J. Appl. Prob. **36,** 232–243 (1999)
Printed in Israel

DELAY IN POLLING SYSTEMS WITH LARGE SWITCH-OVER TIMES

R. D. VAN DER MEI,* *AT&T Labs*

Abstract

We study the delay in cyclic polling systems with mixtures of gated and exhaustive service, and with deterministic switch-over times. We show that, under proper scalings, the waiting-time distribution at each of the queues converges to a uniform distribution over a known interval when the switch-over times tend to infinity.

Keywords: polling systems; switch-over times; delay distribution; asymptotics

AMS 1991 Subject Classification: Primary 60M20; 60K25
Secondary 90B22

1. Introduction

The basic polling system consists of a number of queues and a single server which visits the queues in cyclic order to provide service to the customers waiting at the queues. Polling models find many applications in the areas of maintenance, manufacturing, production and computer-communication systems. During the last three decades, polling models have received much attention in the literature. We refer to [15] for an overview of the applicability of polling models, and to [22] for a fairly recent review of the state-of-the-art in the analysis of polling models. Most variations of polling models do not allow for an exact detailed analysis, and the others usually require the use of numerical techniques to determine performance measures of interest, like the distributions and the moments of the delay incurred at each of the queues. Moreover, numerical techniques can only contribute to the understanding of the behavior of the system to a limited extent. Explicit expressions provide much more insight into the dependence of the performance measures with respect to the system parameters. These observations raise the importance of an exact asymptotic analysis of the performance of polling models.

In this paper it is shown that a class of polling models allows an exact detailed asymptotic analysis when the switch-over times tend to infinity. We derive explicit expressions for the complete asymptotic scaled waiting-time distributions at each of the queues. The results lead to a variety of asymptotic properties which have not been observed before in the literature, providing new insights into the asymptotic behavior of polling systems with large switch-over times.

Polling systems with large switch-over times find applications in the areas of maintenance, flexible manufacturing and production systems. Mack *et al.* [16] use a polling model (with single buffers) to describe a patrolling repairman who inspects a number of machines to check whether a breakdown has occurred and if so, eliminates such breakdowns. Evidently, in polling models the repairman is represented by the server, the breakdowns are represented by the

Received 1 September 1997; revision received 26 January 1998.
* Postal address: Network Design and Performance Analysis Department, AT&T Labs, Advanced Technologies, 200 Laurel Avenue, Middletown, NJ 07748, USA.
Email address: rvandermei@att.com

customers and the times needed by the repairman to travel from one machine to the next are represented by the switch-over times. In [4] similar models are studied in which an operator at a fixed position serves a number of storage locations on a rotating carousel conveyor. Polling models are also used for the modelling of flexible manufacturing and production systems in which machines can be used to perform various types of tasks. Here, the server typically represents the machine, each of the queues represents a different type of job and the switch-over times represent the time needed by the machine to change from one type of operation to the other. Other typical applications of polling models with large switch-over times can be found, for example, in transportation networks [3], public transportation systems [6], mail delivery [17] and elevators [11].

The literature on polling models reveals a striking difference in complexity between different models. This distinction in complexity has been illuminated by Resing [18], who showed that for a class of polling models the joint queue-length process embedded at polling instants (i.e. the moments at which the server arrives at a queue) 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 satisfying an MTBP-structure several numerical algorithms have been proposed to determine the moments of the delay at the queues by solving sets of linear equations (see [21] for some references). Recently, the efficiency of the numerical techniques has been considerably improved by the so-called Descendant Set Approach (DSA). The DSA is an iterative technique which explores the MTBP-structure of the model by making use of the concept of so-called descendant sets (see [13]). Choudhury and Whitt [5] use numerical transform-inversion for the determination of tail probabilities of the waiting times. The key element in the identification of the class of polling models in [18] is that the service policy at each of the queues should satisfy a certain 'branching property'. This property is satisfied by the present model with mixtures of exhaustive and gated service policies. Polling models with service disciplines that do not have a branching structure (e.g. limited-type service disciplines) are usually not exactly analysable and generally require much more computational effort to obtain performance measures such as moments of the delay at each of the queues (see, e.g. [1, 14]).

The motivation of this paper is twofold. First, we have a theoretical interest in studying the impact of switch-over times on the delay incurred at each of the queues. This study aims to provide insights into the behavior of polling systems when the switch-over times get large. Second, polling models in which the switch-over times are large compared to the service times find a variety of specific applications (see above). Evidently, for those applications an exact analysis of the system is very useful.

We consider an asymmetric cyclic polling model with deterministic switch-over times and with general mixtures of exhaustive and gated service. We study waiting-time (and queue-length) distributions when the switch-over times tend to infinity. The waiting times are known to grow without bound when the switch-over times get very large. Therefore, we study the distribution of the scaled waiting-time, which is defined as the waiting time divided by the total switch-over time per cycle of the server along the queues, when the switch-over times tend to infinity. The key result is the observation that the distribution of the length of a queue at polling instants at that queue, divided by total switch-over time per cycle, converges almost surely to a known constant. The derivation of the result is based on the use of the DSA and application of the Strong Law of Large Numbers for Renewal Reward Processes. The result, in turn, is shown to imply that the scaled waiting-times distribution at each queue converges (in distribution) to a uniform distribution over a known interval when the switch-over times

tend to infinity. This leads to explicit expressions for the complete marginal distributions, and hence for the moments also, of the asymptotic scaled waiting-time at each of the queues. The results reveal a variety of properties of the asymptotic scaled waiting times, providing new insights into the behavior of the system. It is shown that the asymptotic scaled waiting-time distribution at each queue (i) is independent of the order in which the queues are visited, (ii) depends on the service-time distributions through their first moments, but is independent of the higher moments of the service times, and (iii) depends on the service discipline at that queue, but is independent of the service disciplines at all other queues.

The remainder of the paper is organized as follows. In Section 2 the model is described in detail. In Section 3 we review the principles of the DSA. In Section 4 the concept of descendant sets is used to derive explicit expressions for the asymptotic scaled waiting-time distributions. In Section 5 the results are illustrated by numerical examples. Finally, in Section 6 we discuss implications of the results and address a number of topics for further research.

2. Model description

Consider a system consisting of N infinite-buffer queues, $Q_1, \ldots, Q_N$, and a single server which 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 first and second moments b_i and $b_i^{(2)}$, respectively. The load offered to Q_i is $\rho_i = \lambda_i b_i$, and the total offered load is equal to $\rho = \sum_{i=1}^{N} \rho_i$. The moments at which the server arrives at Q_i are referred to as polling instants at Q_i. The time intervals during which the server visits at Q_i are called service periods at Q_i. The time interval between two successive polling instants at Q_1 is referred to as a cycle. The service at each queue is either according to the gated policy or the exhaustive policy. In 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. In 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 $G := \{i : Q_i \text{ is served according to the gated policy}\}$ and $E := \{i : Q_i \text{ is served exhaustively}\}$. At each queue the customers are served on a FIFO basis. After completing service at Q_i the server proceeds to Q_{i+1}, incurring a fixed switch-over period of length r_i. Define $r = \sum_{i=1}^{N} r_i$, i.e. the total switch-over time per cycle. 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 $\rho < 1$ (see [8]). In the following, it is assumed that this condition is satisfied, and that the system is in steady state.

Denote by W_i the delay incurred by an arbitrary customer at Q_i when the system is in steady state, and denote by $W_i^*(\cdot)$ the corresponding LST. Similarly, define X_i to be the number of customers at Q_i at an arbitrary polling instant at Q_i, and denote by $X_i^*(\cdot)$ the corresponding Probability Generating Function (PGF). Our main interest is in the behavior of W_i when the switch-over times tend to infinity. The waiting times are known to grow without bound when the switch-over times tend to infinity. Therefore, the analysis is oriented towards the determination of the distribution of

$$\tilde{W}_i := \lim_{r\to\infty} \frac{W_i}{r}, \quad \tilde{X}_i := \lim_{r\to\infty} \frac{X_i}{r}, \quad i = 1, \ldots, N, \tag{1}$$

referred to as the *asymptotic scaled* waiting time at Q_i, and the asymptotic scaled length of Q_i, at polling instants at Q_i, respectively.

Remark 1. It is shown in [19] (Corollary 1) that if the switch-over times are deterministic, then the distributions of X_i and W_i depend on the individual switch-over times $r_1, \ldots, r_N$ *only* through $r = \sum_{i=1}^{N} r_i$, the *total* switch-over time per cycle. This implies that the limits in (1) are well-defined, and that we can assume, without loss of generality, that $r_1 = \cdots = r_{N-1} = 0$, $r_N = r$.

3. The descendant set approach

The waiting-time distribution at Q_i and the queue-length distribution at polling instants at Q_i are related by the following equations (see [20]): for $i = 1, \ldots, N$, $\mathrm{Re}\, s \geq 0$,

$$W_i^*(s) = \frac{1-\rho}{r} \frac{X_i^*(B_i^*(s)) - X_i^*(1 - s/\lambda_i)}{s - \lambda_i + \lambda_i B_i^*(s)} \quad (i \in G),$$
$$W_i^*(s) = \frac{1-\rho}{r} \frac{1 - X_i^*(1 - s/\lambda_i)}{s - \lambda_i + \lambda_i B_i^*(s)} \quad (i \in E). \tag{2}$$

Hence, to determine the distribution of W_i it is sufficient to obtain the distribution of X_i. To this end, note that simple balancing arguments lead to the following closed-form expressions for the quantities $E[X_i]$ (e.g. [23]): for $i = 1, \ldots, N$,

$$E[X_i] = \frac{\lambda_i(1 - \rho_i I_{\{i \in E\}})r}{1-\rho}, \quad \text{and hence,} \quad E[\tilde{X}_i] = \frac{\lambda_i(1 - \rho_i I_{\{i \in E\}})}{1-\rho}. \tag{3}$$

However, the complete distribution, and the higher moments, of X_i cannot be obtained explicitly. The DSA provides a recursive scheme to determine the distribution of X_i and appears to be very useful for performing the asymptotic analysis in this paper. Throughout this section, we discuss the basic ideas of the DSA. The reader is referred to [13] for more details.

3.1. Terminology

All customers in a polling system can be classified into: (i) *originators*, and (ii) *non-originators*. An originator is a customer which arrives at the system during a switch-over period. A non-originator is a customer who arrives at the system during the service of another customer. For a customer C, let the *children set* 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*. The main ideas are the observations that (i) each of the $X_1(P^*)$ customers belongs to the descendant set of exactly *one* originator, and (ii) 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 then on calculating the number of type-1 descendants it has at P^*. Summing 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. Therefore, we number the successive polling instants as follows (counting backwards in time). Let $P_{N,0}$ be the last polling instant at Q_N prior to P^*, and for $i = N-1, \ldots, 1$, 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, 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 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 $\alpha^{(k)}_{i,c}$, $k = 1, 2, \ldots$. Based on the terminology discussed here, and on Remark 1, it is readily verified that the mean and the variance of X_1 can be expressed in terms of the variables $\alpha^{(1)}_{i,c}$ and $\alpha^{(2)}_{i,c}$ as follows (see also [13]):

$$E[X_1] = r \sum_{j=1}^{N} \sum_{c=0}^{\infty} \lambda_j \alpha^{(1)}_{j,c-1}, \quad \mathrm{Var}[X_1] = r \sum_{j=1}^{N} \sum_{c=0}^{\infty} \lambda_j \alpha^{(2)}_{j,c-1}. \tag{4}$$

Expressions for the higher moments of X_1 in terms of the moments of $A_{i,c}$ can be obtained in a similar way (see e.g. [13, 24, 19]).

3.2. Recursive scheme

The DSA is based on recursive relations between the variables $A_{i,c}$, and hence, on their moments $\alpha^{(k)}_{i,c}$. Fix k and consider a tagged (i, c)-customer, present at Q_i at $P_{i,c}$, denoted by $T_i(P_{i,c})$. We want to find the contribution of $T_i(P_{i,c})$ to $X_1(P^*)$. A key observation is that this contribution is equal to the total contribution to $X_1(P^*)$ of all *immediate children* of $T_i(P_{i,c})$, i.e. the customers which arrive during the service of $T_i(P_{i,c})$, leading to the following recursive relations for $A_{i,c}(z)$ (see [13]). For $|z| \leq 1$, $c = 0, 1, \ldots$,

$$A^*_{i,c}(z) = B^*_i\Big(\sum_{j=i+1}^{N} \lambda_j (1 - A^*_{j,c}(z)) + \sum_{j=1}^{i} \lambda_j (1 - A^*_{j,c-1}(z)) \Big) \quad (i \in G), \tag{5}$$

and

$$A^*_{i,c}(z) = \Theta^*_i\Big(\sum_{j=i+1}^{N} \lambda_j (1 - A^*_{j,c}(z)) + \sum_{j=1}^{i-1} \lambda_j (1 - A^*_{j,c-1}(z)) \Big) \quad (i \in E), \tag{6}$$

where $\Theta^*_i(\cdot)$ stands for the LST of the duration of a busy period in an $M/G/1$ system with rate λ_i and service-time distribution with LST $B^*_i(\cdot)$. The first two moments of Θ_i are known to be $\theta_i = b_i/(1 - \rho_i)$ and $\theta^{(2)}_i = b^{(2)}_i/(1 - \rho_i)^3$, respectively. To obtain the initial conditions, notice that if $T_i(P_{i,c})$ has not yet been served at $P_{k,0}$, then $T_i(P_{i,c})$ contributes 1 to $X_1(P^*)$ if $i = 1$ and 0 if $i \neq 1$. Hence, the initial conditions are as follows (see [13]): $A^*_{1,-1}(z) := z$; $A^*_{i,-1}(z) := 1$ $(i = 2, \ldots, N)$. In this way, the distribution of the variables $A_{i,c}$, and their corresponding moments, can be determined recursively. Focusing on the first moments, the variables $\alpha^{(1)}_{i,c}$ can be recursively calculated from the following relations, which follow directly from (5) and (6) (see [13]): for $c = 0, 1, \ldots$,

$$a^{(1)}_{i,c} = b_i\Big[\sum_{j=i+1}^{N} \lambda_j \alpha^{(1)}_{j,c} + \sum_{j=1}^{i} \lambda_j \alpha^{(1)}_{j,c-1} \Big] \quad (i \in G),$$

$$a^{(1)}_{i,c} = \theta_i\Big[\sum_{j=i+1}^{N} \lambda_j \alpha^{(1)}_{j,c} + \sum_{j=1}^{i-1} \lambda_j \alpha^{(1)}_{j,c-1} \Big] \quad (i \in E). \tag{7}$$

Similarly, for the second moments we have the following recursive relations: for $i = 1, \ldots, N$, $c = 0, 1, \ldots,$

$$\alpha_{i,c}^{(2)} = \frac{b_i^{(2)}}{b_i^2}(\alpha_{i,c}^{(1)})^2 + b_i\left[\sum_{j=i+1}^{N} \lambda_j \alpha_{j,c}^{(2)} + \sum_{j=1}^{i} \lambda_j \alpha_{j,c-1}^{(2)}\right] \quad (i \in G), \tag{8}$$

and

$$\alpha_{i,c}^{(2)} = \frac{\theta_i^{(2)}}{\theta_i^2}(\alpha_{i,c}^{(1)})^2 + \theta_i\left[\sum_{j=i+1}^{N} \lambda_j \alpha_{j,c}^{(2)} + \sum_{j=1}^{i-1} \lambda_j \alpha_{j,c-1}^{(2)}\right] \quad (i \in E). \tag{9}$$

The initial conditions are $\alpha_{1,-1}^{(1)} = 1$, $\alpha_{i,-1}^{(1)} = 0$ $(i = 2, \ldots, N)$, and $\alpha_{1,-1}^{(2)} = 1$, $\alpha_{i,-1}^{(2)} = 0$ $(i = 2, \ldots, N)$. Based on these conditions, the variables $\alpha_{i,c}^{(1)}$ and $\alpha_{i,c}^{(2)}$ can be computed recursively according to (7)–(9). The higher moments of $A_{i,c}$ can be computed recursively in a similar manner.

4. Analysis

In this section we use the DSA to derive asymptotic results for the distribution of X_1 when r tends to infinity. Based on the concept of descendant sets, X_1 is known to be the sum of a number of independent random variables (namely, the contributions to $X_1(P^*)$ of the original customers), where this number increases to infinity when r tends to infinity. Based on this observation, the Strong Law of Large Numbers for Renewal Reward Processes is shown to imply that X_1/r converges (almost surely) to a known constant when r tends to infinity. This, in turn, it is shown to imply that the distribution of W_i/r converges to a uniform distribution over a known interval when r tends to infinity.

A family of random variables $\{A_\alpha, \alpha \geq 0\}$ is said to converge *almost surely* to a constant a for $\alpha \to \infty$, denoted as $A_\alpha \overset{\text{a.s.}}{\to} a (\alpha \to \infty)$, if for all $\epsilon > 0$, $P(|A_\alpha - a| > \epsilon) \to 0$ for $\alpha \to \infty$. A family of random variables $\{A_\alpha, \alpha \geq 0\}$ is said to converge *in distribution* to a random variable A for $\alpha \to \infty$, denoted as $A_\alpha \overset{\text{d}}{\to} A (\alpha \to \infty)$, if for all $\epsilon > 0$, $P(|A_\alpha - A| > \epsilon) \to 0$ for $\alpha \to \infty$.

Theorem 1. *For $i = 1, \ldots, N$,*

$$\frac{X_i}{r} \overset{\text{a.s.}}{\to} \frac{\lambda_i(1 - \rho_i I_{\{i \in E\}})}{1 - \rho} \quad (r \to \infty). \tag{10}$$

Proof. Following the terminology discussed in Section 3.1, it suffices to determine the distribution of $X_1 = X_1(P^*)$, i.e. the number of customers at Q_1 at the reference point at Q_1. Recall from Remark 1 that $r_1 = r_2 = \cdots = r_{N-1} = 0$ and $r_N = r$, without loss of generality. Define the cth cycle be the time interval between the polling instants $P_{c,1}$ and $P_{c-1,1}$, $c = 0, 1, \ldots$. Denote by R_c the switch-over period during the cth cycle. In this way, the cth cycle consists of the successive service periods starting at $P_{c,1}, \ldots, P_{c,N}$ (with no switch-over times between these service periods), followed by the switch-over period R_c. Defining Y_c to be the total contribution to $X_1(P^*)$ of all (original) customers which arrive during R_c, we can write

$$X_1 = \sum_{c=0}^{\infty} Y_c. \tag{11}$$

Based on the concept of descendant sets, the random variables Y_c are mutually *independent*. The distribution of Y_c is obtained by conditioning on the queue at which an original customer arrives. To this end, denote by $N_c^{(i)}$ the number of (original) customers which arrive at Q_i during R_c, and moreover, denote by $Y_c^{(i)}$ the total contribution to $X_1(P^*)$ of all $N_c^{(i)}$ type-i originators which arrive during R_c. Hence, we can write: for $c = 0, 1, \ldots,$

$$Y_c = \sum_{i=1}^{N} Y_c^{(i)}, \quad \text{with} \quad Y_c^{(i)} = \sum_{k=1}^{N_c^{(i)}} Y_c^{(i)}(k), \tag{12}$$

where $Y_c^{(i)}(k)$ stands for the contribution to $X_1(P^*)$ of the kth customer which arrives to Q_i during R_c, $k = 1, \ldots, N_c^{(i)}$. Based on the concept of descendant sets, the random variables $Y_c^{(i)}(k)$ are *independent*. We make the following observations: (i) the arrival process of (original) customers which arrive at Q_i during R_c is a Poisson process and hence, is a renewal process, (ii) $N_c^{(i)}$ has a Poisson distribution with mean $r\lambda_i$; and (iii) $Y_c^{(i)}(k)$ has the same distribution as $A_{i,c-1}$, for all k. Combining these observations, and application of the Strong Law of Large Numbers for Renewal Reward Processes (see e.g. [12, Theorem 6.4]) leads to the following result (see Section 6 for some additional remarks): for $i = 1, \ldots, N$, $c = 0, 1, \ldots,$

$$\frac{Y_c^{(i)}}{r} \overset{\text{a.s.}}{\to} \lambda_i \alpha_{i,c-1}^{(1)} \quad (r \to \infty). \tag{13}$$

To proceed, we have to show (see (11)) that the infinite series $\sum_{c=0}^{\infty} Y_c$ converges. To this end, we apply Theorem 2 in [7, Chapter 7, pp. 238], which requires that $\sum_{c=0}^{\infty} E[Y_c^2] < \infty$. This requirement is readily verified to be equivalent to $\sum_{c=0}^{\infty} E[A_{i,c}^2] < \infty$ for all i, which, in turn, is equivalent to $\sum_{c=0}^{\infty} \alpha_{i,c}^{(2)} < \infty$ for all i. To see that the latter requirement is satisfied, it is readily verified from equations (5) and (6) (by multiplication with λ_i and summation over i and c) that $\sum_{c=0}^{\infty} \alpha_{i,c}^{(1)} < \infty$ for all i, which implies that $\sum_{c=0}^{\infty} (\alpha_{i,c}^{(1)})^2 < \infty$ for all i. Equations (7)–(9) can then be used (again by multiplication with λ_i and summation over i and c, see also equations (3.10) and (3.11) in [13]) to show that $\sum_{c=0}^{\infty} \alpha_{i,c}^{(2)} < \infty$ for all i. This implies that $\sum_{c=0}^{\infty} E[Y_c^2] < \infty$, which is a sufficient condition for application of Theorem 2 in [7], which yields the following result: for $i = 1, \ldots, N$,

$$\sum_{c=0}^{\infty} \frac{Y_c^{(i)}}{r} \overset{\text{a.s.}}{\to} \sum_{c=0}^{\infty} \lambda_i \alpha_{i,c-1}^{(1)} \quad (r \to \infty). \tag{14}$$

Combining relations (11)–(13), (4) and (3) implies relation (10) for $i = 1$, which completes the proof.

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

$$\frac{W_i}{r} \overset{\text{d}}{\to} \tilde{W}_i \quad (r \to \infty), \tag{15}$$

where $\tilde{W}_i$ is uniformly distributed over the interval $[\tilde{a}_i, \tilde{b}_i]$, with

$$\tilde{a}_i = \frac{\rho_i}{1-\rho}, \quad \tilde{b}_i = \frac{1}{1-\rho} \quad (i \in G) \quad \text{and} \quad \tilde{a}_i = 0, \quad \tilde{b}_i = \frac{1-\rho_i}{1-\rho} \quad (i \in E). \tag{16}$$

Proof. First we introduce some notation. Let $\hat{W}_i := W_i/r$ be the scaled delay at Q_i, and let $\hat{W}_i^*(\cdot)$ be the corresponding LST. Moreover, let $\hat{X}_i := X_i/r$ and denote by $\hat{X}_i^*(\cdot)$ its corresponding PGF. Consider the case of exhaustive service at Q_i. Then to prove Theorem 2, it is sufficient to show that: for $i = 1, \ldots, N$, $\mathrm{Re}\, s \geq 0$,

$$\lim_{r\to\infty} \hat{W}_i^*(s) = \frac{1}{s\tilde{b}_i}(1 - \mathrm{e}^{-s\tilde{b}_i}), \tag{17}$$

where the right-hand side is the LST of the distribution of $\tilde{W}_i$, i.e. the uniform distribution on the interval $[0; \tilde{b}_i]$. Recall from (16) that $\tilde{b}_i = (1-\rho_i)/(1-\rho)$. To see that this is indeed the case, we observe that Theorem 1 implies that $\hat{X}_i^*(z) = X_i^*(z^{1/r}) \to z^{\lambda_i \tilde{b}_i}$ $(r \to \infty)$, for all $|z| \leq 1$. This, in turn, is easily shown to imply that, for $i = 1, \ldots, N$, $\mathrm{Re}\, s \geq 0$,

$$\lim_{r\to\infty} X_i^*\left(1 - \frac{s}{\lambda_i r}\right) = \mathrm{e}^{-s\tilde{b}_i}. \tag{18}$$

The validity of Theorem 2 then follows from the following equalities: for $i = 1, \ldots, N$, $\mathrm{Re}\, s \geq 0$,

$$\lim_{r\to\infty} \hat{W}_i^*(s) = \lim_{r\to\infty} W_i^*(s/r) = \lim_{r\to\infty} \frac{1-\rho}{r} \frac{1 - X_i^*(1 - s/\lambda_i r)}{s/r - \lambda_i + \lambda_i B_i^*(s/r)} = \frac{1}{s\tilde{b}_i}(1 - \mathrm{e}^{-s\tilde{b}_i}). \tag{19}$$

The first equality follows directly from the fact that $\hat{W}_i^*(s) = E[\mathrm{e}^{-sW_i/r}] = W_i^*(s/r)$. The second equality follows directly from (2), and the third equality is directly obtained from (18) and some straightforward manipulations. This completes the proof of Theorem 2 for the case of exhaustive service at Q_i. The validity of Theorem 2 for the case of gated service at Q_i can be shown in a similar way.

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

$$E[\tilde{W}_i^k] = \frac{1}{k+1} \frac{1 + \rho_i + \cdots + \rho_i^k}{1-\rho} \quad (i \in G),$$
$$E[\tilde{W}_i^k] = \frac{1}{k+1} \frac{(1-\rho_i)^k}{1-\rho} \quad (i \in E). \tag{20}$$

Proof. This follows directly from Theorem 2.

5. Numerical examples

In this section we illustrate the validity of the results. Consider the model with the following parameters: $N = 4$; the ratios between the arrival rates are 4:4:1:1; the service times are exponentially distributed with means $b_1 = 7$, $b_2 = b_3 = b_4 = 1$; $G = \{1, 2\}$, $E = \{3, 4\}$. We consider two models. In model I the load offered to the system is $\rho = 0.3$, and in model II the offered load is $\rho = 0.8$. Recall from Remark 1 that, for given r, the individual switch-over times do not need to be specified. As can be seen, the model is fairly asymmetrical in the arrival rates, the service times and the service disciplines.

We use a numerical tool based on the DSA combined with numerical transform-inversion [5] to compute the probability density function of the delay at the queues. Figure 1 shows

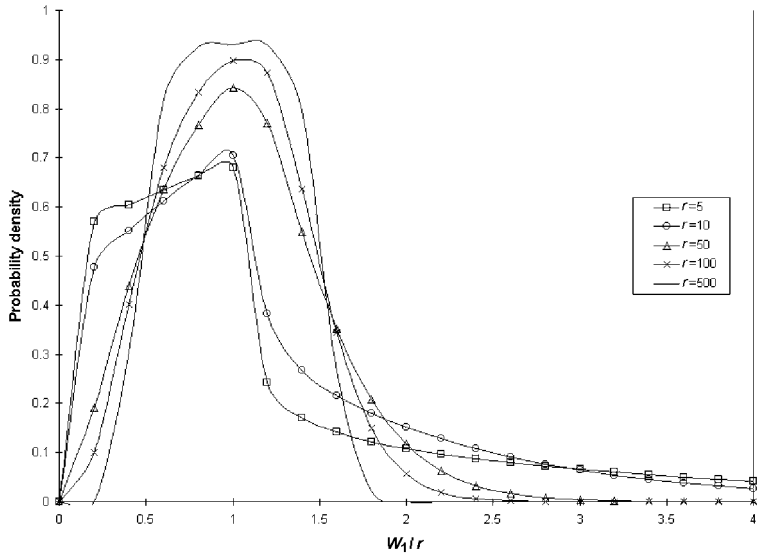

FIGURE 1: Probability density function of the scaled delay for different switch-over times; $\rho = 0.3$.

the probability density function of W_1/r (i.e. the delay at Q_1 divided by r) for different values of r, under Model I. Similarly, Figure 2 shows the probability density function of W_1/r for different values of r, under Model II.

Figures 1 and 2 suggest that the probability distribution of the scaled delay indeed converges to a uniform distribution when r becomes large, as expected on the basis of Theorem 2.

The probability distribution in the limiting case $r \to \infty$ can be determined from Theorem 2. For model I we have $\rho_1 = 21/85$ and $\rho = 0.3$, so that $\tilde{a}_1 = 6/17 \approx 0.3529$ and $\tilde{b}_1 = 10/7 \approx 1.4286$. Similarly, for model II we have $\rho_1 = 56/85$ and $\rho = 0.8$, so that $\tilde{a}_1 = 56/17 \approx 3.2941$ and $\tilde{b}_1 = 5$. The correctness of the asymptotic bound is supported by Figures 1 and 2.

We observe that the shape of the delay distribution for small values of r may differ strongly. Moreover, we notice that in the limiting case $r \to \infty$ the delay is uniformly distributed and has a finite support, whereas for finite r the delay distribution has an infinite support and the tail of the waiting-time distribution may be non-negligible, as illustrated in Figures 1 and 2. This difference in tail behavior between the case of finite r and the limiting case may become apparent if the limiting case is used as an approximation of the moments of the delays in models with finite r.

Comparison of Figures 1 and 2 suggests that the probability density of W_1/r, as function of r, approaches the limiting distribution considerably 'faster' when the offered load is low. Apparently, the rate of convergence to the limiting distribution decreases when ρ increases. This observation may be caused by the fact that for high values of the load the higher moments of $A_{i,c}$ are generally larger, so that the left-hand side of (13) converges more slowly to its limiting value.

6. Implications of the results

The results in Section 4 reveal several properties about how the delay distributions depend on the system parameters. These properties are discussed below.

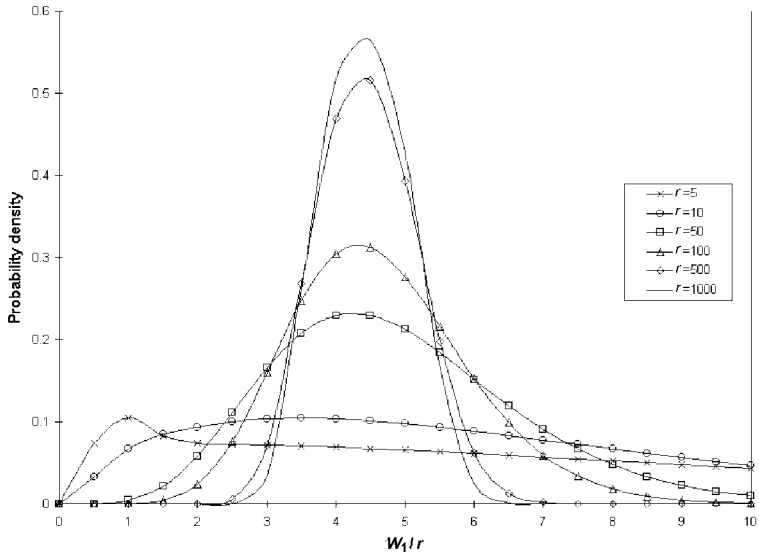

FIGURE 2: Probability density function of the scaled delay for different switch-over times; $\rho = 0.8$.

Lemma 1. *For $i = 1, \ldots, N$, the distribution of $\tilde{W}_i$ is independent of*

1. *the visit order;*
2. *the higher moments of the service-time distributions;*
3. *the service discipline at Q_j (for all $j \neq i$).*

Proof. This follows directly from Theorem 2.

The properties in Lemma 1 are known to be not generally valid for finite r. In this case, the waiting-time distribution at each queue generally *does* depend on the visit order, the higher moments of the service-time distributions and the service disciplines at the other queues. In this perspective, Lemma 1 reveals that the influence of the visit order, the higher moments of the service-time distributions and the service disciplines at the other queues vanish when the switch-over times get large.

Notice that similar insensitivity properties hold for $E[V] = \sum_{i=1}^{N} \rho_i E[W_i]$, i.e. the expected *total* amount of waiting work in the system. The results obtained in [2] show that $E[V]$ is independent of the visit order, the higher moments of the service times and the service disciplines, even for finite r.

Remark 2. The main result of this paper is Theorem 1. The key observation in the proof of Theorem 1 is relation (13), which is based on the application of the Strong Law of Large Numbers for Renewal Reward Processes (RRPs). To clarify this in some more detail, note that an RRP consists of two components (see, e.g. [12, Chapter 6]): (i) a renewal process and (ii) a reward distribution. In the proof of (13) we consider an RRP, where the renewal process is the process of arrivals of original customers at Q_i (which is a Poisson process with rate λ_i per time unit) during the period R_c (of duration r), and where the reward distribution is the distribution of $A_{i,c-1}$. Thus, the total 'reward' over a period of length r, which can be seen as

the total contribution of $X_1(P^*)$ of all original customers which arrive at Q_i during R_c, can be expressed as, for $i = 1, \ldots, N$, $c = 0, 1, \ldots$,

$$Y_c^{(i)}(1) + \cdots + Y_c^{(i)}(N_c^{(i)}), \tag{21}$$

where $N_c^{(i)}$ has a Poisson distribution with mean $r\lambda_i$, and where $Y_c^{(i)}(k)$ are i.i.d. with the same distribution as $A_{i,c-1}$, for all k. The Strong Law of Large Numbers for RRPs basically states that the average reward per time unit incurred over a very long time interval 'flattens out' and converges (almost surely) to its expected value. Accordingly, the contribution to X_1 of all customers which arrive at Q_i during R_c per time unit converges almost surely to the mean value $\lambda_i E[A_{i,c-1}]$, which implies relation (13).

Remark 3. Although the results are quite intuitive, their derivations are not trivial in the sense that the switch-over times do not 'completely dominate' the waiting times (which would be the case, for example, when $\rho \to 0$). Note that the fraction of the time in which the server is serving customers is ρ, independent of the switch-over times. To pursue this somewhat further, we observe that the scaled queue-length process (i.e. the process of queue lengths divided by r) tends to follow a deterministic pattern when r gets large. This is caused by the following two observations. First, the impact of the switch-over times on the evolution of the (scaled) queue-length process tends to become deterministic when r gets large, in the sense that the scaled numbers of customers arriving at each of the queues during the switch-over times tend to become deterministic. Second, the impact of the service periods on the evolution of the scaled queue-length process tends to become deterministic when r (and hence the number of customers at a queue at polling instants at that queue) gets large. This is caused by the branching structure of the gated and exhaustive service disciplines and application of the Theory of Large Numbers. Note that this is not necessarily the case for service disciplines in which the branching structure is violated (e.g. for probabilistically-limited service disciplines [14]). Thus, the impact of *both* the switch-over times and the service periods (and not the switch-over times only) on the evolution of the scaled queue-length process imply that the scaled queue-length process itself tends to become deterministic when r gets large.

We conclude with a number of topics for further research.

The derivation of the results presented here, particularly the proof of Theorem 1, relies on the assumption that the switch-over times are deterministic. However, when the switch-over times are non-deterministic, the key relation (13) is no longer generally valid. To see this, suppose for example that $R_c = 0$ with probability $\frac{1}{2}$ and $R_c = 2r$ with probability $\frac{1}{2}$ (so that $E[R_c] = r$ remains the same). Then, even if $r \to \infty$, we have $Y_c^{(i)} = 0$ with probability $\frac{1}{2}$, and we have $Y_c^{(i)}/r \to \lambda_i \alpha_{i,c-1}^{(1)}$ with probability $\frac{1}{2}$. Hence, relation (13), and, in turn, Theorem 1, are no longer valid. Notice also that the assumptions in Remark 1 are no longer valid without loss of generality, so that $\tilde{W}_i$ and $\tilde{X}_i$ are no longer well-defined in (1). Extension of the results to non-deterministic switch-over times is for further research.

In this paper we have derived expressions for the marginal waiting-time distribution at the different queues, based on the observation that the number of customers at Q_i at polling instants at Q_i becomes deterministic when the switch-over times tend to infinity. We suspect that when the switch-over times get large, the number of customers at Q_j at polling instants at Q_i also becomes deterministic for *all* j, which may lead to expressions for the *joint* asymptotic queue-length and waiting-time distributions. This is left as an open area for further research.

The observation that the waiting-time distribution tends to become only weakly independent of the parameters of the other queues (only through ρ) greatly simplifies the analysis. This, in turn, may be useful for optimization of the system performance with respect to the service disciplines at the queues. Optimization of polling models with large switch-over times is an open issue.

Acknowledgements

The author would like to thank G. L. Choudhury for making the software from [5] available for performing the numerical experiments, and D. P. Heyman for useful discussions about Laws of Large Numbers.

References

[1] Blanc, J. P. C. (1992). Performance evaluation of polling systems by means of the power-series algorithm. *Ann. Oper. Res.* **35,** 155–186.
[2] Boxma, O. J. and Groenendijk, W. P. (1988). Pseudo-conservation laws in cyclic service systems. *J. Appl. Prob.* **24,** 949–964.
[3] Bozer, Y. A. and Srinivasan, M. M. (1991). Tandem configurations for automated guided vehicle systems and the analysis of single loops. *IIE Trans.* **23,** 72–82.
[4] Bunday, B. D. and El-Badri, W. K. (1988). The efficiency of M groups of machines served by a travelling robot: comparison of two models. *Internat. J. Prod. Res.* **26,** 299–308.
[5] Choudhury, G. and Whitt, W. (1996). Computing transient and steady state distributions in polling models by numerical transform inversion. *Perf. Eval.* **25,** 267–292.
[6] Dukhovnyy, I. M. (1979). An approximate model of motion of urban passenger transportation over annular routes. *Eng. Cybern.* **17,** 161–162.
[7] Feller, W. (1971). *An Introduction to Probability Theory and Its Applications*, 2nd edn. Wiley, New York.
[8] Fricker, C. and Jaïbi, M. R. (1994). Monotonicity and stability of periodic polling models. *Queueing Systems* **15,** 211–238.
[9] Fuhrmann, S. W. (1992). A decomposition result for a class of polling models. *Queueing Systems* **11,** 109–120.
[10] Fuhrmann, S. W. and Cooper, R. B. (1985). Stochastic decompositions in the $M/G/1$ queue with generalized vacations. *Operat. Res.* **33,** 1117–1129.
[11] Gamse, B. and Newel, G. F. (1982). An analysis of elevator operation in moderate height buildings—A single elevator. *Transp. Res. B* **16,** 303–319.
[12] Heyman, D. P. and Sobel, M. J. (1982). *Stochastic Models in Operations Research*, Vol. 1. McGraw-Hill, New York.
[13] Konheim, A. G., Levy, H. and Srinivasan, M. M. (1994). Descendant set: an efficient approach for the analysis of polling systems. *IEEE Trans. Commun.* **42,** 1245–1253.
[14] Leung, K. K. (1991). Cyclic service systems with probabilistically-limited service. *IEEE J. Sel. Areas Commun.* **9,** 185–193.
[15] Levy, H. and Sidi, M. (1991). Polling models: applications, modeling and optimization. *IEEE Trans. Commun.* **38,** 1750–1760.
[16] Mack, C., Murphy, T. and Webb, N. L. (1957). The efficiency of N machines unidirectionally patrolled by one operative when walking times and repair times are constants. *J. Roy. Statist. Soc.* B **19,** 166–172.
[17] Nahmias, S. and Rothkopf, M. H. (1984). Stochastic models for internal mail delivery systems. *Management Sci.* **30,** 1113–1120.
[18] Resing, J. A. C. (1993). Polling systems and multitype branching processes. *Queueing Systems* **13,** 409–426.
[19] Srinivasan, M. M., Niu, S.-C. and Cooper, R. B. (1995). Relating polling models with zero and nonzero switch-over times. *Queueing Systems* **19,** 149–168.
[20] Takagi, H. (1986). *Analysis of Polling Systems*. MIT Press, Cambridge, MA.
[21] Takagi, H. (1990). Queueing analysis of polling models: an update. In *Stochastic Analysis of Computer and Communication Systems*, ed. H. Takagi. North-Holland, Amsterdam, 267–318.
[22] Takagi, H. (1997). Queueing analysis of polling models: progress in 1990–1994. In *Frontiers in Queueing: Models and Applications in Science and Technology*, ed. J. H. Dshalalow. CRC Press, Boca Raton, FL, pp. 119–144.
[23] Van der Mei, R. D. and Levy, H. (1998). Mean delay analysis of polling systems in heavy traffic. *Adv. Appl. Prob.* **30,** 586–602.
[24] Van der Mei, R. D. (1998). Polling systems in heavy traffic: higher moments of the delay. To appear in *Queueing Systems*.