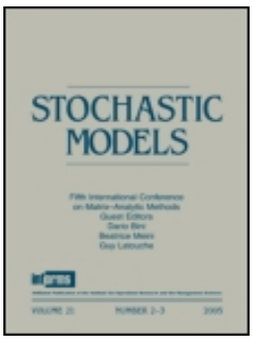

Taylor & Francis
Taylor & Francis Group

ISSN: 1532-6349 (Print) 1532-4214 (Online) Journal homepage: https://www.tandfonline.com/loi/lstm20

POLLING SYSTEMS WITH SIMULTANEOUS BATCH ARRIVALS

R. D. van der Mei

To cite this article: R. D. van der Mei (2001) POLLING SYSTEMS WITH SIMULTANEOUS BATCH ARRIVALS, , 17:3, 271-292, DOI: 10.1081/STM-100002274

To link to this article: https://doi.org/10.1081/STM-100002274

Published online: 15 Feb 2007.

Submit your article to this journal

Article views: 38

View related articles

Citing articles: 2 View citing articles

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=lstm20

STOCHASTIC MODELS, 17(3), 271–292 (2001)

POLLING SYSTEMS WITH SIMULTANEOUS BATCH ARRIVALS

R. D. van der Mei

KPN Research, Leidschendam, The Netherlands, and Free University of Amsterdam, Amsterdam, The Netherlands

ABSTRACT

We study the delay in polling systems with simultaneous batch arrivals. Arrival epochs are generated according to a Poisson process. At any arrival epoch, batches of customers may arrive simultaneously at the different queues, according to a general joint batch-size distribution. The server visits the queues in cyclic order, the service times and the switch-over times are generally distributed, and the service disciplines are general mixtures of gated and exhaustive service. We derive closed-form expressions for the expected delay at each of the queues when the load tends to unity (under proper scalings), in a general parameter setting. The results are strikingly simple and reveal explicitly how the expected delay depends on the system parameters, and in particular, on the batch-size distributions and the simultaneity of the batch arrivals. Moreover, the results suggest simple and fast-to-evaluate approximations for the expected delay in heavily loaded polling systems. Numerical experiments demonstrate that the approximations are highly accurate in medium and heavily loaded systems.

Key Words: Polling systems; Batch arrivals; Simultaneous arrivals; Delay; Heavy traffic.

1. INTRODUCTION

Polling systems are multi-queue systems in which a single server visits the queues in some order to serve the customers waiting at the queues, typically incurring some amount of switch-over time to proceed from one queue to the next. Polling models find a wide variety of applications in the fields of computer-communication systems, production, robotics, maintenance and manufacturing (cf. [17] for an

Copyright © 2001 by Marcel Dekker, Inc. www.dekker.com

ORDER REPRINTS

overview). Because of their wide applicability, polling models have received much attention in the literature since the early 1960s (cf. [22, 23] for overviews). In the vast majority of the papers on polling models, it is assumed that the arrival processes at the different queues are independent unit Poisson processes, i.e., where exactly one customer arrives at a particular queue at a time. However, in many applications customers arrive in batches, and batches of customers may arrive at different queues simultaneously. Neglecting such a correlation structure in the arrival processes may lead to strongly erroneous performance predictions and improper operation of the system. Therefore, in this paper we analyze the impact of the batch-size distributions and the simultaneity of batch arrivals at the different queues on the delay incurred at each of the queues.

Exact detailed results about the delay incurred at each of the queues are scarce, and hopes for exact results are often abandoned in favor of numerical techniques. However, the usefulness of numerical techniques is limited for several reasons. First, numerical techniques to some extent view the system as a "black box", and as such, can only provide limited insight into the dependence of the system performance with respect to the system parameters. Second, typical operation issues are "In which order should the queues be served?" and "What service disciplines should be assigned to the queues?", while the proper operation of the system is particularly critical when the system is heavily loaded. However, the efficiency of the available numerical techniques to predict the performance of the system may degrade dramatically when the system is heavily loaded. These observations raise the importance of an exact analysis of the system in heavy traffic.

The possibility of batched and simultaneous arrivals strongly enhances the modeling and analysis capabilities of polling models. Many examples are found in computer-communications systems. Consider for instance a Local Area Network (LAN), in which the right for transmission, represented by a so-called token, is circulated among the users. If a user wants to initiate a transaction over the network, a transaction request is placed into an output buffer. However, the amount of data that can be sent over the network at a time is limited. Accordingly, transmission requests exceeding the maximum transmission unit (MTU) size are fragmented. When the user gets the right for transmission, a number of (possibly all) fragments are transmitted over the network. In this application, the server represents the right for transmission (token) and the queues represent the output buffers. The customers represent the fragments, and as such, a transmission initiation request placed by a user is represented by a batch arrival of customers. Applications are also found in the area of flexible manufacturing systems. Consider for example a production facility that can produce different types of products, but only one type of product at a time. A large set of wholesalers from time to time place joint replenishment orders for different products. Incoming orders for a given product that can not be processed immediately are placed into a buffer of pending orders for that product. After a number of (possibly all) outstanding orders for a specific product have been processed, the production facility is installed to process the next type of products. In this way, the facility "visits" the buffers of pending orders for the different product

Copyright © Marcel Dekker, Inc. All rights reserved.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

ORDER REPRINTS

types in a round-robin fashion. In this example, the server represents the production facility, the customers represent replenishment orders of one unit of a product and joint replenishment orders are represented by simultaneous batch arrivals.

In the literature, polling systems with simultaneous arrivals have received only little attention. For models with gated or exhaustive service at each of the queues, Levy and Sidi [18] derive a set of linear equations for the expected delay at each of the queues. They also provide a pseudo-conservation law for the system, i.e., an exact expression for a specific weighted sum of the expected waiting times at the different queues. The moments of the delay in the model studied in [18] can also be obtained numerically by means of the Descendant Set Approach, an iterative technique based on the concept of descendant sets [13]. Boxma et al. [4] derive a pseudo-conservation law for discrete-time polling systems with independent batch arrivals and with mixtures of exhaustive, gated and 1-limited service. Chiarawongse and Srinivasan [5] derive a pseudo-conservation law for the continuous-time polling model with compound Poisson arrivals. Several papers have been devoted to polling systems under heavy load. For models with independent renewal arrival processes, Coffman et al. [7, 8], Reiman and Wein [21] and Markowitz [19] use the theory of diffusion processes to analyze the heavy-traffic behavior of polling models. For models with unit Poisson arrivals, Kudoh et al. [15] give explicit expressions for the second moment of the waiting time in fully symmetric systems with gated or exhaustive service at each queue for models with two, three and four queues. They also give conjectures for the heavy-traffic limits of the first two moments of the waiting times for systems with an arbitrary number of queues. For asymmetric polling systems with unit Poisson arrivals, Van der Mei [26, 27] derives closed-form expressions for the moments and the Laplace-Stieltjes Transform of the delay, under heavy traffic scalings. Kroese [14] studies continuous polling systems in heavy traffic with unit Poisson arrivals on a ring and shows that the steady-state number of customers at each queue has approximately a gamma-distribution.

In this paper, we consider a cyclic polling model with general mixtures of gated and exhaustive service and general service-time and switch-over time distributions, in heavy traffic. The correlation structure of the arrivals is modeled as follows. Arrival points are generated according to a Poisson process. At each arrival point, batches of customers may arrive *simultaneously* at the different at the queues, according to a general joint batch-size distribution. We focus on the expected delay at each of the queues when the load offered to the system, denoted by ρ, tends to unity. It is shown that the expected delay, considered as a function of ρ, has a first-order pole at $\rho = 1$ and hence, tends to infinity when the system tends to saturate. Denoting the delay at queue i by W_i, we consider the random variable $(1 - \rho)W_i$, referred to as the scaled delay. We focus on ω_i, defined as the limit of the expected scaled delay when ρ tends to unity, and derive closed-form expressions for ω_i, in a general parameter setting. The expressions explicitly reveal how the expected delay depends on the system parameters, and in particular, how the correlation structure in the arrival process impacts the expected delay incurred at each of the queues. In addition, the results reveal a variety of insensitivity and monotonicity

Copyright © Marcel Dekker, Inc. All rights reserved.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

ORDER REPRINTS

properties of the scaled expected delay with respect to specific system parameters. These observations give insights into the heavy-traffic behavior of polling systems that have not been observed before. Finally, the results suggest simple and fast-to-evaluate approximations for the expected delay in stable polling systems. Numerical results demonstrate that the accuracy of the approximations is very good for medium and heavily loaded systems.

This paper extends the results in [24], where we considered the expected delay in polling systems with independent unit Poisson arrivals, based on the use of the Descendant Set Approach (DSA). The DSA has been found to be very useful in analyzing the heavy-traffic behavior of polling systems. For this reason, the derivation of the results in the present paper proceeds along similar lines. The extension to the case of simultaneous batch arrivals is fairly straightforward. In this context, we emphasize that the added value of the present paper is the variety of new insights obtained into the impact of correlated batch arrivals on the delay and the proposal of simple approximations for the expected delay, rather than its technical contribution.

The remainder of this paper is organized as follows. In section 2 the model is described. In section 3 we discuss the use of the DSA for the present model, use the DSA to analyze the heavy-traffic behavior of the system and derive closed-form expressions for the scaled expected delay in the limiting case. In section 4 we discuss asymptotic properties of the behavior of the system in heavy traffic. In section 5 we propose and test simple approximations for the expected delay at each of the queues. Finally, in section 6 we address a number of topics for further research.

2. MODEL

Consider a system consisting of $N \geq 2$ infinite-buffer stations $Q_1, \ldots, Q_N$ served by a single server that visits and serves the queues in cyclic order. The correlated batch arrival process is modeled as follows. *Arrival points* are generated according a Poisson process with rate λ. At each arrival point, batches of customer may arrive at the different queues according to some probability distribution. More precisely, let $\underline{K} = (K_1, \ldots, K_N)$ be a random vector, where K_i stands for the number of customers arriving to Q_i at an arrival point. Customers at Q_i are referred as type-i customers. The random vector $\underline{K}$ is assumed to be independent of previous or future arrival points. Denote the joint batch-size distribution by $\pi(k_1, \ldots, k_N) := \text{Prob}\{K_1 = k_1, \ldots, K_N = k_N\}(k_i = 0, 1, \ldots, \text{for } i = 1, \ldots, N)$, and denote the corresponding Probability Generating Function (PGF) of $\underline{K}$ by $K^*(\underline{z})$. Denote the arrival rate at Q_i by $\lambda_i := \lambda E[K_i]$, and let $K_{i,i} := E[K_i(K_i - 1)]$ for $i = 1, \ldots, N$ and $K_{i,j} := E[K_i K_j]$ for $i \neq j$. Denote the total arrival rate by $\Lambda := \sum_{i=1}^{N} \lambda_i$. The service time of a customer at Q_i is a random variable B_i with Laplace-Stieltjes Transform (LST) $B_i^*(\cdot)$, and (finite) k-th moment $b_i^{(k)}$, $k = 1, 2, \ldots$. The load offered to Q_i is defined by $\rho_i = \lambda_i b_i^{(1)}$, and the total offered load is equal to $\rho = \sum_{i=1}^{N} \rho_i$.

Copyright © Marcel Dekker, Inc. All rights reserved.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

ORDER REPRINTS

Denote the k-th moment of the service time of an arbitrary customer by $b^{(k)} := (1/\Lambda)\sum_{i=1}^{N}\lambda_i b_i^{(k)}$, $k = 1, 2$. Polling instants are defined as the epochs at which the server arrives at a queue to serve customers waiting at that queue. We consider two types of service disciplines: gated and exhaustive. Under the gated policy only the customers that were present at the polling instant at Q_i are served; customers that arrive at Q_i while it is being served are served during the next visit to Q_i. Under the exhaustive policy the server visits Q_i until it is empty. We allow general mixtures of exhaustive and gated service, but the service policy at each queue remains the same for all visits. Define $E := \{i : Q_i \text{ is served exhaustively}\}$ and $G := \{i : Q_i \text{ is served according to the gated policy}\}$. At each queue the customers are served in the order of arrival. After completing service at Q_i the server immediately proceeds to Q_{i+1}, incurring a switch-over period whose duration is an independent random variable R_i with LST $R_i^*(\cdot)$, mean r_i and (finite) second moment $r_i^{(2)}$. Denote the first two moments of the total switch-over time per cycle by $r := \sum_{i=1}^{N} r_i$ and $r^{(2)} := \sum_{i=1}^{N} r_i^{(2)} + \sum_{i\neq j} r_i r_j$, respectively. All interarrival times and service times are assumed to be mutually independent and independent of the state of the system. It is assumed that the system is stable (i.e., $\rho < 1$) [11] and that the system is in steady state.

Let W_i be the delay incurred by an arbitrary customer at Q_i. We are interested in the behavior of $E[W_i]$, the expected delay at Q_i, when $\rho \uparrow 1$. Throughout, $E[W_i]$ will be considered as a function of ρ, where λ (i.e., the rate at which the arrival points occur) is variable, while the service-time and batch-size distributions remain fixed. It is shown that $E[W_i]$, considered as a function of ρ, has a first-order pole at $\rho = 1$ (see Remark 3.3). Therefore, we consider the expected value of the random variable $(1-\rho)W_i$, referred to as the *scaled* delay at Q_i, when the load tends to unity. Thus, the analysis will be oriented towards the determination of closed-form expressions for the quantities

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

The following notation is useful. For an event E, denote by I_E the indicator function on E. Denote by $\underline{e}_i$ the i-th unit vector, $i = 1, \ldots, N$. Moreover, for each variable x that is a function of ρ, we use the hat-notation $\hat{x}$ to indicate its value at $\rho = 1$.

3. ANALYSIS

In this section we analyze the queueing behavior of the model by means of the Descendant Set Approach (DSA). First, we discuss the use of the DSA for the present model. Then we use the DSA to analyze the heavy-traffic behavior of the system.

Without loss of generality, we focus on $E[W_1]$, the expected delay incurred at Q_1. Denote by X_1 the number of customers at Q_1 at an arbitrary polling instant at

Copyright © Marcel Dekker, Inc. All rights reserved.

ORDER REPRINTS

Q_1, and denote its correspond PGF by $X_1^*(\cdot)$. For the present model, the expected delay at Q_1 at an arbitrary moment can be expressed in terms of the first two moments of X_1 as follows (cf. [18]):

$$E[W_1] = \frac{E\left[X_1^2\right] - E[X_1]}{2\lambda_1 E[X_1]}(1+\rho_1) - \frac{K_{1,1}}{2\lambda_1 E[K_1]} \quad (1 \in G), \tag{2}$$

$$E[W_1] = \frac{E\left[X_1^2\right] - E[X_1]}{2\lambda_1 E[X_1]} + \frac{\lambda_1 b_1^{(2)}}{2(1-\rho_1)} + \frac{(2\rho_1 - 1)K_{1,1}}{2\lambda_1(1-\rho_1)E[K_1]} \quad (1 \in E). \tag{3}$$

Hence, to obtain $E[W_1]$, it suffices to obtain expressions for $E[X_1]$ and $E[X_1^2]$. Simple balancing arguments immediately lead to the following expressions for $E[X_1]$:

$$E[X_1] = \frac{\lambda_1 r}{1-\rho} \quad (1 \in G), \quad E[X_1] = \frac{\lambda_1(1-\rho_1)r}{1-\rho} \quad (1 \in E). \tag{4}$$

The derivation of expressions for $E[X_1^2]$ is more involved and requires the use of the DSA, which is outlined below (cf. [17] for more details).

3.1. Descendant Set Approach

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 the distribution of X_1, the number of customers at Q_1 present at an arbitrary polling instant P^* at Q_1, referred to as the reference point. Define a cycle as the elapsed time between two successive polling instants at Q_1. The key observation is that we can evaluate $X_1(P^*)$ by considering, recursively, contributions to X_1 from waiting customers at all queues of the past polling epochs, working backward from the reference point. Let $T_{i,c}$ be a customer served at Q_i during the c-th cycle. Define the children set of $T_{i,c}$ to be the set of customers arriving during the service of $T_{i,c}$; the descendant set of $T_{i,c}$ is recursively defined to consist of $T_{i,c}$, its children and the descendants of its children. Let $A_{i,c}$ be the number of type-1 customers at the reference point (at Q_1) that are descendants of $T_{i,c}$, and let $A_{i,c}^*(\cdot)$ denote its PGF. In this way, $A_{i,c}$ can be viewed as the contribution of $T_{i,c}$ to $X_1(P^*)$. Denote by $\alpha_{i,c}^{(k)}$, $k = 1, 2$, the first two factorial moments of $A_{i,c}$. Denote by $R_{i,c}$ the switch-over time (from Q_i to the next) immediately after the visit starting at $P_{i,c}$. Let $S_{i,c}$ be the total contribution to $X_1(P^*)$ of all (original) customers that arrive in the system during $R_{i,c}$, and denote the corresponding PGF by $S_{i,c}^*(\cdot)$. Then X_1 can be expressed as the independent sum $X_1 = \sum_{i=1}^{N}\sum_{c=0}^{\infty} S_{i,c}$, or equivalently, for $|z| \le 1$,

$$X_1^*(z) = \prod_{i=1}^{N}\prod_{c=0}^{\infty} S_{i,c}^*(z), \tag{5}$$

Copyright © Marcel Dekker, Inc. All rights reserved.

ORDER REPRINTS

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

$$S^*_{i,c}(z) = R^*_i(\lambda - \lambda K^*(A^*_{1,c-1}(z), \ldots, A^*_{i,c-1}(z), A^*_{i+1,c}(z), \ldots, A^*_{N,c}(z))). \tag{6}$$

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

$$A^*_{i,c}(z) = B^*_i(\lambda - \lambda K^*(A^*_{1,c-1}(z), \ldots, A^*_{i,c-1}(z), A^*_{i+1,c}(z), \ldots, A^*_{N,c}(z))), \tag{7}$$

and for $i \in E$, $c = 0, 1, \ldots$, $|z| \leq 1$,

$$A^*_{i,c}(z) = B^*_i(\lambda - \lambda K^*(A^*_{1,c-1}(z), \ldots, A^*_{i-1,c-1}(z), A^*_{i,c}(z), \ldots, A^*_{N,c}(z))). \tag{8}$$

Since we focus on the number of type-1 customers at the reference point, the initial conditions are $A^*_{1,-1}(z) = z$ and $A^*_{i,-1}(\cdot) = 1$, for $i = 2, \ldots, N$.

Differentiating (5) and (6) with respect to z and substituting $z = 1$ leads to the following expressions for the first two moments of X_1 in terms of the DS variables:

$$E[X_1] = \sum_{i=1}^{N} \sum_{c=0}^{\infty} r_i \left[\sum_{j=i+1}^{N} \lambda_j \alpha^{(1)}_{j,c} + \sum_{j=1}^{i} \lambda_j \alpha^{(1)}_{j,c-1} \right], \tag{9}$$

and

$$\begin{aligned} Var[X_1] = {} & \sum_{i=1}^{N} \sum_{c=0}^{\infty} \left(r^{(2)}_i - r^2_i\right) \left[\sum_{j=i+1}^{N} \lambda_j \alpha^{(1)}_{j,c} + \sum_{j=1}^{i} \lambda_j \alpha^{(1)}_{j,c-1} \right]^2 \\ & + \sum_{i=1}^{N} \sum_{c=0}^{\infty} r_i \left[\sum_{j=i+1}^{N} \lambda_j \alpha^{(2)}_{j,c} + \sum_{j=1}^{i} \lambda_j \alpha^{(2)}_{j,c-1} \right] \\ & + \lambda \sum_{i=1}^{N} \sum_{c=0}^{\infty} r_i \delta_{i,c}, \end{aligned} \tag{10}$$

where

$$\begin{aligned} \delta_{i,c} := {} & \sum_{j=i+1}^{N} \left[\sum_{k=i+1}^{N} K_{j,k} \alpha^{(1)}_{j,c} \alpha^{(1)}_{k,c} + \sum_{k=1}^{i} K_{j,k} \alpha^{(1)}_{j,c} \alpha^{(1)}_{k,c-1} \right] \\ & + \sum_{j=1}^{i} \left[\sum_{k=i+1}^{N} K_{j,k} \alpha^{(1)}_{j,c-1} \alpha^{(1)}_{k,c} + \sum_{k=1}^{i} K_{j,k} \alpha^{(1)}_{j,c-1} \alpha^{(1)}_{k,c-1} \right]. \end{aligned} \tag{11}$$

Similarly, from (7) and (8) we obtain the following recursive relations for the first two factorial moments of $A_{i,c}$: For $i \in G$, $c = 0, 1, \ldots$,

$$\alpha^{(1)}_{i,c} = b^{(1)}_i \left[\sum_{j=i+1}^{N} \lambda_j \alpha^{(1)}_{j,c} + \sum_{j=1}^{i} \lambda_j \alpha^{(1)}_{j,c-1} \right], \tag{12}$$

Copyright © Marcel Dekker, Inc. All rights reserved.

ORDER REPRINTS

and

$$\alpha_{i,c}^{(2)} = b_i^{(1)} \left[\sum_{j=i+1}^{N} \lambda_j \alpha_{j,c}^{(2)} + \sum_{j=1}^{i} \lambda_j \alpha_{j,c-1}^{(2)} \right] + b_i^{(2)} \left[\sum_{j=i+1}^{N} \lambda_j \alpha_{j,c}^{(1)} + \sum_{j=1}^{i} \lambda_j \alpha_{j,c-1}^{(1)} \right]^2 + b_i^{(1)} \lambda \delta_{i,c}, \tag{13}$$

where $\delta_{i,c}$ is defined in (11). Similar relations can be obtained for the case Q_i is served exhaustively: For $i \in E$, $c = 0, 1, \ldots$,

$$\alpha_{i,c}^{(1)} = \frac{b_i^{(1)}}{1 - \rho_i} \left[\sum_{j=i+1}^{N} \lambda_j \alpha_{j,c}^{(1)} + \sum_{j=1}^{i-1} \lambda_j \alpha_{j,c-1}^{(1)} \right], \tag{14}$$

and

$$\alpha_{i,c}^{(2)} = \frac{1}{1 - \rho_i} \left\{ b_i^{(1)} \left[\sum_{j=i+1}^{N} \lambda_j \alpha_{j,c}^{(2)} + \sum_{j=1}^{i-1} \lambda_j \alpha_{j,c-1}^{(2)} \right] \right. \tag{15}$$

$$\left. + b_i^{(2)} \left[\sum_{j=i}^{N} \lambda_j \alpha_{j,c}^{(1)} + \sum_{j=1}^{i-1} \lambda_j \alpha_{j,c-1}^{(1)} \right]^2 + b_i^{(1)} \lambda \tilde{\delta}_{i,c} \right\}, \tag{16}$$

where $\tilde{\delta}_{i,c}$ is the same as $\delta_{i,c}$, except that the summations $\sum_{j=1}^{i}$ and $\sum_{j=i+1}^{N}$ in (11) are replaced by $\sum_{j=1}^{i-1}$ and $\sum_{j=i}^{N}$, respectively (and similarly for the variable k). The initial conditions are $\alpha_{1,-1}^{(k)} = 1$, $\alpha_{i,-1}^{(k)} = 0$ for $i = 2, \ldots, N$, $k = 1, 2$.

Relations (2)-(16) give a complete, but not explicit, characterization of the expected delay at each of the queues. These can relations can be used for the numerical computation of the first two moments of X_1. In the next subsection it will be shown how the DSA can also be used to obtain exact expressions for the (scaled) moments of X_1 under heavy traffic assumptions.

3.2. Heavy Traffic Behavior

In this section we analyze the heavy-traffic behavior of the sequences $\{\alpha_{i,c}^{(k)}, c = 0, 1, \ldots\}$, $k = 1, 2$. The obtained properties will then be used to obtain expressions for the $E[X_1^2]$ and finally for $E[W_1]$. Let us first consider the case $k = 1$. It is readily verified that the recursive relations for $k = 1$ (see (12) and (14)) are identical to the case of independent unit Poisson arrivals, which are analyzed in detail in [24, 26]. The results shown in those papers are outlined below.
Define

$$\gamma := \frac{1}{2} \left[1 - \sum_{i \in E} \hat{\rho}_i^2 + \sum_{i \in G} \hat{\rho}_i^2 \right]. \tag{17}$$

Copyright © Marcel Dekker, Inc. All rights reserved.

ORDER REPRINTS

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

$$\alpha_{i,c}^{(1)} = \xi^{c+1} v_i w + s_{i,c}, \tag{18}$$

where

(1) $\xi < 1$ *if and only if* $\rho < 1; \xi = 1$ *if and only if* $\rho = 1$,
(2) $\lim_{\rho \uparrow 1} \xi = 1$,
(3) $\hat{v}_i := b_i^{(1)}/\gamma \, (i = 1, \ldots, N)$,
(4) $\hat{w} := \lambda_1(1 - \hat{\rho}_1 I_{\{1 \in E\}})$,
(5) $\lim_{\rho \uparrow 1} \frac{1-\xi}{1-\rho} = \gamma^{-1}$,
(6) $|s_{i,c}| < C \xi_*^c$ *for some* $C \, (0 < C < \infty)$ *and* $\xi \, (0 < \xi_* < \xi)$.

Proof: See [24].

Lemma 1 decomposes the variables $\alpha_{i,c}^{(1)}$ into two parts: a dominant and a recessive part. The dominant part (i.e., $\xi^{c+1} v_i w$) dominates the recessive part (i.e., $s_{i,c}$) in the sense that

$$\lim_{\rho \uparrow 1} (1 - \rho) \sum_{i=1}^{N} \sum_{c=0}^{\infty} \alpha_{i,c}^{(1)} = \lim_{\rho \uparrow 1} (1 - \rho) \sum_{i=1}^{N} \sum_{c=0}^{\infty} \xi^{c+1} v_i w \tag{19}$$

$$+ \lim_{\rho \uparrow 1} (1 - \rho) \sum_{i=1}^{N} \sum_{c=0}^{\infty} s_{i,c} = \hat{\lambda}_1 (1 - \hat{\rho}_1 I_{\{1 \in E\}}) + 0, \tag{20}$$

where the first equality follows from (18) and the second follows from the properties stated in Lemma 1. In this way, the impact of the recessive part becomes negligible when the system tends to saturate. This concept of dominance plays an important role in the analysis (see Remark 4.2 for more details).

We now proceed to analyze the sequences $\{\alpha_{i,c}^{(k)}, c = 0, 1, \ldots\}$ for $k = 2$. The following result is crucial to the analysis.

Theorem 1.

$$\lim_{\rho \uparrow 1} (1 - \rho)^2 \sum_{i=1}^{N} \sum_{c=0}^{\infty} \lambda_i \alpha_{i,c}^{(2)} \tag{21}$$

$$= \frac{\hat{\lambda}_1^2 (1 - \hat{\rho}_1 I_{\{1 \in E\}})^2}{2\gamma} \left[\frac{b^{(2)}}{b^{(1)}} + \hat{\lambda} \sum_{i=1}^{N} \sum_{j=1}^{N} b_i^{(1)} b_j^{(1)} K_{i,j} \right]. \tag{22}$$

Copyright © Marcel Dekker, Inc. All rights reserved.

ORDER REPRINTS

Proof: Define $A^{(2)} := \sum_{i=1}^{N} \sum_{c=0}^{\infty} \lambda_i \alpha_{i,c}^{(2)}$. Then for $1 \in G$, relation (13) can be easily shown to imply that

$$A^{(2)} = \frac{1}{1-\rho} \left\{ \sum_{i=1}^{N} \sum_{c=0}^{\infty} \lambda_i b_i^{(2)} \left[\sum_{j=i+1}^{N} \lambda_j \alpha_{j,c}^{(1)} + \sum_{j=1}^{i} \lambda_j \alpha_{j,c-1}^{(1)} \right]^2 \right. \tag{23}$$

$$\left. + \lambda \sum_{i=1}^{N} \sum_{c=0}^{\infty} \lambda_i b_i^{(1)} \delta_{i,c} + \lambda_1 \rho \right\}. \tag{24}$$

Then it can be shown that

$$\lim_{\rho \uparrow 1} (1-\rho) \sum_{i=1}^{N} \sum_{c=0}^{\infty} \lambda_i b_i^{(2)} \left[\sum_{j=i+1}^{N} \lambda_j \alpha_{j,c}^{(1)} + \sum_{j=1}^{i} \lambda_j \alpha_{j,c-1}^{(1)} \right]^2 \tag{25}$$

$$= \frac{\hat{\lambda}_1^2}{2\gamma} \sum_{i=1}^{N} \hat{\lambda}_i b_i^{(2)} = \frac{\hat{\lambda}_1^2}{2\gamma} \hat{\Lambda} b^{(2)} = \frac{\hat{\lambda}_1^2}{2\gamma} \frac{b^{(2)}}{b^{(1)}}. \tag{26}$$

The first equality in (25)-(26) follows from (18), Lemma 1 and several straightforward manipulations. The second equality follows from the definition of $b^{(2)}$ and the last equality follows from the fact that $\hat{\Lambda} b^{(1)} = \hat{\rho} = 1$. Moreover, we have

$$\lim_{\rho \uparrow 1} (1-\rho) \sum_{i=1}^{N} \sum_{c=0}^{\infty} \lambda_i b_i^{(1)} \delta_{i,c} = \frac{\hat{\lambda}_1^2}{2\gamma} \sum_{i=1}^{N} \sum_{j=1}^{N} b_i^{(1)} b_j^{(1)} K_{i,j}, \tag{27}$$

which follows from (11), Lemma 1 and several straightforward manipulations. Combining (23)-(27) with the fact that $\lim_{\rho \uparrow 1} (1-\rho)\lambda_1 \rho = 0$ leads to the result. Similar arguments lead to the results for the case $1 \in E$. □

We will now use Theorem 1 to obtain expressions for ω_i $(i = 1, \ldots, N)$, defined in (1). The variables $E[X_1]$ and $E[X_1^2]$, considered as a function of ρ, possess a first-order pole and a second-order pole at $\rho = 1$, respectively (see Remark 3.3 below). Therefore, we define the corresponding "heavy-traffic residues" as follows:

$$x_1^{(1)} := \lim_{\rho \uparrow 1} (1-\rho) E[X_1], \quad x_1^{(2)} := \lim_{\rho \uparrow 1} (1-\rho)^2 E\left[X_1^2\right]. \tag{28}$$

The following result gives expressions for $x_1^{(1)}$ and $x_1^{(2)}$ in closed form.

Theorem 2.

$$(1) \quad x_1^{(1)} = \hat{\lambda}_1 \left(1 - \hat{\rho}_1 I_{\{1 \in E\}}\right) r, \tag{29}$$

$$(2) \quad x_1^{(2)} = \frac{r\hat{\lambda}_1^2}{2\gamma} \frac{b^{(2)}}{b^{(1)}} + r^2 \hat{\lambda}_1^2 + \frac{r\hat{\lambda}_1^2}{2\gamma} \hat{\lambda} \sum_{i=1}^{N} \sum_{j=1}^{N} b_i^{(1)} b_j^{(1)} K_{i,j} \quad (1 \in G), \tag{30}$$

Copyright © Marcel Dekker, Inc. All rights reserved.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

ORDER REPRINTS

$$(3)\quad x_1^{(2)} = \frac{r\hat{\lambda}_1^2(1-\hat{\rho}_1)^2}{2\gamma}\frac{b^{(2)}}{b^{(1)}} + r^2\hat{\lambda}_1^2(1-\hat{\rho}_1)^2 \tag{31}$$

$$+ \frac{r\hat{\lambda}_1^2(1-\hat{\rho}_1)^2}{2\gamma}\hat{\lambda}\sum_{i=1}^{N}\sum_{j=1}^{N} b_i^{(1)}b_j^{(1)}K_{i,j} \quad (1 \in E). \tag{32}$$

Proof: Part 1 follows directly from (4). Parts 2 and 3 follow from Theorem 1 and several straightforward manipulations. □

The following relations express ω_1 in terms of the variables $x_1^{(k)}$, $k = 1, 2$.

Theorem 3.

$$\omega_1 = \frac{x_1^{(2)}}{2\hat{\lambda}_1 x_1^{(1)}}(1+\hat{\rho}_1)\,(1 \in G), \quad \omega_1 = \frac{x_1^{(2)}}{2\hat{\lambda}_1 x_1^{(1)}} \quad (1 \in E). \tag{33}$$

Proof: Follows directly from (2) and (3), the assumption that $N \geq 2$ (see also Remark 3.2 below) and several straightforward manipulations. □

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

Theorem 4 (Main Result). *For $i \in G$,*

$$\omega_i = \frac{1+\hat{\rho}_i}{1-\sum_{m\in E}\hat{\rho}_m^2+\sum_{m\in G}\hat{\rho}_m^2}\frac{b^{(2)}}{2b^{(1)}} + \frac{1}{2}r(1+\hat{\rho}_i) \tag{34}$$

$$+ \frac{\hat{\lambda}(1+\hat{\rho}_i)\sum_{j=1}^{N}\sum_{k=1}^{N} b_j^{(1)}b_k^{(1)}K_{j,k}}{2(1-\sum_{m\in E}\hat{\rho}_m^2+\sum_{m\in G}\hat{\rho}_m^2)}. \tag{35}$$

For $i \in E$,

$$\omega_i = \frac{1-\hat{\rho}_i}{1-\sum_{m\in E}\hat{\rho}_m^2+\sum_{m\in G}\hat{\rho}_m^2}\frac{b^{(2)}}{2b^{(1)}} + \frac{1}{2}r(1-\hat{\rho}_i) \tag{36}$$

$$+ \frac{\hat{\lambda}(1-\hat{\rho}_i)\sum_{j=1}^{N}\sum_{k=1}^{N} b_j^{(1)}b_k^{(1)}K_{j,k}}{2(1-\sum_{m\in E}\hat{\rho}_m^2+\sum_{m\in G}\hat{\rho}_m^2)}. \tag{37}$$

Proof: Follows directly from Theorems 2 and 3, and the fact that we assumed $i = 1$ without loss of generality. □

Remark 3.1. Theorem 4 is supported by known results in the literature. For the special case of independent unit Poisson arrivals, the results generalize the heavy-traffic results obtained in [24]. For the case $N = 2$, $G = \emptyset$ and independent unit arrivals the results correspond to those obtained by Coffman et al. [7, 8], who also derive the results for non-Poisson renewal-type arrival processes. For systems with joint batch arrivals, the only available explicit results are the formulations

Copyright © Marcel Dekker, Inc. All rights reserved.

ORDER REPRINTS

of the pseudo-conservation laws (cf. [18]), giving a closed-form expression for $\sum_{i=1}^{N} \rho_i E[W_i]$. The expressions obtained in the present paper are in line with the results in [18].

Remark 3.2. In the derivation of the results (in particular, Theorem 3) it is assumed that $N \geq 2$. For the case $N = 1$, it follows directly from the pseudo-conservation law in [18] that

$$E[W_1] = \frac{\rho}{(1-\rho)} \frac{b_1^{(2)}}{2b_1^{(1)}} + \frac{r^{(2)}}{2r} + \frac{r\rho}{2(1-\rho)} I_{\{1\in G\}} + \frac{b_1^{(1)}}{(1-\rho)} \frac{K_{1,1}}{2E[K_1]}. \tag{38}$$

Then, by premultiplying by a factor $(1-\rho)$ and taking the limit for $\rho \uparrow 1$, it is readily seen that

$$\omega_1 = \frac{b_1^{(2)}}{2b_1^{(1)}} + \frac{r}{2} I_{\{1\in G\}} + \frac{b_1^{(1)} K_{1,1}}{2E[K_1]}. \tag{39}$$

Remark 3.3. In equation (1) it is assumed that $E[W_1]$, considered as a function of ρ, has a first-order pole at $\rho = 1$. It is interesting to note that the results presented above actually *prove* that this is indeed the case. To this end, note first that equation (4) directly implies that $E[X_1]$ has a first-order pole at $\rho = 1$. Moreover, combining Lemma 1, Theorem 1 and relations (10)-(11) it can be shown that $Var[X_1]$, and hence also $E[X_1^2]$, possesses a second-order pole at $\rho = 1$. Equations (2) and (3) then immediately imply that $E[W_1]$ has a first-order pole at $\rho = 1$.

4. ASYMPTOTIC PROPERTIES

The results derived in the previous section reveal a number of interesting observations, which are discussed below. First, Theorem 4 suggests the following decomposition: For $i = 1, \ldots, N$,

$$\omega_i = I_i + II_i + III_i + IV_i, \tag{40}$$

where

$$I_i := \frac{\hat{\eta}_i}{2(1 - \sum_{m\in E} \hat{\rho}_m^2 + \sum_{m\in G} \hat{\rho}_m^2)} \frac{b^{(2)}}{b^{(1)}}, \tag{41}$$

$$II_i := \frac{\hat{\eta}_i r}{2}, \tag{42}$$

$$III_i := \frac{\hat{\lambda}\hat{\eta}_i \sum_{j=1}^{N} b_j^{(1)^2} K_{j,j}}{2(1 - \sum_{m\in E} \hat{\rho}_m^2 + \sum_{m\in G} \hat{\rho}_m^2)}, \tag{43}$$

$$IV_i := \frac{\hat{\lambda}\hat{\eta}_i \sum_{j=1}^{N} \sum_{k\neq j} b_j^{(1)} b_k^{(1)} K_{j,k}}{2(1 - \sum_{m\in E} \hat{\rho}_m^2 + \sum_{m\in G} \hat{\rho}_m^2)}, \tag{44}$$

Copyright © Marcel Dekker, Inc. All rights reserved.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

ORDER REPRINTS

with $\eta_i := 1 + \rho_i$ for $i \in G$ and $\eta_i := 1 - \rho_i$ for $i \in E$. We make the following observations: (1) I_i is independent of the switch-over times, the batch-size distributions and the simultaneity of the arrivals, (2) II_i depends on the switch-over times, but is independent of the batch-size distributions and the simultaneity of the arrivals of customers at the different queues, (3) III_i depends on the marginal batch-size distributions at the different queues, but is independent of the switch-over times and the simultaneity of the arrivals, and (4) IV_i depends on the simultaneity of the arrivals, but is independent of the switch-over times and the marginal batch-size distributions. These observations leads to the following result.

Property 1 (Decomposition). *For $i = 1, \ldots, N$, we can write $\omega_i = I_i + II_i + III_i + IV_i$, defined in* (41)-(44), *where*

(1) *I_i is the scaled expected delay for the model with zero switch-over times and independent unit Poisson arrivals,*
(2) *II_i quantifies the impact of the switch-over times,*
(3) *III_i quantifies the impact of the marginal batch-size distributions,*
(4) *IV_i quantifies the impact of the simultaneity of the arrival epochs at the different queues.*

Theorem 4 also reveals a number of insensitivity properties of the expected (scaled) delay with respect to the system parameters.

Property 2 (Insensitivity). *For $i = 1, \ldots, N$,*

(1) *ω_i is independent of the visit order,*
(2) *ω_i depends on the second-order moments of the joint batch-size distribution $K_{j,k}$ $(j, k = 1, \ldots, N)$ only through the weighted sum $\sum_{j=1}^{N} \sum_{k=1}^{N} b_j^{(1)} b_k^{(1)} K_{j,k}$,*
(3) *ω_i depends on the second moments of the service-time distributions only through $b^{(2)}$,i.e., the second moment of the service time of an arbitrary customer,*
(4) *ω_i depends on the switch-over time distributions only through r, i.e.,the total expected switch-over time per cycle.*

In addition, Theorem 4 immediately leads to the following monotonicity properties.

Property 3 (Monotonicity). *For $i = 1, \ldots, N$, ω_i is strictly increasing in $b^{(2)}$, r and $K_{j,k}$ $(j, k = 1, \ldots, N)$.*

Remark 4.1. Although the monotonicity results in Property 3 are quite intuitive, rigorous proofs of similar monotonicity properties are scarce in the polling literature. Altman et al. [1] show that the waiting times, cycle times and intervisit times are stochastically increasing in the arrival rates, service times and switch-over times. Levy and Sidi [18] compare the expected delay in a similar model for stable polling systems (i.e., for $\rho < 1$). Based on extensive numerical experimentation,

Copyright © Marcel Dekker, Inc. All rights reserved.

ORDER REPRINTS

they *conjecture* that the expected delay in the simultaneous batched Poisson case is always larger than in the base model. In this context, our results actually *prove* that this conjecture is (asymptotically) correct under heavy-traffic assumptions. In addition to these quite intuitive monotonicity results, several counter-intuitive monotonicity results have been obtained. For example, Cooper et al. [9] consider models in which the server incurs a switch-over time to move to a queue and an additional setup time in case the queue is not empty. The results in [9] show that the addition of (state dependent) switch-over times does not necessarily lead to an increase of the expected delay in stable systems (i.e., with $\rho < 1$). Similar counter-intuitive results have been obtained by Altman and Yechiali [2], who consider a model in which a state-independent Bernoulli scheme is used to decide whether the server actually enters a queue to render service at thatqueue. If so, the server incurs a setup time before starting to serve customers (if any); otherwise, the server immediately proceeds to the next queue. The results in [2] show that the Bernoulli scheme that minimizes the expected amount of waiting work in the system allows for the possibility that queues are skipped with a non-zero probability. The results in the present paper suggest that the possibility of decreasing the expected delay by increasing the switch-over times vanishes when the system tends to saturate.

Remark 4.2. The insensitivity properties listed in Property 2 are known to be not generally valid for stable polling systems, where the expected delay generally *does* depend on the visit order, the individual service-time and switch-over time distributions and *all* second moments of the joint batch-size distribution. In this perspective, the impact of these parameters can be viewed as lower-order effects in heavy traffic. For stable systems, explicit expressions for the expected delay can be obtained (for a given number of queues) by algebraically solving sets of linear equations (cf. [17]). However, these expressions are very cumbersome, even for systems with a small number of queues. In this paper it is shown that the (scaled) heavy-traffic limits of these expressions are remarkably simple. The underlying reason for this observation is a concept of *dominance*, which plays a role when the load tends to unity. The basic observation is that the sequences $\{\alpha_{i,c}^{(k)}, c = 0, 1, \ldots\}$, $k = 1, 2$, can be decomposed into a dominant and a recessive part, in the sense that the summations $\lim_{\rho\uparrow 1} (1-\rho)^k \sum_{i=1}^{N} \sum_{c=0}^{\infty} \alpha_{i,c}^{(k)}$, and hence also $x_1^{(k)}$ $(k = 1, 2)$ and ω_1, are completely determined by the dominant part (for $k = 1$, see also equation (19)). The impact of the recessive part becomes negligible in the limiting case, while the dominant part is relatively easy to analyze (see Lemma 1 and Theorem 1). Hence, the impact of parameters that only occur in the recessive part (such as the visit order, the higher moments of the service-time distributions and the switch-over times) vanish when the system tends to saturate. These observations explain the remarkable simplicity of the results and the observed insensitivity properties.

As an illustration of the dominance effect, consider for example the fully symmetric case, with all queues served exhaustively. In that case we have $b_i^{(k)} = b_1^{(k)}$, $\rho_i = \rho/N$ for $i = 1, \ldots, N, k = 1, 2, K_{i,i} = K_{1,1}$, $i = 1, \ldots, N$, and $K_{i,j} = K_{1,2}$,

Copyright © Marcel Dekker, Inc. All rights reserved.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

ORDER REPRINTS

$i, j = 1, \ldots, N, i \neq j$. Then from the results in [18] it may be shown that,

$$E[W_1] = \frac{\rho}{(1-\rho)} \frac{b_1^{(2)}}{2b_1^{(1)}} + \frac{r^{(2)}}{2r^{(1)}} + \frac{r\rho}{2(1-\rho)} \left(1 - \frac{1}{N}\right) \tag{45}$$

$$+ \frac{\lambda \left(b_1^{(1)}\right)^2 (NK_{1,1} + N(N-1)K_{1,2})}{2(1-\rho)}. \tag{46}$$

Premultiplication by $(1-\rho)$ and taking the limit for $\rho \uparrow 1$ it is ready seen that

$$\omega_1 = \frac{b_1^{(2)}}{2b_1^{(1)}} + \frac{r(N-1)}{2N} + \frac{1}{2}\hat{\lambda}\left(b_1^{(1)}\right)^2 \left(NK_{1,1} + N(N-1)K_{1,2}\right), \tag{47}$$

which is readily seen to be identical to the results in Theorem 4. The validity of the results for the case of gated service can be verified similarly. In the context of the above-mentioned dominance in heavy traffic, relations (45)-(47) illustrate, for instance, that the impact of the second moments of the switch-over times (in the symmetric case represented by $r^{(2)}$ only) vanishes when the load tends to unity, and as such can be considered of lower order in heavy traffic.

Remark 4.3. The expected delay at a queue can be related to the first two moments of the cycle times. Defining a cycle to be the time interval between two successive polling instants at Q_i, it is well-known that $E[C_i] = r/(1-\rho)$ for all i, and for the case of gated service at Q_i we have $E[W_i] = (1+\rho_i)E[RC_i]$, where $E[RC_i] = E[C_i^2]/2E[C_i]$, and where RC_i is the residual cycle time at Q_i. Similarly, defining a cycle time as the time between two successive departure instant from Q_i, for the case of exhaustive service at Q_i we have $E[W_i] = (1-\rho_i)E[RC_i]$. In this context, Theorem 4 indicates that, for $i = 1, \ldots, N$,

$$\lim_{\rho \uparrow 1} (1-\rho)^2 Var[C_i] = \frac{r}{2\gamma} \left[\frac{b^{(2)}}{b^{(1)}} + \hat{\lambda} \sum_{j=1}^{N} \sum_{k=1}^{M} b_j^{(1)} b_k^{(1)} K_{j,k} \right], \tag{48}$$

independent of i. It is known that in the variances of the cycle times for the different queues are not generally identical for stable polling systems (i.e., with $\rho < 1$). Apparently, the differences in the variances of the cycle times at the different queues vanish when the system tends to saturate.

The asymptotic properties discussed in Properties 1-3 and Remarks 4.1-4.3 have not been observed before in the general context of the present paper and provide new insights into the heavy-traffic behavior of polling systems with batched and simultaneous arrivals.

Copyright © Marcel Dekker, Inc. All rights reserved.

ORDER REPRINTS

5. APPROXIMATION

Theorem 4 suggests the following simple approximation for the expected delay at each of the queues: For $i = 1, \ldots, N$, $\rho < 1$,

$$E[W_i(app)] := \frac{\omega_i}{1-\rho}, \tag{49}$$

where ω_i is given by Theorem 4. Note that Theorem 4 implies that the approximation in (49) is asymptotically exact when the load tends to unity. To assess the accuracy of the approximation, in terms of "For which values of the load is the approximation accurate?", we have performed several numerical experiments. The results are outlined below.

Consider the following 3-queue model (referred to as model 1.A): The service discipline at Q_1 and Q_3 is exhaustive while Q_2 receives gated service. The service times at Q_1 are exponentially distributed with mean 1, while the service times at Q_2 are deterministic with mean 10, and the service times at Q_3 is gamma-distributed with mean 3 and squared coefficient of variation 10. The switch-over times from Q_1 to Q_2 are exponentially distributed with mean 0.05, the switch-over times from Q_2 to Q_3 are deterministic with mean 1 and the switch-over times from Q_3 to Q_1 are gamma-distributed with mean 2.5 and squared coefficient of variation 5. At any arrival epoch, the joint batch-size distribution is as follows: $\pi(1,2,4) = 1/2$, $\pi(2,1,0) = 1/4$ and $\pi(0,1,2) = 1/4$, or equivalently $\underline{K}^*(z_1,z_2,z_3) = z_1z_2^2z_3^4/2 + z_1^2z_2/4 + z_2z_3^2/4$. It is readily verified the mean batch sizes (including the possibility of batches of size 0) are given by $E[K_1] = 1$, $E[K_2] = 3/2$ and $E[K_3] = 5/2$, and that $K_{1,1} = K_{2,2} = 1/2$, $K_{1,2} = K_{2,1} = 3/2$, $K_{1,3} = K_{3,1} = 5/2$, $K_{2,3} = K_{3,2} = 5$ and $K_{3,3} = 13/2$. Note that the model is fairly asymmetric in the service times, the switch-over times and the batch-size distribution.

We consider the following 3 variants of the model: the "corresponding" models with independent batched Poisson arrivals (1.B), with independent unit Poisson arrivals and non-zero switch-over times (1.C), and with independent unit Poisson arrivals and zero switch-over times (1.D). To be specific, model 1.B is obtained from model 1.A by taking $K_{i,j} = 0$ for all i, j with $i \neq j$. Model 1.C is similar to model 1.B, except that the arrival rates at the different queues are Poisson with rates $\lambda_1 = \lambda$, $\lambda_2 = 3\lambda/2$ and $\lambda_3 = 5\lambda/2$ vand $K_{i,j} = 0$ for *all* i, j. In model 1.D is obtained from model 1.C by adding the assumption that $r = 0$. In the context of the decomposition in section 5, it is readily seen that in the limiting case $\rho \uparrow 1$, for model 1.A we have $\omega_i = I_i + II_i + III_i + IV_i$, for model 1.B we have $\omega_i = I_i + II_i + III_i$, whereas for model 1.C we have $\omega_i = I_i + II_i$ and for model 1.D, $\omega_i = I_i$. We implemented the DSA for the present model, according to equations (2)-(16), to calculate the exact values of the expected delay. Figure 1 shows the exact and approximated values of $(1-\rho)E[W_1]$ (i.e., the scaled expected delay at Q_1), as function of ρ, for models 1.A, 1.B, 1.C and 1.D. The solid lines indicate the exact results and the approximations are indicated by the dotted lines. Figure 1 demonstrates that the

Copyright © Marcel Dekker, Inc. All rights reserved.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

ORDER REPRINTS

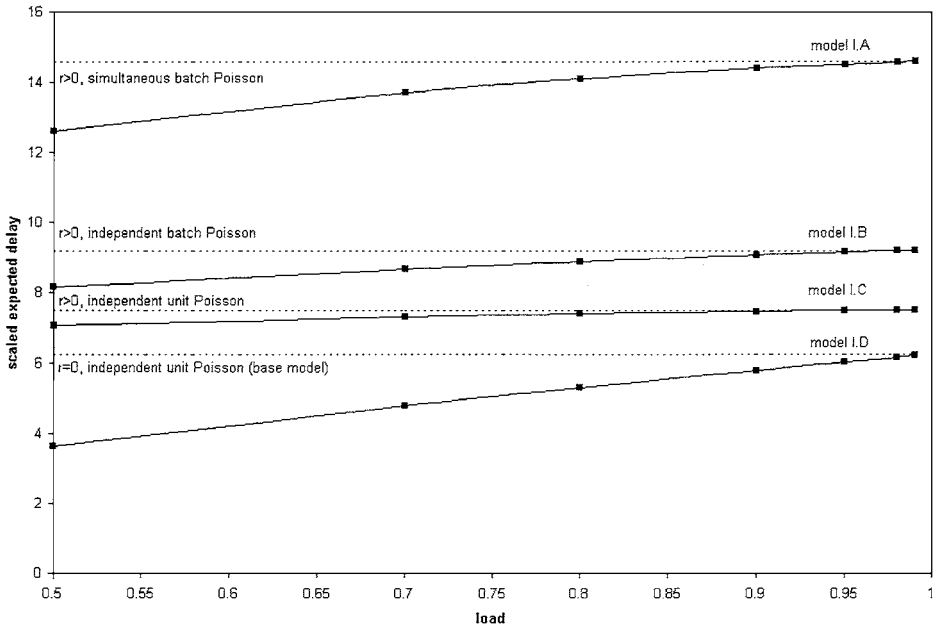

Figure 1. Exact and approximated values of the scaled expected delay as a function of the load (for models 1.A to 1.D).

exact and the approximated values of the scaled expected delay indeed converge to the same value when the load tends to unity for each of the models considered, which confirms the validity of the results. To quantify the accuracy of the approximation, we define the relative error of the approximation, $err\%$, by

$$err\% := \text{abs}\left(\frac{E[W_1(app)] - E[W_1]}{E[W_1]}\right) \times 100\%. \tag{50}$$

Figure 2 plots the relative error of the approximation as a function of the load. The results in Figure 2 demonstrate that the approximation is accurate for practical heavy-traffic scenarios. Let us qualify the quality of the approximation "good" when the relative error is less than 10% and "very good" when the error is less than 5%. Then we observe that for models 1.A and 1.B the quality of the approximation is considered good when the load exceeds only 60% and very good when the load is 75% or more. The results for model 1.C are very good in all cases considered. The worst quality is observed for model 1.D, but even in that case, the quality of the approximation is good when the load exceeds 88% and very good when the load exceeds 93%. The results demonstrate the usefulness of the asymptotic results for practical heavy-traffic scenarios.

To assess the accuracy of the approximation for large and highly asymmetric systems, we also consider a 10-queue model with the following parameters (referred to as model 2.A): The service times at queues 1, 2, 5, 6, 9 and 10 are exponentially distributed with mean 1, while the service times at queue 3 is

Copyright © Marcel Dekker, Inc. All rights reserved.

ORDER REPRINTS

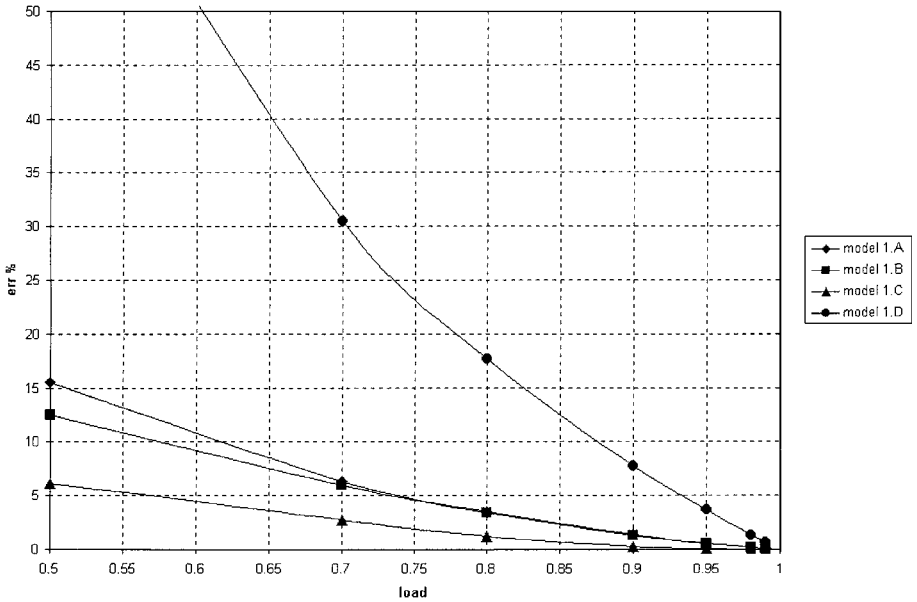

Figure 2. Relative error of the approximation of the expected delay as a function of the load (for models 1.A to 1.D).

deterministically distributed with mean 12, the service times at queue 4 are gamma-distributed with mean 1 and squared coefficient of variation 10 and the service times at queue 8 are exponentially distributed with mean 7. The switch-over times from Q_8 to Q_9 have a gamma distribution with mean 0.05 and squared coefficient of variation 5, the switch-over times from Q_9 to Q_{10} are deterministically distributed with mean 3, the switch-over times from Q_{10} are gamma-distributed with mean 0.7 and squared coefficient of variation 2, and all other switch-over times are exponentially distributed with mean 0.05. Queues 1, 2, 4 and 9 receive gated service, and all other queues are served exhaustively. The batch-size distribution is as follows: $\pi(1, 1, 1, 1, 1, 1, 1, 1, 1, 1) = 1/4$, $\pi(10\underline{e}_{10}) = 1/4$, $\pi(\underline{e}_1) = \pi(\underline{e}_3) = \pi(\underline{e}_4) = \pi(\underline{e}_6) = \pi(\underline{e}_7) = \pi(\underline{e}_8) = 1/20$, $\pi(3\underline{e}_2) = 1/10, \pi(2\underline{e}_5) = 3/40$ and $\pi(5\underline{e}_9) = 1/40$. It is readily verified that the mean batch-sizes for queues 1 to 10 are given by 12/40, 22/40, 12/40, 12/40, 16/40, 12/40, 12/40, 12/40, 15/40 and 90/40, respectively. Moreover, one may verify that $K_{2,2} = 6$, $K_{5,5} = 1$, $K_{9,9} = 20$, $K_{10,10} = 90$, $K_{i,i} = 0$ for $i = 1, 3, 4, 6, 7, 8$, and $K_{i,j} = 1/4$ for all $i \neq j$. Clearly, models 2.A-2.D are highly asymmetric in the service times, switch-over times, service policies and batch-size distributions.

We have calculated the exact and approximated values of the expected delay for different variants of the model, similarly to model 1 discussed above. More precisely, model 2.B is the same as model 2.A, but with independent batch-arrival processes. Model 2.C is the same as model 2.B, but with independent unit Poisson arrivals, and model 2.D is the same as model 2.C, but with zero switch-over times.

Copyright © Marcel Dekker, Inc. All rights reserved.

ORDER REPRINTS

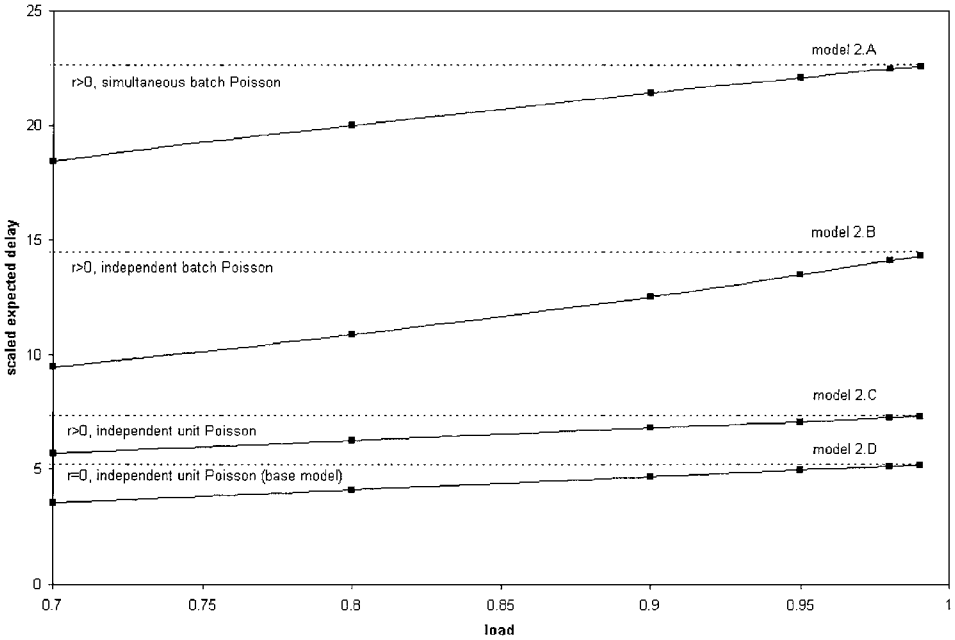

Figure 3. Exact and approximated values of the scaled expected delay as a function of the load (for models 2.A to 2.D).

Figure 3 shows the exact (solid lines) and approximated (dotted lines) values of the scaled expected delay at Q_1 as a function of the load, for model 2.A to 2.D. Figure 4 shows the relative error of the approximation as a function of the load. The results in Figure 3 indicate that the scaled expected delay figures indeed converge to ω_1, which confirms the validity of the exact asymptotic results. Figure 4 demonstrates that the approximation is fairly accurate for practical heavy-traffic scenarios. More precisely, qualifying the accuracy of the approximation "good" or "very good" when the error is less than 10% and 5%, respectively, Figure 4 shows that the accuracy of the approximation for model 2.A is good when the load exceeds only 84%, and even very good when the load 91% or more. The accuracy of the approximation for the other models is slightly worse, but is still considered good when the load exceeds 87-93%. Recall that the system parameters for models 2.A to 2.D are highly asymmetric and as such can be seen as a "worst-case" approximation. In most cases considered the accuracy of the approximation was found to be considerably better.

Remark 5.1. The results in Figures 1 and 3 also illustrate that the impact of batched and simultaneous arrivals may be significant and that neglecting these correlations in the arrival process may lead to highly erroneous performance predictions. These observations address the importance of the inclusion of batched and simultaneous arrivals in the model.

Copyright © Marcel Dekker, Inc. All rights reserved.

ORDER REPRINTS

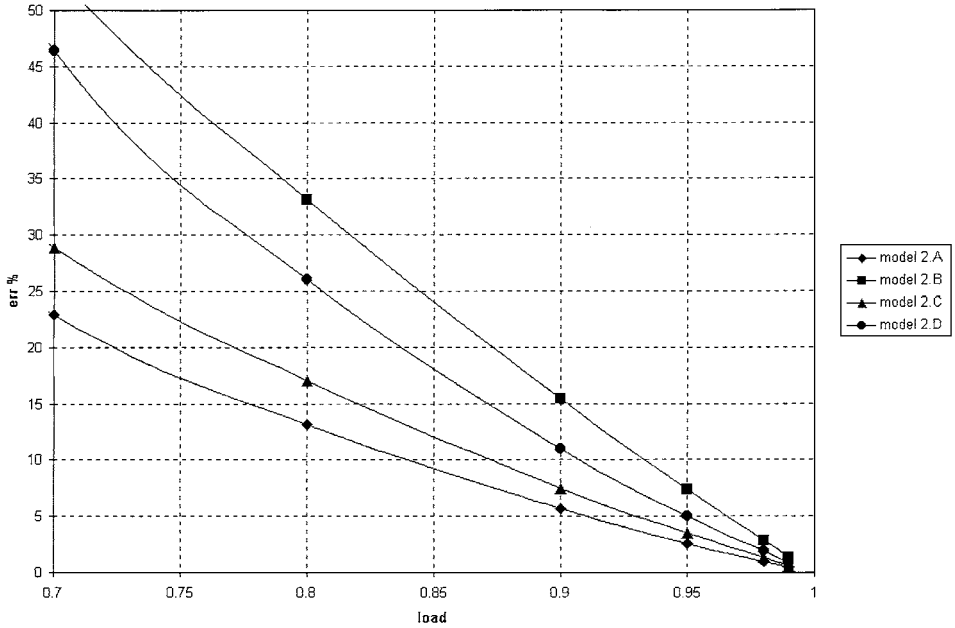

Figure 4. Relative error of the approximation of the expected delay as a function of the load (for models 2.A to 2.D).

Remark 5.2. For the special case of independent unit Poisson arrivals (i.e., with $K_{i,j} = 0$ for all i, j) approximations similar to (49) have been suggested by Boxma and Meister [3], Everitt [10] and Groenendijk [12]. These approxmations are derived along the following three steps: (1) obtain (possibly approximate) relations between de $E[W_i]$ and $Var[C_i]$ (see also Remark 4.3), (2) assume that $Var[C_i]$ is the same for all i, and (3) use the pseudo-conservation law (i.e., a closed-form expression for a specific weighted sum of the expected waiting times) to obtain the approximation in closed form. In this context, it is interesting to note that the results in the present paper actually prove that the variances of the cycle times for the different queues are asymptotically the same when the load tends to unity, see equation (48). These observations indicate that the approximations in [3, 10, 12] are also asymptotically exact when the load tends to unity (for mixtures of gated and exhaustive service).

To summarize, the numerical examples discussed above demonstrate that the impact of simultaneous batch arrivals may be significant and that the approximations (49), covering the impact of both batched and simultaneous arrivals, are useful in practical heavy-traffic scenarios.

6. TOPICS FOR FURTHER RESEARCH

In the present paper it is assumed that the joint batches arrive according to a Poisson arrival process. It would be interesting to analyze the queueing behavior

Copyright © Marcel Dekker, Inc. All rights reserved.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

ORDER REPRINTS

when the Poisson assumption is relaxed. The results in [7, 8] (for independent unit renewal processes) suggest that simple expressions can still be obtained for renewal-type arrival processes. Moreover, recent studies have demonstrated that many traffic streams in telecommunications networks exhibit even more complex correlation structures (cf. , e.g., [16, 20] and references therein). Extension of the heavy-traffic results to more general correlation structures in the arrival processes is a challenging topic for further research. It should be noted, however, that the implicit branching structure of the model studied is the present paper may be violated under more general arrival processes, so that the DSA-based approach discussed in the present paper may not be applicable.

In some applications the most important performance measure is the probability that the delay at a queue exceeds a certain threshold, rather than the average delay. To this end, the analysis presented in this paper may be extended to obtain expressions for the higher moments of the delay, and possibly for the complete distribution of the (scaled) delay incurred at each of the queues. For the case of independent unit Poisson arrivals, simple closed-form expressions for the higher moments, and even the Laplace-Stieltjes Transform, of the delay were obtained in [25, 26, 27]. Extension of the results to the higher moments and the probability distribution of the delay is an interesting topic for further research.

ACKNOWLEDGMENT

A partial and preliminary version of this paper has appeared in [28].

REFERENCES

[1] Altman, E., Konstantopoulos, P. and Liu, Z. (1992). Stability, monotonicity and invariant quantities in general polling systems. *Queueing Systems* **11**, 35–57.
[2] Altman, E. and Yechiali, U. (1993). Cyclic Bernoulli polling. *Z.O.R. Methods and Models in Operations Research* **38**, 55–76.
[3] Boxma, O.J. and Meister, B. (1987). Waiting-time approximations for cyclic-service systems with switch-over times. *Perf. Eval.* **7**, 299–308.
[4] Boxma, O.J., Groenendijk, W.P. and Weststrate, J.A. (1990). A pseudo-conservation law for service systems with a polling table. *IEEE Trans. Commun.* **38**, 1865–1870.
[5] Chiarawongse, J. and Srinivasan, M.M. (1991). On pseudo-conservation laws for the cyclic server system with compound Poisson arrivals. *Oper. Res. Lett.* **10**, 453–459.
[6] Choudhury, G. and Whitt, W. (1996). Computing transient and steady state distributions in polling models by numerical transform inversion. *Perf. Eval.* **25**, 267–292.
[7] Coffman, E.G., Puhalskii, A.A. and Reiman, M.I. (1995). Polling systems with zero switch-over times: a heavy-traffic principle. *Ann. Appl. Prob.* **5**, 681–719.
[8] Coffman, E.G., Puhalskii, A.A. and Reiman, M.I. (1998). Polling systems in heavy-traffic: a Bessel process limit. *Math. Oper. Res.* **23**, 257–304.
[9] Cooper, R.B., Niu, S.-C. and Srinivasan, M.M. (1997). Setups in polling models: does it make sense to set up if no work is waiting? *J. Appl. Prob.* **36**, 585–592.

Copyright © Marcel Dekker, Inc. All rights reserved.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

ORDER REPRINTS

[10] Everitt, D. (1986). Simple approximations for token rings. *IEEE Trans. Commun.* **34**, 719–721.
[11] Fricker, C. and Jaïbi, M.R. (1994). Monotonicity and stability of periodic polling models. *Queueing Systems* **15**, 211–238.
[12] Groenendijk, W.P. (1988). Waiting-time approximations for cyclic-service systems with mixed service strategies. In: *Teletraffic Science for New Cost-Effective Systems*, ed. M. Bonatti (North-Holland, Amsterdam), 1434–1441.
[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] Kroese, D.P. (1997). Heavy traffic analysis for continuous polling models. *J. Appl. Prob.* **34**, 720–732.
[15] Kudoh, S., Takagi, H. and Hashida, O. (2000). Second moments of the waiting time in symmetric polling systems. *J. Oper. Res. Soc. of Japan* **43**, 306–316.
[16] Leland, W., Taqqu, M., Willinger, W. and Wilson, D. (1994). On the self-similar nature of Ethernet traffic. *IEEE/ACM Trans. Netw.* **2**, 1–15.
[17] Levy, H. and Sidi, M. (1991). Polling models: applications, modeling and optimization. *IEEE Trans. Commun.* **38**, 1750–1760.
[18] Levy, H. and Sidi, M. (1991). Polling systems with simultaneous arrivals. *IEEE Trans. Commun.* **39**, 823–827.
[19] Markowitz, D. (1995). *Dynamic Scheduling of Single-Server Queues with Setups*: *A Heavy Traffic Approach.* Ph.D. Thesis, Operations Research Center, MIT, Cambridge, MA.
[20] Paxson, V. and Floyd, S. (1995). Wide area traffic: the failure of Poisson modeling. *IEEE/ACM Trans. Netw.* **3**, 226–244.
[21] Reiman, M.I. and Wein, L.M. (1998). Dynamic scheduling of a two-class queue with setups. *Oper. Res.* **46**, 532–547.
[22] 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.
[23] 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), 119–146.
[24] Van der Mei, R.D. and Levy, H. (1998). Expected delay in polling systems in heavy traffic. *Adv. Appl. Prob.* **30**, 586–602.
[25] Van der Mei, R.D. (1999). Polling systems in heavy traffic: higher moments of the delay. *Queueing Systems* **31**, 365–394.
[26] Van der Mei, R.D. (2000). Polling systems with switch-over times in heavy traffic: moments of the delay. *Queueing Systems* **36**, 381–404.
[27] Van der Mei, R.D. (1999). Distribution of the delay in polling systems in heavy traffic. *Perf. Eval.* **38**, 133–148.
[28] Van der Mei, R.D. (2000). Polling systems with correlated batch arrivals. In: *Proc. 1st Polish-German Symposium on Teletraffic* (Dresden, September 25-26), 161–168.

Received: February 1, 2000
Revised: September 24, 2000
Accepted: October 12, 2000

Copyright © Marcel Dekker, Inc. All rights reserved.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

Request Permission or Order Reprints Instantly!

Interested in copying and sharing this article? In most cases, U.S. Copyright Law requires that you get permission from the article's rightsholder before using copyrighted content.

All information and materials found in this article, including but not limited to text, trademarks, patents, logos, graphics and images (the "Materials"), are the copyrighted works and other forms of intellectual property of Marcel Dekker, Inc., or its licensors. All rights not expressly granted are reserved.

Get permission to lawfully reproduce and distribute the Materials or order reprints quickly and painlessly. Simply click on the "Request Permission/Reprints Here" link below and follow the instructions. Visit the U.S. Copyright Office for information on Fair Use limitations of U.S. copyright law. Please refer to The Association of American Publishers' (AAP) website for guidelines on Fair Use in the Classroom.

The Materials are for your personal use only and cannot be reformatted, reposted, resold or distributed by electronic means or otherwise without permission from Marcel Dekker, Inc. Marcel Dekker, Inc. grants you the limited right to display the Materials only on your personal computer or personal wireless device, and to copy and download single copies of such Materials provided that any copyright, trademark or other notice appearing on such Materials is also retained by, displayed, copied or downloaded as part of the Materials and is not removed or obscured, and provided you do not edit, modify, alter or enhance the Materials. Please refer to our Website User Agreement for more details.

Order now!

Reprints of this article can also be ordered at
http://www.dekker.com/servlet/product/DOI/101081STM100002274