ, defined in (16). This observation implies that in heavy traffic the impact of the correlation structure in the arrival process and the variability in the service times on the waiting-time distributions are to some extent *interchangeable*. Notice that equations (2) and (4) imply that this interchangeability property is not generally true for stable polling systems (i.e., with $\rho < 1$). In addition, we observe that the limiting waiting-time distribution is independent of the third and higher cross-moments of the joint batch-size distribution. This observation is caused by the fact that all terms related to those higher-order cross-moments are in $\Theta_{i,c}^{(k)}$ in (25), and hence are lower-order terms.

The asymptotic insensitivity results listed in property 1, and remark 5.1, are generally not true for stable systems (i.e., with $\rho < 1$). Apparently, the impacts the higher moments of the joint batch-size distribution and the higher moments of the service-time and switch-over time distributions and the visit order on the delay incurred at each of the queues vanish when the system tends to saturate, and as such can be viewed as *lower-order effects* in this context. We emphasize that these findings have not been observed before in the general context of the present paper and provide new fundamental insights into the heavy-traffic behavior of polling systems with correlated arrival streams.

6. Approximation

Theorem 5 and equations (43) and (44) suggest the following simple approximations for the distribution and the moments of the delay at each of the queues: for $i = 1, \ldots, N$, $\rho < 1, x > 0$,

$$\text{Prob}\{W_i > x\} \approx \text{Prob}\big\{\widetilde{W}_i > x(1-\rho)\big\}, \tag{47}$$

and for $i \in G, k = 1, 2, \ldots,$

$$E\big[W_i^k\big] \approx \frac{1}{(1-\rho)^k}\left\{\frac{1+\hat{\rho}_i+\cdots+\hat{\rho}_i^k}{k+1}\prod_{j=1}^{k}\left(r+\frac{j\varphi}{2\delta}\right)\right\}, \tag{48}$$

and for $i \in E, k = 1, 2, \ldots,$

$$E\big[W_i^k\big] \approx \frac{1}{(1-\rho)^k}\left\{\frac{(1-\hat{\rho}_i)^k}{k+1}\prod_{j=1}^{k}\left(r+\frac{j\varphi}{2\delta}\right)\right\}, \tag{49}$$

where δ and φ are defined in (15) and (16), respectively. The tail probabilities of $\widetilde{W}_i$ at the right-hand side of (47) can be obtained very efficiently by applying (one-dimensional) numerical transform inversion to the expressions given in theorem 5. The approximations for the moments of W_i, which follow directly from (43) and (44), are even given in closed form by expressions (48) and (49). Hence, the approximations in (47)–(49) are not only simple, but also very fast-to-evaluate. Notice also that theorem 5 indicates that

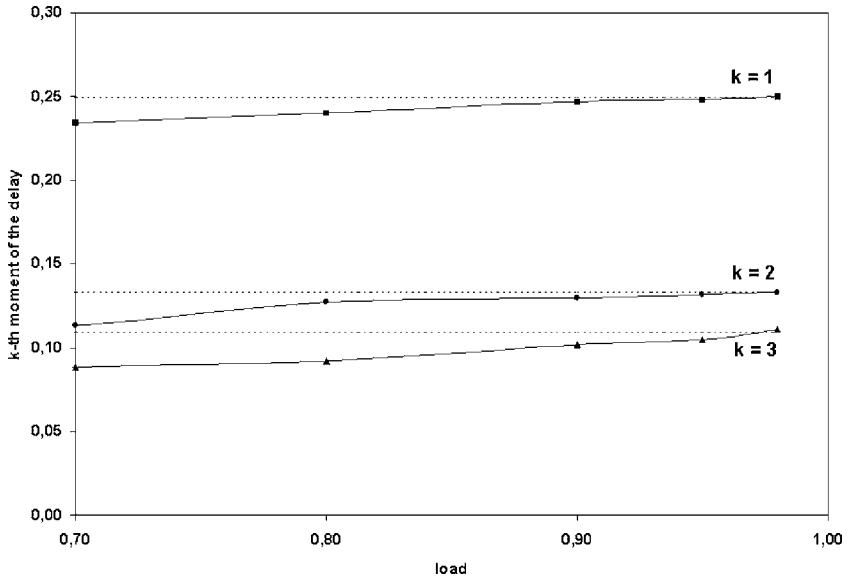

Figure 1. Simulated and approximated values of $(1-\rho)E[W_1^k]$ as a function of the load ($k = 1, 2, 3$).

the approximations are asymptotically exact when the load tends to unity. To assess the accuracy of the approximations for $\rho < 1$, we have performed numerical experiments, comparing the approximations with simulations. The results are outlined below.

Consider a symmetric 5-queue model in which all queues are served exhaustively, the service times are exponentially distributed with mean 0.10 and the switch-over times are exponentially distributed with mean 0.05. The joint batch-size distribution is as follows: $\pi(1, 1, 0, 0, 0) = 1/5$, $\pi(0, 1, 1, 0, 0) = 1/5$, $\pi(0, 0, 1, 1, 0) = 1/5$, $\pi(0, 0, 0, 1, 1) = 1/5$ and $\pi(1, 0, 0, 0, 1) = 1/5$, or equivalently $\underline{K}^*(z_1, z_2, z_3, z_4, z_5) = (z_1z_2 + z_2z_3 + z_3z_4 + z_4z_5 + z_1z_5)/5$. It is readily verified the mean batch sizes (including the possibility of batches of size 0) are given by $E[K_i] = 2/5$ ($i = 1, \ldots, 5$), that $K_{1,2} = K_{2,3} = K_{3,4} = K_{4,5} = K_{5,1} = K_{2,1} = K_{3,2} = K_{4,3} = K_{5,4} = K_{1,5} = 1$ and $K_{i,j} = 0$ in all other cases, and that $\hat{\rho}_i = 0.2$ ($i = 1, \ldots, 5$), $b^{(1)} = 0.10$, $b^{(2)} = 0.02$, $r = 0.25$ and $\hat{\lambda} = 5$. Substituting these parameters in (44) is easily verified to imply that for all i we have $\omega_i^{(1)} = 9/40$, $\omega_i^{(2)} = 4/3$ and $\omega_i^{(3)} = 11/100$. Figure 1 shows the exact and approximated values of $(1-\rho)^k E[W_1^k]$ as a function of ρ, for $k = 1, 2, 3$.[1] The solid lines indicate the simulated results and the approximations are indicated by the dotted lines. The approximations have been obtained from (48) and (49). To asses the quality of the approximations, let us qualify the quality of the approximation "fair" when the (absolute value of the) relative error is less than 20%, "good" when the relative error is 5–10% and "very good" when the error is less than 5%. Close examination of the results plotted in figure 1 shows that overall the approximations are "good" to "very good" when the load

[1] Notice that for fully symmetric systems a closed-form expression for the case $k = 1$ follows immediately from the pseudo-conservation law derived in [15].

is 80% or more. When the load drops about 70%, the accuracy of the approximations tends to become "fair". The accuracy is also found to decrease further when the load is 60% or less. We also observe that the quality of the approximations for given load tends to degrade for higher values of k. This observation was to be expected, because the differences between the tail probabilities of actual waiting-time distribution and the approximations based on the exact asymptotic results are magnified by taking higher moments. We have also performed numerical experiments with asymmetric model instances. The quality of the approximations was found to be only slightly worse than in the fully symmetric case (except for extremely asymmetric model instances, in which cases the approximations are only accurate when ρ is close to 1). The quality of the approximations of the first few (say $k = 1$ to 4) moments by (48)–(49) was still found to be "good" to "very good" when the load is 75–90% or more. On the other hand, for $k \geqslant 5$ the accuracy of the approximations tends to degrade (for fixed ρ), and is only considered good when the load is very close to 1. Hence, the approximations for the moments in (48) and (49) are particularly useful for approximating the first few moments of the delay. By the time of writing of this paper, we are also performing extensive simulation experiments to asses the quality of the approximations for the tail probabilities in (47). Initial results suggest that the approximation (visually) converges to the limiting distribution rather rapidly, except for the very small tail probabilities, which are beyond the scope of this paper.

Remark 6.1. The importance of the approximations (47)–(49) is raised by the fact that the efficiency of simulations or existing numerical techniques (cf. [4,9]) to evaluate the tail probabilities and the moments of the delay works well for lightly and medium loaded systems, but tends to degrade significantly when the system is heavily loaded. In fact, during the simulation experiments we found that extremely long simulation runs had to be run to obtain reliably and estimations (especially for the higher moments of the delay), which addresses exactly the importance of the approximations proposed in (47)–(49): the applicability of the approximations discussed above *complement* the applicability of the existing numerical techniques.

To summarize, the numerical examples discussed above demonstrate that (a) the approximations (47)–(49), covering the impact of both batched and simultaneous arrivals, are accurate for medium and heavily loaded systems, (b) the approximations in (48)–(49) are particularly useful for approximating the first few moments of the delay, and (c) the applicability of the approximations in (47)–(49) complements the applicability of the existing numerical techniques.

7. Topics for further research

A fundamental property of the model considered in the paper is that the joint queue-length processes at polling instants at a fixed queue can be described as a multi-type branching process (MTBP) with immigration in each state [16]. In this paper, we explored the DSA to obtain heavy-traffic results, showing the occurrence of the gamma in

the limiting case. Interestingly, in the theory of MTBPs (cf. [8]) the gamma distribution occurs as the limiting distribution in the so-called critical case (which in the context of the present model corresponds to the case $\rho = 1$), for a general class of MTBPs. In the context of polling models, the results in [8] suggest that the results in this paper may be generalized to a much broader class of polling models, including models with non-cyclic periodic server routing, general branching-type service disciplines (e.g., binomial gated, fractional exhaustive), polling models with customer routing, amongst others. Extension of the results to more general branching-type polling models is a challenging topic for further research.

In this paper it is assumed that all moments of the service times, switch-over times and the (cross-)moments of the joint batch-size distributions are finite. However, theorem 5 (see also property 1) shows that the limiting waiting-time distributions depend only on the first moment of the switch-over times and on the first two moments of the service times and batch-size distributions. This observation suggests that the heavy-traffic results in this paper may be obtained under weaker assumptions about the finiteness of the moments mentioned. This suggestion is supported by the results obtained by Coffman et al. [7] for the case of two-queue models with exhaustive service and independent renewal arrival processes. Derivation of such results in the context of correlated batch arrivals is an interesting topic for further research.

A related area for further research is to analyze the impact of heavy-tailed (say, with infinite variance) batch-size distributions, service-time distributions and switch-over time distributions on the distributions of the delay in heavy traffic. In this context, interesting results have been obtained by Boxma et al. [2], who study the tail behavior of the waiting times in polling systems with so-called regularly varying service times and switch-over times, and by Boxma and Cohen [1], who derive the heavy-traffic limiting distribution for the waiting times in the single-server queue with a class of heavy-tailed service-time distributions.

An interesting feasibility problem is "Can the system be operated such that $\text{Prob}\{W_i > x_i\} < \alpha_i$ $(i = 1, \ldots, N)$?", for given values of x_i and α_i. To the best of the author's knowledge, this type of problems has not been studied before in the literature on polling models. The results in this paper open possibilities for obtaining (approximative) solutions for solving feasibility problems under heavy-traffic assumptions, which is a challenging new area for further research.

Experimental studies have demonstrated that in many applications the arrival processes are non-Poisson. Therefore, it would be interesting to investigate whether exact expressions can still be obtained for the waiting-time distribution when the Poisson assumption is relaxed. In this context, notice that the joint arrival process considered in the present paper is (although not independent unit Poisson) still of "Poisson-type", and that the model fits within the realm of multi-type branching processes, so that the DSA-based approach discussed in section 3 applies. However, for non-Poisson-type arrival processes, the branching structure of the evolution of the joint queue-length process is generally violated, so that the solution method in section 3 is no longer applicable. In this context, encouraging results are obtained by Coffman et al. [7], who derive sim-

ple expressions for the heavy-traffic limit of the waiting-time distribution for a class of non-Poisson arrival processes.

An interesting extension to the standard polling model is the feature of retrial customers. For various variants of this type of models, Langaris [12–14] shows that the expected delay can be obtained by solving sets of linear equations. In this context, it is an interesting topic for further research to investigate whether explicit heavy-traffic results can be obtained for polling models with retrial customers.

Acknowledgments

The author wishes to thank Frank Phillipson for developing the simulation program used to obtain the numerical results. The author would also like to thank the anonymous referees, whose suggestions have led to a significant improvement of the presentation of the paper.

References

[1] O.J. Boxma and J.W. Cohen, The single server queue: heavy tails and heavy traffic, in: *Self-Similar Network Traffic and Performance Evaluation*, eds. K. Park and W. Willinger (Wiley, New York, 2000) pp. 143–169.
[2] O.J. Boxma, Q. Deng and J.A.C. Resing, Polling systems with regularly varying service and/or switch-over times, Adv. Performance Anal. 3 (2000) 71–107.
[3] O.J. Boxma and W.P. Groenendijk, Waiting times in discrete-time cyclic-service systems, IEEE Trans. Commun. 36 (1988) 64–70.
[4] G. Choudhury and W. Whitt, Computing transient and steady state distributions in polling models by numerical transform inversion, Performance Evaluation 25 (1996) 267–292.
[5] K.L. Chung, *A Course in Probability*, 2nd. ed. (Academic Press, 1974).
[6] E.G. Coffman, A.A. Puhalskii and M.I. Reiman, Polling systems with zero switch-over times: a heavy-traffic principle, Ann. Appl. Probab. 5 (1995) 681–719.
[7] E.G. Coffman, A.A. Puhalskii and M.I. Reiman, Polling systems in heavy-traffic: a Bessel process limit, Math. Oper. Res. 23 (1998) 257–304.
[8] J.H. Foster, Branching processes involving immigration, Ph.D. Thesis, University of Wisconsin (1969). (A copy is available from the author upon request.)
[9] A.G. Konheim, H. Levy and M.M. Srinivasan, Descendant set: an efficient approach for the analysis of polling systems, IEEE Trans. Commun. 42 (1994) 1245–1253.
[10] D.P. Kroese, Heavy traffic analysis for continuous polling models, J. Appl. Probab. 34 (1997) 720–732.
[11] S. Kudoh, H. Takagi and O. Hashida, Second moments of the waiting time in symmetric polling systems, J. Oper. Res. Soc. Japan 43 (2000) 306–316.
[12] C. Langaris, A polling model with retrial customers, J. Oper. Res. Soc. Japan 40 (1997) 489–508.
[13] C. Langaris, Gated polling models with customers in orbit, Math. Comput. Modelling 30 (1999) 171–187.
[14] C. Langaris, Markovian polling systems with mixed service disciplines and retrial customers, Top 7 (1999) 305–323.
[15] H. Levy and M. Sidi, Polling systems with simultaneous arrivals, IEEE Trans. Commun. 39 (1991) 823–827.

[16] J.A.C. Resing, Polling systems and multitype branching processes, Queueing Systems Theory Appl. 13 (1993) 409–426.
[17] H. Takagi, *Analysis of Polling Systems* (MIT Press, Cambridge, MA, 1986).
[18] H. Takagi, Queueing analysis of polling models: an update, in: *Stochastic Analysis of Computer and Communication Systems*, ed. H. Takagi (North-Holland, Amsterdam, 1990) pp. 267–318.
[19] R.D. van der Mei, Distribution of the delay in polling systems in heavy traffic, Performance Evaluation 31 (1999) 163–182.
[20] R.D. van der Mei, Polling systems in heavy traffic: higher moments of the delay, Queueing Systems Theory Appl. 31 (1999) 265–294.
[21] R.D. van der Mei, Polling systems with switch-over times under heavy load: moments of the delay, Queueing Systems Theory Appl. 36 (2000) 381–404.
[22] R.D. van der Mei, Polling systems with simultaneous batch arrivals, Stochastic Model. 17 (2001) 271–292.