Queueing Systems 36, 381–404, 2000

Polling Systems with Switch-over Times under Heavy Load: Moments of the Delay

R.D. VAN DER MEI * r.d.vandermei@kpn.com
KPN Research, Department of Network Planning, Leidschendam, Netherlands

Received May 1990; revised February 2000

Abstract. Consider an asymmetric cyclic polling system with general service-time and switch-over time distributions, and with general mixtures of exhaustive and gated service, in heavy traffic. We obtain explicit expressions for all moments of the steady-state delay at each of the queues, under heavy-traffic scalings. The expressions are strikingly simple: they depend on only a few system parameters, and moreover, can be expressed as finite products of simple known terms. The exact results provide new and useful insights into the behavior of polling systems in heavy traffic. In addition, the results suggest simple and fast approximations for the moments of the delay in stable polling systems. Numerical experiments demonstrate the usefulness of the approximations for moderately and heavily loaded systems.

Keywords: polling systems, delay, moments, heavy traffic, approximations

1. Introduction

A polling system is a multi-queue single-server system in which the server visits the queues in cyclic order to process requests pending at the queues. Polling models occur naturally in the modeling of systems in which service capacity (e.g., CPU, bandwidth, processing power, labor) is shared by different types of users, each type having specific traffic characteristics and performance requirements. Polling models find many applications in the areas of computer-communication networks, production, manufacturing and maintenance (cf. [17]). Since the late 1960s polling models have received much attention in the literature (cf. [25,26]). The majority of the available literature is devoted to the analysis of the expected delay at the queues. Relatively little attention has been paid to the higher moments of the delay, knowledge of which is often important to operate the system properly. Exact analysis of the delay in polling models is only possible in some cases, and even in those cases numerical techniques are usually required to obtain the main performance measures of interest, such as the moments of the delay incurred at each of the queues. However, the use of numerical techniques for the analysis of polling models has several drawbacks. First, numerical techniques do not reveal explicitly how the system performance depends on the system parameters and can therefore contribute to the understanding of the system behavior only to a limited extent. Exact closed-

* Part of the work was done when the author was with AT&T Labs, Network Design and Performance Analysis Department, Middletown, NJ, USA.

form expressions provide much more insight into the dependence of the performance measures on the system parameters. Second, the efficiency of each of the numerical algorithms degrades significantly for heavily loaded, highly asymmetric systems with a large number of queues, while the proper operation of the system is particularly critical when the system is heavily loaded. These observations raise the importance of an exact asymptotic analysis of the higher moments of the delay in polling models in heavy traffic.

The literature reveals a remarkable difference in the complexity of different polling models. Resing [20] shows that for a class of polling models the joint queue-length process embedded at polling instants at a fixed queue constitutes a multi-type branching process (MTBP) with immigration. The theory of MTBPs leads to expressions for the generating function of the joint queue-length process at polling instants. For polling models with an MTBP-structure several numerical algorithms have been proposed to determine the (first few) moments of the delay at the queues by solving sets of linear equations (cf., e.g., [10,21,22]). Alternatively, Konheim et al. [12] propose the so-called Descendant Set Approach (DSA), an iterative technique that exploits the MTBP-structure of the model by making use of the concept of descendant sets, to obtain the moments of the delay. Choudhury and Whitt [6] use numerical transform-inversion to calculate tail probabilities and transient performance measures of the delay. Polling models that do not have an MTBP-structure generally require much more computational effort (cf., e.g., [2,16]). Several papers have been devoted to polling models in heavy traffic. In two excellent pieces of work, Coffman et al. [7,8] study a two-queue model with exhaustive service at both queue and show that, under heavy-traffic assumptions and scalings, the total amount of unfinished work converges to a known process. These observations lead to explicit expressions for the moments of the delay at both queues. They also suggest that, based on a partial conjecture, the analysis can be extended to systems with more than two queues. Reiman and Wein [19] and Markowitz [18] study the problem of determining optimal dynamic scheduling problems, by approximating the dynamic scheduling problems by diffusion control problems. Kudoh et al. [14] 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 about the heavy-traffic limits of the first two moments of the waiting times for systems with an arbitrary number of queues. Kroese [13] studies continuous polling systems in heavy traffic and shows that the steady-state number of waiting customers has approximately a gamma-distribution.

Consider an asymmetric cyclic polling model with generally distributed service times and switch-over times (with finite moments), and with general mixtures of exhaustive and gated service. We focus the moments of the delay incurred at each of the queues in heavy traffic. All queues become instable when the system load (denoted by ρ) approaches 1. More precisely, the kth moment of the delay at a queue possesses a pole of order k at $\rho = 1$. Therefore, the main performance measures of interest will be the moments of the scaled delay, defined as $(1 - \rho)$ times the delay, when ρ tends to 1. We derive closed-form expressions for the moments of the (scaled) delay incurred at each of the queues when ρ tends to 1. The key role in the derivation of the result is played by the DSA, an iteractive technique that provides a new view on the evolution of the system based on the concept of descendant sets. The expressions obtained appear to be remarkably simple: (1) they depend only on relatively few (aggregated) system parameters, and (2) can be expressed as finite products of simple known terms. The results reveal a variety of asymptotic properties, which provide new insights into the behavior of polling system under heavy load. The validity of these properties is illustrated by numerical examples. In addition, the expressions obtained suggest simple and fast approximations for the moments of the delay at each of the queues in stable polling systems. The accuracy of the approximations is tested by numerical experiments. The results show that the approximations are accurate for many practical heavy-traffic scenarios.

This paper generalizes, and explicitly uses, the results in [28], where we considered the special case of zero switch-over times. Extension of the results to the case of nonzero switch-over times is of both practical and theoretical importance for several reasons. First, in many applications (e.g., production, manufacturing, maintenance) the switch-over times are non-negligible and generally have a significant impact on the delay. In view of those applications, the practical importance of extension of the analysis to the case of nonzero switch-over times is evident. Second, exact analysis of the moments of the delay presented in this paper *explicitly* reveals how the moments of the delay depend on the switch-over times, and leads to new insights into the behavior of heavily-loaded polling systems with nonzero switch-over times (see section 6). Third, the efficiency of the existing numerical techniques for evaluating the moments of the delay may degrade dramatically for heavily-loaded systems with a large number of queues. This raises the need for simple and fast-to-evaluate approximations for the moments of the delay, incorporating models with nonzero switch-over times. The exact asymptotic analysis in this paper leads to simple, fast-to-evaluate and accurate approximations for the moments of the delay (see section 7). The contributions listed here make the added value of the present paper compared to [28] evident.

The remainder of this paper is organized as follows. In section 2 the model is described. In section 3 the basic principles of the DSA are reviewed. In section 4 we give a number of preliminary results which are useful in the derivation of the main results. In section 5 we use these preliminary results to derive explicit expressions for the (scaled) moments of the delay in heavy traffic. In section 6 we discuss a number of new insights into the heavy-traffic behavior of the system, and illustrate the results by numerical examples. In section 7 we propose and test simple approximations for the moments of the waiting times in heavy traffic, and address the practicality of the asymptotic results. Finally, in section 8 we address a number of topics for further research.

2. Model description

Consider a system consisting of $N \geqslant 2$ 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 Laplace–Stieltjes Transform (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 a time epoch at which the server visits Q_i. 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 $E := \{i\colon\ Q_i$ is served exhaustively$\}$ and $G := \{i\colon\ Q_i$ 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 time R_i, with LST $R_i^*(\cdot)$ and finite kth moment $r_i^{(k)}$, $k = 1, 2, \ldots$. Denote by $r := \sum_{i=1}^{N} r_i^{(1)} > 0$ the expected total switch-over time per cycle of the server along the queues. All interarrival times, service times and switch-over 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. [11]).

Let W_i be the delay incurred by an arbitrary customer at Q_i. Our main interest is in the behavior of $E[W_i^k]$, the kth moment of the delay at Q_i, in heavy traffic. Throughout, $E[W_i^k]$ 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 remain fixed. It is known that when $\rho \uparrow 1$, all queues become unstable and hence $E[W_i^k]$ tends to infinity (for all i and k). More precisely, $E[W_i^k]$ has a pole of order k at $\rho = 1$: for $i = 1, \ldots, N$, $k = 1, 2, \ldots$,

$$E\big[W_i^k\big] = \frac{\omega_i^{(k)}}{(1-\rho)^k} + \mathrm{O}\big((1-\rho)^{-(k-1)}\big), \quad \rho \uparrow 1. \tag{1}$$

Based on (1), the analysis will be oriented towards the determination of

$$\omega_i^{(k)} := \lim_{\rho \uparrow 1} (1-\rho)^k E\big[W_i^k\big], \quad i = 1, \ldots, N,\ k = 1, 2, \ldots. \tag{2}$$

Throughout, the following notation is used. For an event E, denote by I_E the indicator function on E. Denote by $\mathbf{e}_i$ the ith unit vector, $i = 1, \ldots, N$. Define the inner product of the vectors $\mathbf{x}$ and $\mathbf{y}$ by $\mathbf{x}^{\mathrm{T}}\mathbf{y} := \sum_{i=1}^{N} x_i y_i$. Finally, for each variable x (for scalars) or $\mathbf{x}$ (for vectors) or $\mathbf{X}$ (for matrices) that is a function of ρ, we denote its values evaluated at $\rho = 1$ by $\hat{x}$, $\hat{\mathbf{x}}$ and $\widehat{\mathbf{X}}$, respectively.

3. The descendant set approach

Denote by X_i the number of customers at Q_i at an arbitrary polling instant at Q_i when the system is in steady state, and denote its corresponding Probability Generating Function (PGF) by $X_i^*(\cdot)$. Denoting the LST of the waiting time at Q_i by $W_i^*(\cdot)$, the waiting-time distribution at Q_i is related to the distribution of X_i through the following expressions (cf. [34]): for $\mathrm{Re}\, s \geqslant 0$,

$$\begin{aligned} W_i^*(s) &= \frac{1-\rho}{r}\,\frac{X_i^*(B_i^*(s)) - X_i^*(1-s/\lambda_i)}{s-\lambda_i(1-B_i^*(s))}, \quad & i \in G, \\ W_i^*(s) &= \frac{1-\rho}{r}\,\frac{1 - X_i^*(1-s/\lambda_i)}{s-\lambda_i(1-B_i^*(s))}, & i \in E. \end{aligned} \tag{3}$$

Consequently, to obtain expressions for the moments of W_i, it is suffices to obtain expressions for the moments of X_i. To this end, note that straightforward balancing arguments indicate that $E[X_i]$ can be expressed in following closed form (cf., e.g., [25]):

$$E[X_i] = \frac{\lambda_i r}{1-\rho}, \quad i \in G, \qquad E[X_i] = \frac{\lambda_i(1-\rho_i)r}{1-\rho}, \quad i \in E. \tag{4}$$

However, in general, the higher moments of X_i cannot be obtained explicitly. In the literature, there are several (numercial) techniques to obtain the moments of the delay. In this paper, we focus of the Descendant Set Approach (DSA), an iterative technique based on the concept of so-called descendant sets. The DSA is shown to be particularly useful in obtaining the results presented in this paper. In this section we outline the use of the DSA for the present model (see [12] for more details).

The customers in a polling system can be classified as originators and non-originators. An originator is a customer that arrives at the system during a switch-over period. A non-originator is a customer that arrives at the system during the service of another customer. For a customer C, define the children set to be the set of customers arriving during the service of C; the descendant set of C is recursively defined to consist of C, its children and the descendants of its children. The DSA is focused on the determination of the moments of the delay at a fixed queue, say Q_1. To this end, the DSA concentrates on the determination of $X_1(P^*)$, defined as the number of customers at Q_1 present at an arbitrary fixed polling instant P^* at Q_1. P^* is referred to as the reference point at Q_1. The main ideas are the observations that (1) each of the $X_1(P^*)$ customers belongs to the descendant set of exactly one originator, and (2) the evolutions of the descendant sets of different originators are stochastically independent. Therefore, the DSA concentrates on an arbitrary tagged customer T which arrived at Q_i in the past and on calculating the number of type-1 descendants it has at P^*. Summing up these numbers over all past originators yields $X_1(P^*)$ and hence X_1, because P^* is chosen arbitrarily.

The DSA considers the Markov process embedded at the polling instants of the system. To this end, we number the successive polling instants as follows. Let $P_{N,0}$ be the last polling instant at Q_N prior to P^*, and for $i = N-1, \dots, 1$, let $P_{i,0}$ be recursively defined as the last polling instant at Q_i prior to $P_{i+1,0}$. In addition, for $c = 1, 2, \ldots,$ we define $P_{i,c}$ to be the last polling instant at Q_i prior to $P_{i,c-1}$, $i = 1, \ldots, N$. The DSA is oriented towards the determination of the contribution to $X_1(P^*)$ of an arbitrary customer present at Q_i at $P_{i,c}$. To this end, define an (i, c)-customer to be a type-i customer present at Q_i at $P_{i,c}$. Moreover, for a tagged (i, c)-customer $T_{i,c}$, we define $A_{i,c}$ to be the number of type-1 descendants it has at P^*. In this way, $A_{i,c}$ can be viewed as the contribution of $T_{i,c}$ to $X_1(P^*)$. Denote the PGF of $A_{i,c}$ by $A^*_{i,c}(\cdot)$, and denote the kth factorial moment of $A_{i,c}$ by $\alpha^{(k)}_{i,c}$, $k = 1, 2, \ldots.$

To express the distribution of X_1 in terms of the distributions of the DS variables $A_{i,c}$, denote by $R_{i,c}$ the switch-over period from Q_i to Q_{i+1} immediately after the service period at Q_i starting at $P_{i,c}$. Moreover, denote by $S_{i,c}$ the total contribution to X_1 of all customers that arrive at the system during $R_{i,c}$ (note that, by definition, these customers are original customers), and denote its PGF by $S^*_{i,c}(\cdot)$. In this way, $S_{i,c}$ can be seen as the contribution of $R_{i,c}$ to X_1. It is readily verified that we can write $X_1 = \sum_{i=1}^{N} \sum_{c=0}^{\infty} S_{i,c}$, and since the variables $S_{i,c}$ are mutually independent, we can write, for $|z| \leqslant 1$,

$$X^*_1(z) = \prod_{i=1}^{N} \prod_{c=0}^{\infty} S^*_{i,c}(z). \tag{5}$$

Because $S_{i,c}$ is the total contribution to X_1 of all (original) customers that arrive during $R_{i,c}$, the distribution of $S_{i,c}$ can be expressed in terms of the distributions of the DS variables $A_{i,c}$ as follows (cf. also [12]): for $i = 1, \ldots, N$, $c = 0, 1, \ldots, |z| \leqslant 1$,

$$S^*_{i,c}(z) = R^*_i\left(\sum_{j=i+1}^{N} \lambda_j \bigl[1 - A^*_{j,c}(z)\bigr] + \sum_{j=1}^{i} \lambda_j \bigl[1 - A^*_{j,c-1}(z)\bigr] \right). \tag{6}$$

The observation that the contribution X_1 of a tagged (i, c)-customer $T_{i,c}$ is equal to the total sum of the contributions to X_1 of all its (immediate) children leads to the following recursive relations between the generating functions $A^*_{i,c}(\cdot)$, depending on the service discipline at Q_i (cf. also [12, equations (3.1) and (3.6)]): for $c = 0, 1, \ldots, |z| \leqslant 1$,

$$\begin{aligned} A^*_{i,-1}(z) &= z I_{\{i=1\}}, \\ A^*_{i,c}(z) &= B^*_i\left(\sum_{j=i+1}^{N} \bigl[\lambda_j - \lambda_j A^*_{j,c}(z)\bigr] + \sum_{j=1}^{i} \bigl[\lambda_j - \lambda_j A^*_{j,c-1}(z)\bigr] \right), \quad i \in G, \end{aligned} \tag{7}$$

and

$$\begin{aligned} A^*_{i,-1}(z) &= z I_{\{i=1\}}, \\ A^*_{i,c}(z) &= \Theta^*_i\left(\sum_{j=i+1}^{N} \bigl[\lambda_j - \lambda_j A^*_{j,c}(z)\bigr] + \sum_{j=1}^{i-1} \bigl[\lambda_j - \lambda_j A^*_{j,c-1}(z)\bigr] \right), \quad i \in E, \end{aligned} \tag{8}$$

where $\Theta_i^*(\cdot)$ stands for the LST of Θ_i, defined as the duration of an $M/G/1$-busy period with arrival rate λ_i and service-time distribution $B_i^*(\cdot)$. In this way, relations (5)–(8) give a full, but implicit, characterization of the distribution of X_1.

4. Preliminaries

In this section we discuss a number of preliminary results that will be used in the next section to obtain the main result of the paper. In section 4.1 we obtain expressions for higher-order derivatives of composite functions. In section 4.2 these expressions are used to obtain recursive relations for the DS variables $\alpha_{i,c}^{(k)}$. In section 4.3 we review several useful properties of the sequences $\{\alpha_{i,c}^{(k)},\ c = 0, 1, \ldots\}$ in the limiting case $\rho \uparrow 1$.

4.1. Derivatives of composite functions

In the analysis with the DSA (discussed in section 3), a key role is played by composite functions (see for example equations (6)–(8)). To obtain expressions for the higher moments of the DS variables based on these relations, we need to find expressions for the higher-order derivatives of composite functions. For a function h, denote by $[h]^{(k)}(x)$ the kth derivative of h, evaluated at x (assuming existence). We consider a general composite function of the form $f(g(x))$, and assume that all derivatives of f, g and $f(g)$ with respect to x exist. Then higher-order derivatives of $f(g(x))$ can be obtained by applying standard chain rules for differentiation. To give an expression for the kth derivative of the composite function $f(g(x))$ with respect to x, we introduce the following notation: for $k = 1, 2, \ldots,$

$$S_k := \left\{ \mathbf{m}^{(k)} = (m_1, \ldots, m_k)\colon\ m_j \text{ are non-negative integers with } \sum_{j=1}^{k} j m_j = k \right\}. \tag{9}$$

The following result gives an expression for $[f(g)]^{(k)}(x)$, the kth derivative of the function $f(g)$ with respect to x.

Theorem 1. For $k = 1, 2, \ldots,$ the kth derivative of composite function of the form $f(g(x))$ is a linear combination of the form

$$\big[f(g)\big]^{(k)}(x) = \sum_{\mathbf{m}^{(k)} \in S_k} c_k\big(\mathbf{m}^{(k)}\big)[f]^{(l_k)}\big(g(x)\big) \prod_{i=1}^{k} \big([g]^{(i)}(x)\big)^{m_i}, \quad l_k := \sum_{j=1}^{k} m_j, \tag{10}$$

where the coefficients $c_k(\mathbf{m}^{(k)})$ are defined by the following recursive relations:

$$c_1(1) := 1, \tag{11}$$

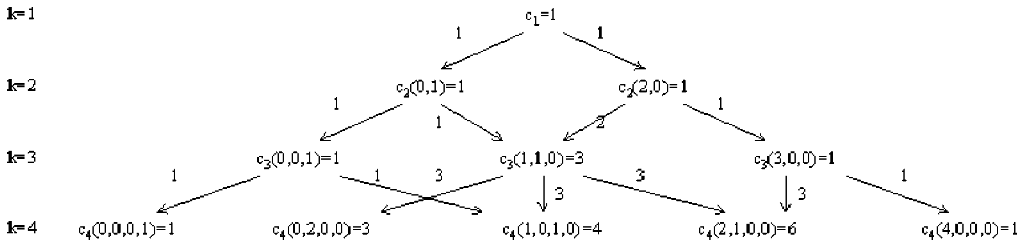

Figure 1. Coefficients of the derivatives of a composite function.

and for $k = 2, 3, \ldots,$[1]

$$\begin{aligned} c_k\big(\mathbf{m}^{(k)}\big) = {} & c_{k-1}(m_1 - 1, m_2, \ldots, m_{k-1}) I_{\{m_1>0\}} \\ & + \sum_{j=1}^{k-1} (m_j + 1) c_{k-1}(m_1, \ldots, m_{j-1}, m_j + 1, m_{j+1} - 1, m_{j+2}, \ldots, m_{k-1}) \\ & \times I_{\{m_{j+1}>0\}}. \end{aligned} \tag{12}$$

Proof. The result can be obtained by repeatedly applying standard rules for differentiation. For compactness of the paper, the details of the proof are omitted.

For example, it is readily verified that $S_1 = \{1\}$, $S_2 = \{(2, 0), (0, 1)\}$, $S_3 = \{(3, 0, 0), (1, 1, 0), (0, 0, 1)\}$ and $S_4 = \{(4, 0, 0, 0), (2, 1, 0, 0), (1, 0, 1, 0), (0, 2, 0, 0), (0, 0, 0, 1)\}$. The corresponding coefficients $c_k(\mathbf{m}^{(k)})$ are given in figure 1. □

4.2. *Recursive relations for higher moments of DS variables*

In this section we apply theorem 1 to obtain recursive relations for $\alpha_{i,c}^{(k)}$, the moments of the DS variables $A_{i,c}$. To this end, denote the kth moment of Θ_i by $\theta_i^{(k)} := E[\Theta_i^k]$, for $i = 1, \ldots, N$, $k = 1, 2, \ldots$. Moreover, to cover the cases of both gated and exhaustive service at Q_i, the following notation is useful: for $c = 0, 1, \ldots,$ $k = 1, 2, \ldots,$

$$\begin{aligned} \gamma_{i,c}^{(k)} &:= \sum_{j=i+1}^{N} \lambda_j \alpha_{j,c}^{(k)} + \sum_{j=1}^{i} \lambda_j \alpha_{j,c-1}^{(k)}, \quad i \in G, \\ \gamma_{i,c}^{(k)} &:= \sum_{j=i+1}^{N} \lambda_j \alpha_{j,c}^{(k)} + \sum_{j=1}^{i-1} \lambda_j \alpha_{j,c-1}^{(k)}, \quad i \in E, \end{aligned} \tag{13}$$

and denote for $k = 1, 2, \ldots,$

$$\varphi_i^{(k)} := E\big[B_i^k\big], \quad i \in G, \qquad \varphi_i^{(k)} := E\big[\Theta_i^k\big], \quad i \in E. \tag{14}$$

[1] For $j = k - 1$, the argument of $c_{k-1}(\cdot)$ in (13) should be read as $(m_1, \ldots, m_j + 1)$. Similarly, for $j = k - 2$, the argument should be read as $(m_1, \ldots, m_j + 1, m_{j+1} - 1)$.

Notice that the assumption that $b_i^{(k)}$ is finite implies that $\theta_i^{(k)}$ is also finite, so that $\varphi_i^{(k)}$ is finite for all i, k. Note also that for $i \in G$ we have $\varphi_i^{(k)} = b_i^{(k)}$, $k = 1, 2, \ldots$, and that for $i \in E$ we have $\varphi_i^{(1)} = b_i^{(1)}/(1-\rho_i)$ and $\varphi_i^{(2)} := b_i^{(2)}/(1-\rho_i)^3$.

Recursive relations between the variables $\alpha_{i,c}^{(k)}$ can be obtained by repeatedly differentiating relations (7) and (8). Noting that (7) and (8) are composite functions of the form $f(g(x))$, we can apply theorem 1 to obtain expressions for $\alpha_{i,c}^{(k)}$. More precisely, taking for $i \in G$,

$$f(z) := B_i^*(z), \qquad g(z) := \sum_{j=i+1}^{N} \lambda_j\bigl(1 - A_{j,c}^*(z)\bigr) + \sum_{j=1}^{i} \lambda_j\bigl(1 - A_{j,c-1}^*(z)\bigr),$$

and for $i \in E$,

$$f(z) := \Theta_i^*(z), \qquad g(z) := \sum_{j=i+1}^{N} \lambda_j\bigl(1 - A_{j,c}^*(z)\bigr) + \sum_{j=1}^{i-1} \lambda_j\bigl(1 - A_{j,c-1}^*(z)\bigr),$$

application of theorem 1 leads to the following recursive relations: for $i = 1, \ldots, N$, $c = 0, 1, \ldots$, $k = 1, 2, \ldots$,

$$\alpha_{i,c}^{(k)} = \sum_{\mathbf{m}^{(k)} \in S_k} c_k\bigl(\mathbf{m}^{(k)}\bigr)\varphi_i^{(l_k)} \prod_{j=1}^{k} \bigl(\gamma_{i,c}^{(j)}\bigr)^{m_j}, \quad \text{with } l_k = \sum_{j=1}^{k} m_j, \tag{15}$$

supplemented with the initial conditions $\alpha_{1,-1}^{(k)} = 1$, $\alpha_{i,-1}^{(k)} = 0$, for $i = 2, 3, \ldots, N$, $k = 1, 2, \ldots$. For example, it is readily verified that the recursive relations for $k = 1, 2, 3$ and 4 are given by the following expressions (see also [12, equations (3.2), (3.3) and (4.4)] for $k = 1, 2, 3$, respectively): for $i = 1, \ldots, N$, $c = 0, 1, \ldots$,

$$\begin{aligned} &\alpha_{i,c}^{(1)} = \varphi_i^{(1)}\gamma_{i,c}^{(1)}, \qquad \alpha_{i,c}^{(2)} = \varphi_i^{(1)}\gamma_{i,c}^{(2)} + \varphi_i^{(2)}\bigl(\gamma_{i,c}^{(1)}\bigr)^2, \\ &\alpha_{i,c}^{(3)} = \varphi_i^{(1)}\gamma_{i,c}^{(3)} + 3\varphi_i^{(2)}\gamma_{i,c}^{(1)}\gamma_{i,c}^{(2)} + \varphi_i^{(3)}\bigl(\gamma_{i,c}^{(1)}\bigr)^3, \end{aligned} \tag{16}$$

$$\alpha_{i,c}^{(4)} = \varphi_i^{(1)}\gamma_{i,c}^{(4)} + 3\varphi_i^{(2)}\bigl(\gamma_{i,c}^{(2)}\bigr)^2 + 4\varphi_i^{(2)}\gamma_{i,c}^{(1)}\gamma_{i,c}^{(3)} + 6\varphi_i^{(3)}\bigl(\gamma_{i,c}^{(1)}\bigr)^2\gamma_{i,c}^{(2)} + \varphi_i^{(4)}\bigl(\gamma_{i,c}^{(1)}\bigr)^4. \tag{17}$$

Note that for $k = 1, 2, \ldots$, the vector $\mathbf{m}^{(k)} = (0, \ldots, 0, 1)$ is the only term in S_k with $l_k = 1$ and that $c_k(0, \ldots, 0, 1) = 1$. It is convenient to rewrite relation (15) by splitting off the (unique) term with $l_k = 1$: for $i = 1, \ldots, N$, $c = 0, 1, \ldots$, $k = 1, 2, \ldots$,

$$\begin{aligned} &\alpha_{i,c}^{(k)} = \varphi_i^{(1)}\gamma_{i,c}^{(k)} + \delta_{i,c}^{(k)}, \quad \text{with } \delta_{i,c}^{(k)} := \sum_{\mathbf{m}^{(k)} \in S_k:\ l_k > 1} c_k\bigl(\mathbf{m}^{(k)}\bigr)\varphi_i^{(l_k)} \prod_{j=1}^{k} \bigl(\gamma_{i,c}^{(j)}\bigr)^{m_j}, \\ &\qquad \text{with } l_k = \sum_{j=1}^{k} m_j. \end{aligned} \tag{18}$$

To analyze the behavior of the sequences $\{\alpha_{i,c}^{(k)},\ c = 0, 1, \ldots\}$, it is convenient to write relations (18) in matrix notation. To this end, for $k = 1, \ldots, N$, define $\mathbf{M}_k$ to be the matrix whose (i, j)th element equals $I_{\{i=j\}}$ for $i \neq k$ and whose (k, j)th element is given by $b_k^{(1)}\lambda_j$ if $k \in G$, and by $I_{\{k\neq j\}}\theta_k^{(1)}\lambda_j$ if $k \in E$. Define, for $c = -1, 0, 1, \ldots,$ $k = 1, 2, \ldots,$

$$\mathbf{M} := \mathbf{M}_1 \cdots \mathbf{M}_N \quad \text{and} \quad \boldsymbol{\alpha}_c^{(k)} := \big(\alpha_{1,c}^{(k)}, \ldots, \alpha_{N,c}^{(k)}\big)^{\mathrm{T}}. \tag{19}$$

Then (18), and the initial conditions, can be expressed as follows: for $c = 0, 1, \ldots,$ $k = 1, 2, \ldots,$

$$\boldsymbol{\alpha}_{-1}^{(k)} := \mathbf{e}_1, \qquad \boldsymbol{\alpha}_c^{(k)} = \mathbf{M}\boldsymbol{\alpha}_{c-1}^{(k)} + \mathbf{x}_c^{(k)} = \mathbf{M}^{c+1}\mathbf{e}_1 + \sum_{j=0}^{c} \mathbf{M}^j \mathbf{x}_{c-j}^{(k)},$$
$$\text{with } \mathbf{x}_c^{(k)} := \sum_{m=1}^{N} \delta_{m,c}^{(k)} \mathbf{M}_1 \cdots \mathbf{M}_{m-1}\mathbf{e}_m. \tag{20}$$

The third equality in (20) follows by iterating the second equality in (14) c times.

4.3. *Heavy-traffic properties of the DS variables*

The sequences $\{\alpha_{i,c}^{(k)},\ c = 0, 1, \ldots\},\ k = 1, 2, \ldots,$ are determined completely by relations (20). Therefore, an important role is played by the matrix $\mathbf{M}$. According to the Frobenius theorem (cf. [1]), $\mathbf{M}$ has a unique, real and strictly positive eigenvalue ξ, referred to as the maximum eigenvalue. Note that ξ is, generally, a function of ρ, say $\xi = f(\rho)$. It is shown in [33] that

(a) $f(1) = 1$,

(b) $f(\cdot)$ is continuous at $\rho = 1$, and

(c) $f'(1) = 1/\gamma$, with $\gamma := (1 - \sum_{m\in E} \hat{\rho}_m^2 + \sum_{m\in G} \hat{\rho}_m^2)/2$.

By exploring these properties, it is shown in [33] that if we define, for $i = 1, \ldots, N,\ k = 1, 2, \ldots,$

$$\kappa_i^{(k)} := \frac{k!}{2^{k-1}} \frac{\hat{\lambda}_1^k (1 - \hat{\rho}_1 I_{\{1\in E\}})^k}{\gamma^k} \left(\frac{b^{(2)}}{b^{(1)}}\right) b_i^{(1)}, \tag{21}$$

then the sequences $\{\alpha_{i,c}^{(k)},\ c = 0, 1, \ldots\}$ can be decomposed as follows: for $i = 1, \ldots, N$, $c = 0, 1, \ldots, k = 1, 2, \ldots,$

$$\alpha_{i,c}^{(k)} = f_{i,c}^{(k)} + g_{i,c}^{(k)}, \quad \text{with } f_{i,c}^{(k)} = \xi^c \sum_{j=0}^{k-1} \pi_{i,j}^{(k)} \xi^{jc}, \tag{22}$$

where

$$\text{(a)} \quad \lim_{\rho\uparrow 1} (1-\rho)^{k-1} \pi_{i,j}^{(k)} = \tilde{\pi}_{i,j}^{(k)} := (-1)^j \binom{k-1}{j} \kappa_i^{(k)}, \quad j = 0, 1, \ldots, k-1, \tag{23}$$

$$\text{(b)} \quad \sum_{c=0}^{\infty} g_{i,c}^{(k)} = \mathrm{O}\big((1-\rho)^{-(k-1)}\big), \quad \rho \uparrow 1. \tag{24}$$

Relation (22) decomposes the sequences $\{\alpha_{i,c}^{(k)},\ c = 0, 1, \ldots\}$ in a dominant part (i.e., $f_{i,c}^{(k)}$) and a recessive part (i.e., $g_{i,c}^{(k)}$). The dominant part can be exactly analyzed, whereas the impact of recessive part becomes negligible in the limiting case $\rho \uparrow 1$ (in a sense discussed below), and as such does not play a role in the limiting case. This concept of dominance will play a key role in the derivation of the results in the next section.

5. Analysis

In this section we use the results discussed in the previous section to obtain the main results of the paper. The derivation proceeds along two steps. First, we use the preliminary results discussed in section 4 to obtain expressions for the limiting scaled moments of X_1 (defined in section 3). Then, these results are used to obtain expressions for $\omega_i^{(k)}$.

The following notation is useful. Defining $x_1^{(k)} := E[X_1^k]$, it is known that $x_1^{(k)}$ possesses a kth order pole at $\rho = 1$ (although a rigorous proof has not been given in the literature). Therefore, it is convenient to define, for $k = 1, 2, \ldots,$

$$\tilde{x}_1^{(k)} := \lim_{\rho \uparrow 1} (1-\rho)^k E\big[X_1^k\big], \tag{25}$$

referred to as the heavy-traffic residue of $E[X_1^k]$. Notice that $\tilde{x}_1^{(k)}$ is also equal to the heavy-traffic residue of the kth factorial moment of X_1, i.e., $\tilde{x}_1^{(k)} = \lim_{\rho\uparrow 1}(1-\rho)^k E[X_1(X_1-1)\cdots(X_1-k+1)]$, for $k = 1, 2, \ldots.$

5.1. *Deterministic switch-over times*

Let us first assume that all switch-over times are deterministic. In that case, relations (5) and (6) imply that $X_1^*(z)$ can be expressed in terms of the distribution of the DS variables $A_{i,c}$ as follows (see also equation (30) in [23]): for $|z| \leqslant 1$,

$$X_1^*(z) = \exp\big\{-r\Lambda H_1^*(z)\big\}, \quad \text{where } H_1^*(z) := \sum_{i=1}^{N} \frac{\lambda_i}{\Lambda} \sum_{c=-1}^{\infty} \big[1 - A_{i,c}^*(z)\big]. \tag{26}$$

The moments of X_1 can be expressed in terms of the moments of $A_{i,c}$ by repeatedly differentiating (26). To this end, note that $X_1^*(z)$ in (26) is a composite function of the form $f(g(z))$, with $f(z) := \mathrm{e}^{-z}$ and $g(z) := r\Lambda H_1^*(z)$. Hence, applying theorem 1 we can write, for $|z| \leqslant 1$, $k = 1, 2, \ldots,$

$$\begin{aligned}
\big[X_1^*\big]^{(k)}(z) &= \sum_{\mathbf{m}^{(k)} \in S_k} c_k(\mathbf{m}^{(k)}) \exp\big\{-r\Lambda H_1^*(z)\big\} r^{l_k} \Lambda^{l_k} \prod_{i=1}^{k} \big(\big[H_1^*\big]^{(i)}(z)\big)^{m_i}, \\
\text{with } l_k &= \sum_{j=1}^{k} m_j,
\end{aligned} \tag{27}$$

where the coefficients $c_k(\mathbf{m}^{(k)})$ are recursively defined in theorem 1. Defining, for $k = 1, 2, \ldots,$

$$h_1^{(k)} := -\frac{\partial^k}{\partial z^k} H_1^*(z)\bigg|_{z=1} = \sum_{i=1}^{N} \frac{\lambda_i}{\Lambda} \sum_{c=-1}^{\infty} \alpha_{i,c}^{(k)}, \qquad \tilde{h}_1^{(k)} := \lim_{\rho \uparrow 1} (1-\rho)^k h_1^{(k)}, \tag{28}$$

equation (27) indicates that we can write, for $k = 1, 2, \ldots,$

$$\tilde{x}_1^{(k)} = \sum_{\mathbf{m}^{(k)} \in S_k} c_k\big(\mathbf{m}^{(k)}\big) r^{l_k} \hat{\Lambda}^{l_k} \exp\big\{-r\widehat{\Lambda}\big\} \prod_{i=1}^{k} \big(\tilde{h}_1^{(i)}\big)^{m_i}, \quad \text{with } l_k = \sum_{j=1}^{k} m_j. \tag{29}$$

Then, combining relations (22)–(24) with the properties of the maximum eigenvalue of $\mathbf{M}$ listed in section 4.3 the following result was shown in [29]: for $k = 1, 2, \ldots,$

$$\tilde{h}_1^{(k)} = \frac{(k-1)!}{2^{k-1}} \frac{\hat{\lambda}_1^k (1 - \hat{\rho}_1 I_{\{1\in E\}})^k}{\widehat{\Lambda} \gamma^{k-1}} \left(\frac{b^{(2)}}{b^{(1)}}\right)^{k-1} \quad \text{with } \gamma := 1 - \sum_{m\in E} \hat{\rho}_m^2 + \sum_{m \in G} \hat{\rho}_m^2. \tag{30}$$

It is important to note that relations (28)–(30), together with theorem 1, provide an *implicit*, not explicit, characterization of $\tilde{x}_1^{(k)}$, $k = 1, 2, \ldots.$ The following result uses these relations to obtain an explicit expression for $\tilde{x}_1^{(k)}, k = 1, 2, \ldots.$

Theorem 2. If all switch-over times are deterministic, then for $k = 1, 2, \ldots,$

$$\tilde{x}_1^{(k)} = \hat{\lambda}_1^k \big(1 - \hat{\rho}_1 I_{\{1\in E\}}\big)^k \prod_{j=0}^{k-1} \left[r + j\left(\frac{b^{(2)}/b^{(1)}}{2\gamma}\right)\right], \quad \text{with } \gamma := 1 - \sum_{m\in E} \hat{\rho}_m^2 + \sum_{m\in G} \hat{\rho}_m^2. \tag{31}$$

Proof. Via induction in k. For $k = 1$, it follows directly from (4) that $\tilde{x}_1^{(1)} = r\hat{\lambda}_1$ for $1 \in G$ and $\tilde{x}_1^{(1)} = r\hat{\lambda}_1(1 - \hat{\rho}_1)$ for $1 \in E$. To proceed, assume that theorem 2 is valid for some $k \geqslant 1$. To prove the validity of the results for $k + 1$, consider an arbitrary term in the summation in (27). Then by using standard rules for differentiation it is readily verified that

$$\begin{aligned}
&\frac{\partial}{\partial z}\left\{\exp\big\{-r\Lambda H_1^*(z)\big\} \prod_{i=1}^{k} \big(\big[H_1^*\big]^{(i)}(z)\big)^{m_i}\right\} \\
&\quad = -r\Lambda\big[H_1^*\big]^{(1)}(z) \exp\big\{-r\Lambda H_1^*(z)\big\} \prod_{i=1}^{k} \big(\big[H_1^*\big]^{(i)}(z)\big)^{m_i} + \exp\big\{-r\Lambda H_1^*(z)\big\} \\
&\qquad \times \sum_{j=1}^{k} \left(m_j \big(\big[H_1^*\big]^{(j)}(z)\big)^{m_j - 1} \big[H_1^*\big]^{(j+1)}(z) \prod_{i\neq j} \big(\big[H_1^*\big]^{(i)}(z)\big)^{m_i}\right) && (32)\\
&\quad = \left(-r\Lambda\big[H_1^*\big]^{(1)}(z) + \sum_{j=1}^{k} m_j \frac{[H_1^*]^{(j+1)}(z)}{[H_1^*]^{(j)}(z)}\right) \exp\big\{-r\Lambda H_1^*(z)\big\} \prod_{i=1}^{k} \big(\big[H_1^*\big]^{(i)}(z)\big)^{m_i}. && (33)
\end{aligned}$$

Evaluating (32)–(33) at $z = 1$, premultiplying by $(1-\rho)^{k+1}$ and taking the limit for $\rho \uparrow 1$, we obtain

$$\lim_{\rho\uparrow 1}(1-\rho)^{k+1}\left[\frac{\partial}{\partial z}\left\{\exp\left\{-r\Lambda H_1^*(z)\right\}\prod_{i=1}^{k}\left(\left[H_1^*\right]^{(i)}(z)\right)^{m_i}\right\}\right]_{z=1}$$

$$= \lim_{\rho\uparrow 1}(1-\rho)\left(r\Lambda h_1^{(1)} + \sum_{j=1}^{k} m_j \frac{h_1^{(j+1)}}{h_1^{(j)}}\right)\exp\{-r\Lambda\}\lim_{\rho\uparrow 1}(1-\rho)^k\prod_{i=1}^{k}\left(h_1^{(i)}\right)^{m_i} \quad (34)$$

$$= \left(r\widehat{\Lambda}\tilde{h}_1^{(1)} + \sum_{j=1}^{k} m_j \frac{\tilde{h}_1^{(j+1)}}{\tilde{h}_1^{(j)}}\right)\exp\left\{-r\widehat{\Lambda}\right\}\prod_{i=1}^{k}\left(\tilde{h}_1^{(i)}\right)^{m_i} \quad (35)$$

$$= \hat{\lambda}_1\left(1-\hat{\rho}_1 I_{\{1\in E\}}\right)\left[r + k\left(\frac{b^{(2)}/b^{(1)}}{2\gamma}\right)\right]\exp\left\{-r\widehat{\Lambda}\right\}\prod_{i=1}^{k}\left(\tilde{h}_1^{(i)}\right)^{m_i}. \quad (36)$$

The first equality is obtained by evaluating (32)–(33) at $z = 1$, and using the definition of $h_1^{(k)}$ in (28). The second equality follows from the definition of $\tilde{h}_1^{(i)}$ in (28). The third equality follows from (30), and the fact that (by definition) for $\mathbf{m}^{(k)} \in S_k$ we have $\sum_{j=1}^{k} jm_j = k$, see (9). Finally, the proof of theorem 2 is completed by the following equations:

$$\tilde{x}_1^{(k+1)} = \lim_{\rho\uparrow 1}(1-\rho)^{k+1}\left[\frac{\partial}{\partial z}\sum_{\mathbf{m}^{(k)}\in S_k} c_k\left(\mathbf{m}^{(k)}\right)r^{l_k}\Lambda^{l_k}\right.$$
$$\left.\times\left\{\exp\left\{-r\Lambda H_1^*(z)\right\}\prod_{i=1}^{k}\left(\left[H_1^*\right]^{(i)}(z)\right)^{m_i}\right\}\right]_{z=1} \quad (37)$$

$$= \sum_{\mathbf{m}^{(k)}\in S_k} c_k\left(\mathbf{m}^{(k)}\right)\lim_{\rho\uparrow 1}(1-\rho)^{k+1}r^{l_k}\Lambda^{l_k}$$
$$\times\left[\frac{\partial}{\partial z}\left\{\exp\left\{-r\Lambda H_1^*(z)\right\}\prod_{i=1}^{k}\left(\left[H_1^*\right]^{(i)}(z)\right)^{m_i}\right\}\right]_{z=1} \quad (38)$$

$$= \hat{\lambda}_1\left(1-\hat{\rho}_1 I_{\{1\in E\}}\right)\left[r + k\left(\frac{b^{(2)}/b^{(1)}}{2\gamma}\right)\right]$$
$$\times\sum_{\mathbf{m}^{(k)}\in S_k} c_k\left(\mathbf{m}^{(k)}\right)r^{l_k}\widehat{\Lambda}^{l_k}\exp\left\{-r\widehat{\Lambda}\right\}\prod_{i=1}^{k}\left(\tilde{h}_1^{(i)}\right)^{m_i} \quad (39)$$

$$= \hat{\lambda}_1\left(1-\hat{\rho}_1 I_{\{1\in E\}}\right)\left[r + k\left(\frac{b^{(2)}/b^{(1)}}{2\gamma}\right)\right]\tilde{x}_1^{(k)}$$

$$= \hat{\lambda}_1^{k+1}\left(1-\hat{\rho}_1 I_{\{1\in E\}}\right)^{k+1}\prod_{j=0}^{k}\left[r + j\left(\frac{b^{(2)}/b^{(1)}}{2\gamma}\right)\right]. \quad (40)$$

The first equality follows from (25) and (27). The second equality is trivial, and the third equality follows from (34)–(36). The fourth equation follows from (29), and the last equality follows from the induction assumption. This completes the proof of theorem 2. □

5.2. Non-deterministic switch-over times

In the sequel we drop the assumption that all switch-over times are deterministic, and assume that the switch-over times have a general distribution (with finite moments). The following result shows that in the limiting case $\rho \uparrow 1$ the (scaled) moments of X_1 are *insensitive* to the higher moments of the switch-over times.

Theorem 3. For $i = 1, \dots, N,\ k = 1, 2, \dots$, the variable $\tilde{x}_1^{(k)}$ is independent of $r_i^{(l)}$ for $l \geqslant 2$.

Proof. For compactness of the presentation, we only give an outline of the proof. Similar to the decomposition in (18), the variables $s_{i,c}^{(k)}$ can be expressed in terms of the variables $\alpha_{i,c}^{(k)}$ as follows: for $i = 1, \dots, N,\ c = 0, 1, \dots, k = 1, 2, \dots$,

$$s_{i,c}^{(k)} = r_i^{(1)} \left[\sum_{j=i+1}^{N} \lambda_j \alpha_{j,c}^{(k)} + \sum_{j=1}^{i} \lambda_j \alpha_{j,c-1}^{(k)} \right] + \zeta_{i,c}^{(k)}, \tag{41}$$

where $\zeta_{i,c}^{(k)}$ is a linear combination of terms of the form

$$\begin{aligned} & r_i^{(l_k)} \prod_{l=1}^{k} \left(\sum_{j=i+1}^{N} \lambda_j \alpha_{j,c}^{(l)} + \sum_{j=1}^{i} \lambda_j \alpha_{j,c-1}^{(l)} \right)^{m_l}, \\ & \mathbf{m}^{(k)} = (m_1, \dots, m_k) \in S_k, \quad \text{with } l_k := \sum_{j=1}^{k} m_j > 1. \end{aligned} \tag{42}$$

Then, using (41), (42), (28), and the properties discussed in section 4.3, one may verify that

$$\begin{aligned} & \lim_{\rho \uparrow 1} (1-\rho)^k \sum_{i=1}^{N} \sum_{c=0}^{\infty} s_{i,c}^{(k)} = r \tilde{h}_1^{(k)} + 0 = r \tilde{h}_1^{(k)}, \\ & \sum_{i=1}^{N} \sum_{c=0}^{\infty} \zeta_{i,c}^{(k)} = \mathrm{O}\big((1-\rho)^{-(k-1)}\big), \quad \rho \uparrow 1. \end{aligned} \tag{43}$$

To proceed, by exploring (5) one may verify that $E[X_1^k]$ can be expressed as a linear combination of terms of the form $\prod_{p=1}^{k} \Gamma_p^{n_p}$, with $\underline{n}^{(k)} = (n_1, \dots, n_k) \in S_k$, where Γ_p is a linear combination of terms of the form

$$\sum_{i=1}^{N} \sum_{c=0}^{\infty} \prod_{j=1}^{p} \big(s_{i,c}^{(j)}\big)^{m_j}, \quad \text{with } \mathbf{m}^{(p)} = (m_1, \dots, m_p) \in S_p,\ p = 1, \dots, k.$$

For example, it is readily verified that $E[X_1] = \Gamma_1$, $E[X_1^2] = \Gamma_1^2 + \Gamma_2$ and $E[X_1^3] = \Gamma_1^3 + 3\Gamma_1\Gamma_2 + \Gamma_3$, with

$$\Gamma_1 := \sum_{i,c} s_{i,c}^{(1)}, \qquad \Gamma_2 := \sum_{i,c} s_{i,c}^{(2)} - \sum_{i,c}\bigl(s_{i,c}^{(1)}\bigr)^2, \quad \text{and} \quad \Gamma_3 := \sum_{i,c} s_{i,c}^{(3)} - \sum_{i,c}\bigl(s_{i,c}^{(1)}\bigr)^3$$

(see also [12], equations (3.9a), (3.9b) and (4.6), respectively). Now, by using (41), (42), and the properties in section 4.3, one may verify that, for $\mathbf{m}^{(p)} \in S_p$,

$$\sum_{i=1}^{N}\sum_{c=0}^{\infty}\prod_{j=1}^{p}\bigl(s_{i,c}^{(j)}\bigr)^{m_j} = \mathrm{O}\bigl((1-\rho)^{-(p-l_p+1)}\bigr), \quad \rho \uparrow 1. \tag{44}$$

Hence, $E[X_1^k]$ can be expressed as a linear combination of (cross-)terms of the form $\prod_{p=1}^{k}\Gamma_p^{n_p}$, $\mathbf{n}^{(k)} \in S_k$, that have a pole at $\rho = 1$ of the order

$$\sum_{p=1}^{k} n_p(p - l_p + 1) = k - \sum_{p=1}^{k} n_p(l_p - 1), \quad \text{with } \mathbf{n}^{(k)} \in S_k.$$

Then, premultiplying $E[X_1^k]$ by a factor $(1-\rho)^k$ and letting $\rho \uparrow 1$, all cross-terms for which there exists p $(1 \leqslant p \leqslant k)$ such that $n_p > 0$ and $l_p := \sum_{j=1}^{p} m_j > 1$, with $\mathbf{m}^{(p)} \in S_p$, vanish in the limiting case. The only terms that do not vanish in the limiting case are those for which $\sum_{p=1}^{k} n_p(l_p - 1) = 0$, i.e., for which for each p with $n_p > 0$ we have $l_p = 1$, or equivalently, $m_p = 1$, $m_j = 0$ for $j \leqslant p-1$. (Notice that this also implies that $\Gamma_p = \mathrm{O}((1-\rho)^{-p})$, $\rho \uparrow 1$, for $p = 1, \ldots, k$.) Consequently, $\tilde{x}_1^{(k)}$ (defined in (25)) can expressed as a linear combination of terms of the form $\prod_{p=1}^{k}\widetilde{\Gamma}_p^{n_p}$, $\mathbf{n}^{(k)} \in S_k$, where

$$\widetilde{\Gamma}_p = \lim_{\rho\uparrow 1}(1-\rho)^p \sum_{i=1}^{N}\sum_{c=0}^{\infty} s_{i,c}^{(p)} = r\tilde{h}_1^{(p)}, \quad \text{for } p = 1, \ldots, k$$

(which is in line with equation (29)). Theorem 3 then follows immediately from equations (43) and (30), and the fact that $r\tilde{h}_1^{(k)}$ is independent of $r_i^{(l)}$ for $l > 1$. □

5.3. *Main result*

We are now ready to obtain the moments of the asymptotic (scaled) waiting times at each of the queues.

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

$$\omega_1^{(k)} = \begin{cases} \dfrac{\tilde{x}_1^{(k+1)}}{(k+1)\hat{\lambda}_1^k \tilde{x}_1^{(1)}}\bigl(1 + \hat{\rho}_1 + \cdots + \hat{\rho}_1^k\bigr), & 1 \in G; \\[2ex] \dfrac{\tilde{x}_1^{(k+1)}}{(k+1)\hat{\lambda}_1^k \tilde{x}_1^{(1)}}, & 1 \in E. \end{cases} \tag{45}$$

Proof. Follows from (3) after several straightforward manipulations. □

Theorem 5. For $i = 1, \ldots, N$, $k = 1, 2, \ldots,$

$$\omega_i^{(k)} = \frac{(1 + \hat{\rho}_i + \cdots + \hat{\rho}_i^k) I_{\{i \in G\}} + (1 - \hat{\rho}_i)^k I_{\{i \in E\}}}{k+1} \times \prod_{j=1}^{k} \left[r + j \left(\frac{b^{(2)}/b^{(1)}}{1 - \sum_{m \in E} \hat{\rho}_m^2 + \sum_{m \in G} \hat{\rho}_m^2} \right) \right]. \tag{46}$$

Proof. Follows directly from theorems 2–4, where we assumed $i = 1$ without loss of generality. □

Remark 5.1. Theorem 5 is supported by a number of results in the literature. For the case of exhaustive service at all queues (i.e., $G = \emptyset$, $E = \{1, \ldots, N\}$), the results are in line with ones obtained in [3,7]. For the case $k = 1$, with general mixtures of gated and exhaustive service at the queues, the results correspond to the ones obtained in [33]. For the case $r = 0$, the results correspond to those in [28]. For the case of large switch-over times, the results are supported by the results [29].

Remark 5.2. An interesting observation is made by Srinivasan et al. [23], who show that in the case of deterministic switch-over times the complete waiting-time distributions depend on the individual switch-over time distributions only through r, i.e., the total expected switch-over time per cycle. For non-deterministic switch-over times, however, the waiting-time distributions generally *do* depend on the individual switch-over time distributions (characterized by $R_i^*(\cdot)$, $i = 1, \ldots, N$), even when r is kept fixed. In fact, the kth moments of the waiting times generally depend on the first $k + 1$ moments of all individual switch-over time distributions. In this perspective, the results in this paper show that the impact of the individual switch-over time distributions *vanishes* in heavy traffic, and as such can be viewed as "lower-order" effects.

Remark 5.3. Another interesting observation is made by Cooper et al. [9], who compare the performance of polling models with state-independent (SI) and state-dependent (SD) switch-over times. In this case of SD, the server incurs a switch-over time to move to queue and an additional setup timr in case the queue is not empty (to model the time needed by the server to "warm up" before starting to serve customers at that queue). Notice that in the model considered in the present paper, the setup times are assumed to be zero. Although naive intuition would lead one to believe that the SI should perform better than SD, the results in [9] show that (for the expected delay in a two-queue model with exhaustive service at both queues) the SD performs

(i) better if the switch-over times are constant and the setup times are non-deterministic,

(ii) worse if the switch-over times are non-deterministic and the setup times are constant, and

(iii) equally if both the switch-over times and setup time are constants.

The results in the present paper show that the relative differences in the expected delay between these models vanish in heavy traffic, and that the variability in the switch-over times does not play a role (see also section 6). This observation was to be expected, since the probability that a setup occurs when the server arrives at a queue tends to zero when the load tends to unity.

Remark 5.4. The derivation of theorem 5 via theorems 1–4 is rather straightforward. Alternatively, theorem 5 can be obtained by combining the following results: (1) expressions for the (scaled) moments of the delay in the case of zero switch-over times (cf. [28]), and (2) relations between the waiting times in systems with and nonzero switch-over times (cf. [4,23]). Denoting the delay at Q_1 in the zero switch-over time case by $W_1^{(0)}$, the waiting times in the case of nonzero switch-over times can be expressed as the independent sum $W_1 = W_1^{(0)} + Y_1$, where the LST of Y_1, denoted by $Y_1^*(s)$, can be expressed in terms of the DS variables as follows (see relations (39) and (29) in [23]): for $\operatorname{Re} s \geqslant 0$,

$$\begin{aligned} Y_1^*(s) &= \frac{X_1^*(B_1^*(s)) - X_1^*(1 - s/\lambda_1)}{\left[H_1^*(1 - s/\lambda_1) - H_1^*(B_1^*(s))\right] \Lambda r} \quad (1 \in G), \\ Y_1^*(s) &= \frac{1 - X_1^*(1 - s/\lambda_1)}{H_1^*(1 - s/\lambda_1)\Lambda r} \quad (1 \in E). \end{aligned} \tag{47}$$

Notice that these relations are *implicit* in the sense that the moments of Y_1 are *not* simple expressions in terms of the system parameters only, and *still* depend on the moments of X_1, which requires analysis of the system. The derivation of the results in theorem 5 from (47) requires the derivation of expressions for the higher-order derivatives of (47), applying theorem 3, and using the expressions for the asymptotic scaled moments for the case of zero switch-over times in [28]. Note, that the first step involves the determination of higher-order derivatives of functions of the form $f(x)/g(x)$, in addition to the higher-order derivatives of composite functions of the form $f(g(x))$, which is much more involved than the relatively simple derivation discussed in this paper.

6. Implications and illustration of the results

Theorem 5 reveals a variety of properties about the dependence of the moments of the asymptotic scaled delay with respect to the system parameters. These properties are discussed below. Numerical examples are presented to illustrate the validity of the results.

Property 6.1. For $i = 1, \dots, N,\ k = 1, 2, \dots,$

(a) $\omega_i^{(k)}$ is independent of the visit order,

(b) $\omega_i^{(k)}$ is independent of lth moment of the service-time distributions at the queues for $l > 2$,

(c) $\omega_i^{(k)}$ depends on the second moments of the service-time distributions only through the second moment of an arbitrary customer,

(d) $\omega_i^{(k)}$ depends on the moments of the switch-over time distributions only through the r, the first moment of the total switch-over time per cycle.

Property 6.1 is known to be not generally valid for stable systems (i.e., for $\rho < 1$), where the visit order and the third and higher moments of the service times *do* have an impact on the moments of the waiting-times; in fact, it is known that the kth moments of the delay generally depend on the first $k + 1$ moments of the service times. Hence, property 6.1 shows that the influence of these parameters on the waiting-time distributions *vanishes* when ρ tends to 1. A natural question is "how high should the load be" in order for the asymptotic results to be useful in practice. To investigate this, and to illustrate the validity of property 6.1, we have performed a variety of numerical experiments. To this end, we used a computer program corresponding to [6] to determine the moments of the delay in a variety of models. The results of the experiments are outlined below.

Consider a 10-queue model (referred to as model I) 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_2 and Q_6 gamma-distributed with mean 1 and squared coefficient of variation 10, while the service times at Q_4 and Q_8 are deterministically distributed with mean 25 and the service times at all other queues are exponentially distributed with mean 1; $E = \{2, 3, 5, 10\}$; $G = \{1, 4, 6, 7, 8, 9\}$; the switch-over time from Q_6 to Q_7 is 0.91, while all other switch-over times are 0.01. As can be seen, the model is highly asymmetric in the arrival rates, service times, service policies and switch-over times. For comparison, we also consider model II, which is identical to model I, except that the visit order is $(1, 8, 3, 4, 6, 5, 7, 2, 9, 10)$. According to part (a) of property 6.1, the relative difference between the moments of the waiting times vanish when ρ tends to 1. Table 1 shows the kth moment of the delay at Q_1 in both models for $k = 1, 2$ and 5, and for different values of ρ. The relative difference in the results is defined as follows:

$$\Delta := \mathrm{abs}\left(\frac{E[W_1^k(\mathrm{II})] - E[W_1^k(\mathrm{I})]}{E[W_1^k(\mathrm{I})]}\right) \times 100. \tag{48}$$

The notation "xey" stands for "x times 10^y". Table 1 shows that the relative differences in the moments of the waiting times vanish when the system tends to saturate, as expected on the basis of theorem 5. Thus, the impact of the visit order on the moments of the delay tends to "die out" when the system is heavily loaded. Notice also that the "rate" at which the relative error vanishes decreases for higher values of k. In other words, the minimal value of ρ for which the asymptotic results are visible increases for increasing k.

To illustrate the validity of parts (b) and (c) of property 6.1, we consider model III, which is identical to model I, except that the third and higher moments of the service times are different. More precisely, we assume that the service times at Q_1 to Q_9 are

Table 1
Moments of the waiting times in identical models, but with different visit orders.

ρ	$k = 1$			$k = 2$			$k = 5$		
	I	II	Δ	I	II	Δ	I	II	Δ
0.70	1.75e1	1.85e1	5.7	9.69e2	1.06e3	3.8	8.76e4	9.84e4	12.3
0.80	2.94e1	3.08e1	4.8	2.52e3	2.70e3	7.1	3.59e5	3.92e5	9.2
0.90	6.55e1	6.71e1	2.4	1.16e4	1.21e4	4.3	3.42e6	3.61e6	5.6
0.95	1.38e2	1.39e2	0.7	4.96e4	5.07e4	2.2	2.98e7	3.07e7	3.0
0.98	3.54e2	3.56e2	0.6	3.23e5	3.25e5	0.6	4.90e8	4.96e8	1.2
0.99	7.15e2	7.17e2	0.3	1.31e6	1.31e6	0.4	3.99e9	4.01e9	0.5

Table 2
Moments of the waiting times with different higher moments of the service times, but with the same $b^{(1)}$ and $b^{(2)}$.

ρ	$k = 1$			$k = 2$			$k = 5$		
	I	III	Δ	I	III	Δ	I	III	Δ
0.70	1.75e0	1.73e1	1.1	9.69e2	9.81e2	1.2	1.85e9	2.48e9	34.1
0.80	2.94e1	2.92e1	0.7	2.52e3	2.54e3	0.8	1.86e10	2.17e10	16.7
0.90	6.55e1	6.51e1	0.6	1.16e4	1.16e4	0.4	7.73e11	8.18e11	5.8
0.95	1.38e2	1.37e2	0.5	4.96e4	4.97e4	0.2	2.80e13	2.87e13	2.5
0.98	3.54e2	3.54e2	0.2	3.23e5	3.22e5	0.2	2.93e15	2.96e15	1.0
0.99	7.15e2	7.14e2	0.1	1.31e6	1.31e6	0.1	9.61e16	9.64e16	0.3

deterministic, while the service times at Q_{10} are gamma-distributed with mean 1 and squared coefficient of variation 21. Notice that both in model I and III we have $b^{(1)} = 85/37$ and $b^{(2)} = 1338/37$. Table 2 shows the kth moment of the delay at Q_1 in models I and III for $k = 1$, 2 and 5, and for different values of ρ. The relative difference is defined as in (48), where II is replaced by III.

Table 2 shows that the relative differences in the moments of the waiting times vanishes when ρ tends to 1, as expected from theorem 5. Thus, the impact of the higher moments of the service times indeed vanishes when the system in heavily loaded. Again, we observe that the relative difference decreases to 0 at a lower rate for higher values of k.

Remark 6.1. LaPadula and Levy [15] study "very large" polling systems with exhaustive and gated service at each of the queues. They propose an algorithm for transforming polling systems with a very large number of queues into an approximate system with a small number of queues. The proposed transformation is based on the following assumptions: (i) the expected delay is insensitive to the visit order, and (ii) the service-time distributions are the *same* for all queues. In this perspective, we observe that property 6.1 (part (a)) shows that assumption (i) is asymptotically correct when $\rho \uparrow 1$. In addition, parts (b) and (c) show that in the limiting case $\rho \uparrow 1$, assumption (ii) is obsolete: the expected delay *does* depend on the individual service-time distributions only through $b^{(1)}$ and $b^{(2)}$, the first two moments of the distribution of the aggregated service times (weighted proportionally to the arrival rates). Numerical results presented in [15] show that the expected delay in the approximate system lead to accurate approximations for the performance of the original system. Property 6.1 provides theoretical support for the approximative approach in [15].

7. Approximation

In general, the determination the moments of the delay (e.g., by using the numerical transform-inversion in [6]) may be very time consuming when the utilization of the system is high, especially for systems with a large number of queues. This raises the need for simple and fast-to-evaluate approximations for the moments of the delay in stable heavily-loaded systems. Based on theorem 5, we propose the following approximation for the kth moment of the delay at Q_i: for $\rho < 1$, $i = 1, \ldots, N$, $k = 1, 2, \ldots,$

$$E\left[W_i^k(app)\right] := \frac{(1+\rho_i+\cdots+\rho_i^k)I_{\{i\in G\}}+(1-\rho_i)^k I_{i\in E\}}}{(1-\rho)^k(k+1)} \times \prod_{j=1}^{k}\left[r+j\left(\frac{b^{(2)}/b^{(1)}}{1-\sum_{m\in E}\rho_m^2+\sum_{m\in G}\rho_m^2}\right)\right]. \tag{49}$$

Evidently, the time required to evaluate the approximations in (49) is negligible. Theorem 5 implies that the approximations (49) are asymptotically exact in the limiting case $\rho \uparrow 1$. A natural question is "how high should the load be" in order for the approximation to be accurate for systems with $\rho < 1$. To this end, we have performed numerical experiments to test the accuracy of the approximations for different values of the load of the system. The relative error of the approximation of the kth moment of the waiting times at Q_i is defined as follows:

$$\text{error\%} := \text{abs}\left(\frac{E[W_i^k(app)] - E[W_i^k]}{E[W_i^k]}\right) \times 100\%. \tag{50}$$

The results of the experiments are outlined below.

Consider a fully symmetric 5-queue model with exponentially distributed service times with mean 1, where the switch-over times from Q_3 to Q_4 are deterministic with mean 1, whereas all other switch-over times are 0. Figure 2 shows the relative error (defined in (50)) in the approximation for the kth moment of the waiting time at each queue for $k = 2, 3$ and 5, and as a function of the offered load, for the case of exhaustive service at all queues. To investigate the accuracy of the approximations for large, highly asymmetric systems, we also consider the 10-queue model (Model I) defined in section 6. Figure 2 shows the relative error in the approximation of the kth moment of the waiting time at Q_1 for $k = 1, 2, 3$ and 5, as a function of the load. The results in figures 2 and 3 reveal a number of observations. First, we observe that the approximations become more accurate when the load is increased, and that the relative error tends to 0 when the system tends to saturate. These observations were expected on the basis of the asymptotic

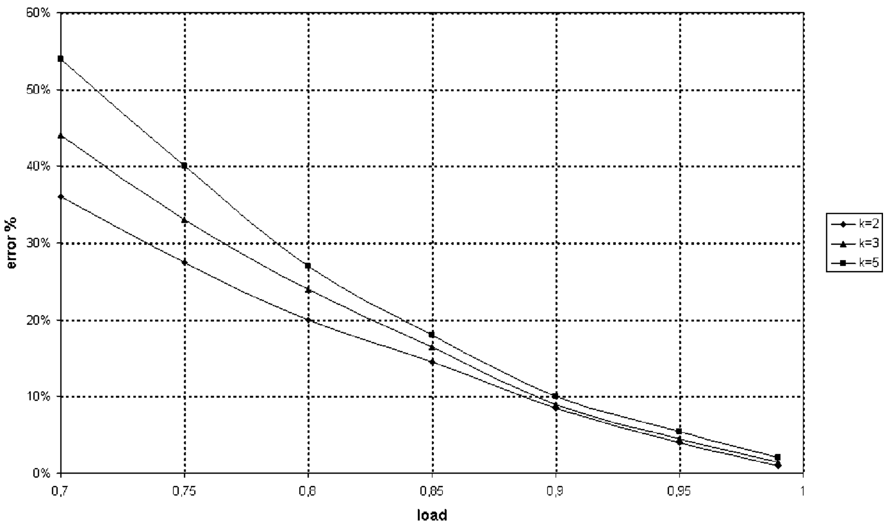

Figure 2. Accuracy of the approximations of moments of the delay as a function of the load in a fully symmetric 5-queue model.

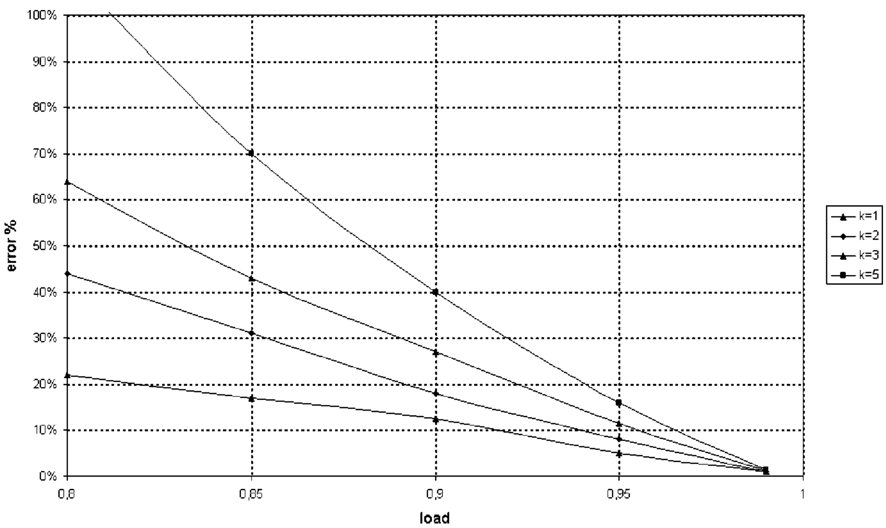

Figure 3. Accuracy of the approximations of moments of the delay as a function of the load in a highly asymmetric 10-queue model.

results shown in this paper (theorem 5). Second, we observe that the accuracy of the approximations tends to degrade for larger values of k; that is, the approximations tends to become less accurate for the higher moments of the delay. This observation was also to be expected, since deviations from the limiting waiting-time distribution (for $\rho \uparrow 1$) are "magnified" by taking higher moments.

To asses the usefulness of the approximations, and the practicality of the results, let us consider an approximation "accurate" if the relative error is less than 10%. The results in figure 2 indicate that for $k = 2$ the approximation is accurate when the load is 88% or more, while the results for $k = 3$ and $k = 5$ are accurate when the load exceeds 90%. The results in figure 3 are slightly less accurate. For $k = 1$ the approximation is accurate when the load is 92% or more, while for $k = 2, 3$, and 5 the load is accurate when the offered load exceeds 94%, 96% and 97%, respectively. Notice that the observed degradation of the accuracy of the approximations compared to the results in figure 1 was to be expected, because the model considered in figure 3 is highly asymmetric, and as such can be seen as a "worst-case" model. Summarizing, to give a rough indication of the accuracy of the approximations, the numerical results suggest that the approximations for the first few moments are accurate when the load is in the range 85–92% or larger, while for the approximations of the higher moments to be accurate, the load should typically be in the range 90–96% or more.

8. Topics for further research

In the derivation of the results we explicitly assume that all moments of the service times and switch-over times are finite. However, theorem 5 shows that the impact of the third and higher moments of the service times and the second and higher moments of the switch-over times vanishes in the limiting case. This suggests that the assumptions on the finiteness of higher the moments of the service and switch-over times may be relaxed. This observation is supported by the results in [7,8], where similar results are obtained under weaker assumptions on the finiteness of the moments. A related challenging area of further research is the derivation of heavy-traffic results when even the second moments of the service times are infinite. Boxma and Cohen [5] obtain exact expressions for the complete waiting-time distribution in a single-server model with different tail behaviors of the service-time distributions, under heavy-traffic assumptions. Generalization of these results to multi-queue polling models is an interesting topic for further research.

The model allows only exhaustive and gated service at each of the queues. It is an open question whether a similar analysis can be performed to derive closed-form expressions similar to (20) for models with other service disciplines. The approach presented in this paper is based on the DSA which, in turn, relies on the branching structure of the model. So, extension of the analysis to more general types of models seems to be possible for more general models with a branching structure (cf. [20]). Explicit expressions for the (scaled) expected delay have been derived for models with a more general class of branching-type service disciplines in [32].

In this paper we considered the moments of the (scaled) delay incurred at each of the queues. These results may be used to obtain expressions for the complete (scaled) waiting-time distributions in the limiting case. We refer to [30,31] for a derivation of the Laplace–Stieltjes Transform of the complete waiting-time distribution for the model discussed in the present paper.

In the present paper it is assumed that the arrival processes are Poisson. Although this assumption is rather realistic in some cases, in many applications it is not. Recent studies of traffic processes in telecommunications systems have revealed that traffic streams may be highly bursty and that the impact of autocorrelation of the system performance is significant. Therefore, it is for further research to extend the analysis to non-Poisson arrival streams. In this perspective, note that the heavy-traffic results presented in Coffman et al. [7,8] are valid for a class of non-Poisson arrival processes.

Acknowledgements

The author wishes to thank G.L. Choudhury for making the computer program corresponding to [6] available for performing the numerical experiments. The author also wishes to thank H. Levy for many interesting discussions on the topic. A partial and preliminary version of this paper appeared in [27].

References

[1] K.B. Athreya and P.E. Ney, *Branching Processes* (Springer, Berlin, 1971).
[2] J.P.C. Blanc, Performance analysis and optimization with the power-series algorithm, in: *Performance Evaluation of Computer and Communication Systems*, eds. L. Donatiello and R. Nelson (Springer, Berlin, 1993) pp. 53–80.
[3] J.P.C. Blanc and R.D. van der Mei, Optimization of polling systems with Bernoulli schedules, Performance Evaluation 22 (1995) 139–158.
[4] S.C. Borst and O.J. Boxma, Polling models with and without switchover times, Oper. Res. 45 (1997) 536–543.
[5] 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).
[6] G. Choudhury and W. Whitt, Computing transient and steady state distributions in polling models by numerical transform inversion, Performance Evaluation 25 (1996) 267–292.
[7] 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.
[8] 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.
[9] R.B. Cooper, S.-C. Niu and M.M. Srinivasan, Setups in polling models: does it make sense to set up if no work is waiting? J. Appl. Probab. 36 (1997) 585–592.
[10] M.J. Ferguson, Computation of the variance of the waiting times for token rings, IEEE J. Selected Areas Commun. 4 (1986) 775–782.
[11] C. Fricker and M.R. Jaïbi, Monotonicity and stability of periodic polling models, Queueing Systems 15 (1994) 211–238.
[12] 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.
[13] D.P. Kroese, Heavy traffic analysis for continuous polling models, J. Appl. Probab. 34 (1997) 720–732.
[14] S. Kudoh, H. Takagi and O. Hashida, Second moments of the waiting time in symmetric polling systems.
[15] C.A. LaPadula and H. Levy, Customer delay in very large multi-queue single-server systems, Performance Evaluation 26 (1996) 201–218.

[16] K.K. Leung, Cyclic service systems with probabilistically-limited service, IEEE J. Selected Areas Commun. 9 (1991) 185–193.
[17] H. Levy and M. Sidi, Polling models: applications, modeling and optimization, IEEE Trans. Commun. 38 (1991) 1750–1760.
[18] D. Markowitz, Dynamic scheduling of single-server queues with setups: A heavy traffic approach, Ph.D. thesis, Operations Research Center (MIT, Cambridge, MA, 1995).
[19] M.I. Reiman and L.M. Wein, Dynamic scheduling of a two-class queue with setups, Oper. Res. 46 (1998) 532–547.
[20] J.A.C. Resing, Polling systems and multitype branching processes, Queueing Systems 13 (1993) 409–426.
[21] D. Sarkar and W.I. Zangwill, Expected waiting time for nonsymetric cyclic queueing systems – Exact results and applications, Mangm. Sci. 35 (1989) 1463–1474.
[22] M.M. Srinivasan, H. Levy and A.G. Konheim, The individual station technique for the analysis of cyclic polling models, Naval Res. Logist. 73 (1993) 79–101.
[23] M.M. Srinivasan, S.-C. Niu and R.B. Cooper, Relating polling models with zero and nonzero switchover times, Queueing Systems 19 (1995) 149–168.
[24] H. Takagi, *Analysis of Polling Systems* (MIT Press, Cambridge, MA, 1986).
[25] 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.
[26] H. Takagi, Queueing analysis of polling models: progress in 1990–1994, in: *Frontiers in Queueing: Models, Methods and Problems*, ed. J.H. Dshalalow (CRC Press, Boca Raton, FL, 1997).
[27] R.D. van der Mei, Delay in polling systems in heavy traffic, in: *Proc. of SPIE Conf. on Performance and Control of Network Systems*, eds. W. Lai and R.B. Cooper, Boston, MA, 2–4 November 1998, pp. 168–175.
[28] R.D. van der Mei, Polling systems in heavy traffic: higher moments of the delay, Queueing Systems 31 (1999) 265–294.
[29] R.D. van der Mei, Delay in polling systems with large switch-over times, J. Appl. Probab. 36 (1999) 232–243.
[30] R.D. van der Mei, Waiting-time distributions in polling systems in heavy traffic, in: *Teletraffic Engineering in a Competitive World*, eds. P. Key and D. Smith (Elsevier, Amsterdam, 1999) pp. 325–334.
[31] R.D. van der Mei, Distribution of the delay in polling systems in heavy traffic, Performance Evaluation 38 (1999) 133–148.
[32] R.D. van der Mei and H. Levy, Polling systems in heavy traffic: exhaustiveness of the service policies, Queueing Systems 27 (1997) 227–250.
[33] R.D. van der Mei and H. Levy, Expected delay in polling systems in heavy traffic, Adv. in Appl. Probab. 30 (1998) 586–602.
[34] K.S. Watson, Performance evaluation of cyclic strategies – a survey, in: *Proc. of Performance '84*, ed. E. Gelenbe (North-Holland, Amsterdam, 1985) pp. 521–533.