Adv. Appl. Prob. **30,** 586–602 (1998)
Printed in Northern Ireland

EXPECTED DELAY ANALYSIS OF POLLING SYSTEMS IN HEAVY TRAFFIC

R.D. VAN DER MEI,* *AT&T Labs*
H. LEVY,** *Tel-Aviv University*

Abstract

We study the expected delay in a cyclic polling model with mixtures of *exhaustive* and *gated* service in heavy traffic. We obtain closed-form expressions for the mean delay under standard heavy-traffic scalings, providing new insights into the behaviour of polling systems in heavy traffic. The results lead to excellent approximations of the expected waiting times in practical heavy-load scenarios and moreover, lead to new results for optimizing the system performance with respect to the service disciplines.

Keywords: polling systems; heavy traffic; expected delay; optimization

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

1. Introduction

The basic polling system consists of a number of queues attended by a single server which visits the queues in cyclic order to render service to the customers waiting in the queues. Polling models find many applications in computer-communication systems, and are also widely applicable in the areas of maintenance, manufacturing and production. During the last three decades, the analysis of polling models has received much attention in the literature. The reader is referred to [14] for an overview of the applicability of polling models, and to [22] for a review of the state-of-the-art analysis of polling models. Some variations of polling models do not allow for an exact and detailed analysis, and the others usually require the use of numerical techniques to determine performance measures of interest.

The ultimate goal of performance modeling and analysis is to obtain the 'best' possible system performance. The proper operation of the system is particularly critical when the system is heavily loaded. However, the efficiency of each of the numerical algorithms degrades significantly for heavily loaded, highly asymmetric systems with a large number of queues. Moreover, numerical techniques can only contribute to the understanding of the behaviour of the system to a limited extent. Exact closed-form expressions provide much more insight into the dependence of the performance measures on the system parameters. These observations raise the importance of an exact asymptotic analysis of the performance of polling models in heavy traffic.

In this paper we show that a class of polling models allows for an exact analysis under heavy traffic assumptions. The results are useful for understanding the heavy-traffic behaviour of the system. Furthermore, they lead to an excellent approximation of the expected delays

Received 29 April 1996; revision received 29 January 1997.
* Postal address: AT&T Labs, P.O. Box 3030, Holmdel, NJ 07733, USA. E-mail address: hanoch@math.tau.ac.il
** Postal address: Tel-Aviv University, Department of Computer Science, Tel Aviv, Israel.

at practical heavy-load scenarios and to the derivation of rules for optimization of the system performance.

The literature on polling models reveals a striking difference in the complexity between different polling models. Recently, 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 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 a series of numerical algorithms have been proposed to determine the moments of the delay at the queues by solving sets of linear equations (see [21] for references). Recently, the efficiency of the numerical techniques has been considerably improved upon 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 (cf. [12]). Choudhury and Whitt [5] use the numerical transform-inversion to extend the DSA to the determination of tail probabilities and even transient performance measures. Polling models that do not have an MTBP-structure generally require much more computational effort (cf. [2, 13]).

Although the number of papers on polling models is impressive, relatively few papers have been devoted to the exact analysis of polling models in heavy traffic. An exception is made by Coffman et al. [6]. For a two-queue model with exhaustive service at both queues and with zero switch-over times they show that, under standard heavy-traffic assumptions and scalings, the total unfinished work converges to a Reflected Brownian Motion (RBM), whereas the workloads of individual queues change at a rate that becomes infinite in the limit. Based on a partial conjecture, it is shown in [6] that similar properties hold for systems with more than two queues. Moreover, it is shown in [7] that for non-zero switch-over times the scaled process can be described by a Bessel process. Based on these observations, exact expressions can be derived for the main performance measures of interest. Assuming that the observations in [6, 7] also hold for non-exhaustive policies, Reiman and Wein [17] study the problem of determining optimal dynamic scheduling problems for two-queue models, with either setup times or switch-over times under heavy-traffic assumptions, by approximating the dynamic scheduling problems by diffusion control problems. Markowitz [16] extends the results in [17] to the multi-class case.

Consider an asymmetric polling model with general mixtures of exhaustive and gated service. Here, we study the expected delay at each of the queues in a heavy-traffic environment. All the queues become unstable when the system load (denoted by ρ) approaches 1 (cf. [9]). More precisely, the expected delay at the queues possesses a first-order pole at $\rho = 1$. Therefore, the main performance measure of interest will be the limit of the scaled expected delay, defined as $(1 - \rho)$ times the expected delay, when ρ tends to 1. The scaled expected delay indicates the rate at which the expected delay tends to infinity when ρ tends to 1. We derive closed-form expressions for the scaled expected delay at each of the queues. The key role in the derivation of the results is played by the concept of descendant sets. The use of descendant sets has been very useful in obtaining the moments of the delay in stable polling models (cf. [12]). The derivation of closed-form expressions for the expected delay, however, requires a new view on the concept of descendant sets. We numerically examine the quality of these results at practical high load situations. The results indicate that the asymptotic results lead to excellent approximations for the expected delay, when the load exceeds 85%. Finally, we consider the optimization problem of assigning service disciplines (gated or exhaustive) to each of the queues as to minimize an arbitrary weighted sum of the scaled expected delays.

We show that the queues with the highest weight/load ratio should be served exhaustively. Moreover, for the case of zero switch-over times, we obtain a simple and fast algorithm for solving the problem.

The remainder of the paper is organized as follows. In Section 2 the model is described in detail. In Section 3 the DSA is reviewed and some new relations useful for our analysis are derived. In Section 4 closed-form expressions for the scaled expected waiting times are derived. In Section 5 we use numerical results to assess the quality of an expected delay approximation based on the heavy traffic limit results. In Section 6 we address the problem of optimizing the system performance with respect to the service disciplines. Finally, in Section 7 we discuss a number of topics for further research. The proofs of the various results are discussed in the appendix.

2. Model description and the main result

Consider a system consisting of N infinite-buffer queues, $Q_1, \ldots, Q_N$, and a single server which visits and serves them 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 first and second moments b_i and $b_i^{(2)}$. Denote $\boldsymbol{b} = (b_1, \ldots, b_N)$. The first two moments of an arbitrary service time are denoted by

$$b = \sum_{i=1}^{N} \lambda_i b_i / \Lambda \quad \text{and} \quad b^{(2)} = \sum_{i=1}^{N} \lambda_i b_i^{(2)} / \Lambda.$$

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 the *polling instants* at Q_i. 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 $E := \{i : Q_i \text{ is served exhaustively}\}$ and $G := \{i : Q_i$ is served according to the gated policy$\}$.

After completing service at Q_i the server proceeds to Q_{i+1} (where queues are indexed in modulo N), incurring a *switch-over period* whose duration is an independent random variable R_i. The first two moments of R_i are denoted by r_i and $r_i^{(2)}$. Denote the first moment of the total switch-over time in a cycle by $r = \sum_{i=1}^{N} r_i$, and the second moment by $r^{(2)} = \sum_{i=1}^{N} r_i^{(2)} + \sum_{i,j=1}^{N} \sum_{i \neq j} r_i r_j$. It is assumed throughout the paper that $r > 0$, unless indicated otherwise.

All interarrival times, service times and switch-over times are assumed to be mutually independent and independent of the state of the system.

A necessary and sufficient condition for the stability of the system is $\rho < 1$ (cf. [9]). In the following, it is assumed that this condition is satisfied, and that the system is in steady state, unless indicated otherwise.

Denote by W_k the delay incurred by an arbitrary customer at Q_k. Our main interest is in the behaviour of $E[W_k]$, the expected delay at Q_k, in heavy traffic. Throughout, $E[W_k]$ will be considered as function of ρ. More specifically, we assume that the arrival rates are parameterized as $\lambda_i = a_i \rho$, where relative arrival rates a_i $(= \lambda_i / \rho)$ remain fixed. It is known that when $\rho \uparrow 1$, all queues become instable and hence, $E[W_k]$ tends to infinity for all k (cf.

[9]). Although a rigorous proof has not been found in the literature, we assume that $E[W_k]$ has a first-order pole at $\rho = 1$ as follows:

$$E[W_k] = \frac{\omega_k}{1-\rho} + o((1-\rho)^{-1}), \quad (\rho \uparrow 1), \quad \text{for } k = 1, \ldots, N, \tag{1}$$

where $o((1-\rho)^{-1})$ stands for a function of ρ which becomes negligible compared with $(1-\rho)^{-1}$ when $\rho \uparrow 1$. Based on (1), the analysis will be oriented towards the determination of

$$\omega_k = \lim_{\rho \uparrow 1} (1-\rho) E[W_k], \quad k = 1, \ldots, N, \tag{2}$$

i.e. the scaled expected delay at Q_k, also referred to as the heavy-traffic residue of $E[W_k]$ at $\rho = 1$. In words, ω_k indicates the *rate* at which $E[W_k]$ tends to infinity as $\rho \uparrow 1$.

The main result of the paper is the following theorem.

Theorem 1 (Main Result.) *The scaled expected waiting times at each of the queues is given by the following closed-form expression,*

$$\omega_i = \frac{1+\rho_i}{\sum_{j\in G} \rho_j(1+\rho_j) + \sum_{j\in E} \rho_j(1-\rho_j)} \frac{b^{(2)}}{2b} + \tfrac{1}{2} r(1+\rho_i), \quad i \in G,$$

$$\omega_i = \frac{1-\rho_i}{\sum_{j\in G} \rho_j(1+\rho_j) + \sum_{j\in E} \rho_j(1-\rho_j)} \frac{b^{(2)}}{2b} + \tfrac{1}{2} r(1-\rho_i), \quad i \in E.$$

The result implies that for $\rho \approx 1$ the expected waiting times can be approximated by the following expression:

$$E[W_i] \approx \frac{1}{1-\rho} \left[\frac{(1+\rho_i) I_{\{i\in G\}} + (1-\rho_i) I_{\{i\in E\}}}{\sum_{j\in G} \rho_j(1+\rho_j) + \sum_{j\in E} \rho_j(1-\rho_j)} \frac{b^{(2)}}{2b} + \tfrac{1}{2} r((1+\rho_i) I_{\{i\in G\}} + (1-\rho_i) I_{\{i\in E\}}) \right]. \tag{3}$$

Sections 3 and 4 will be mainly devoted to the derivation of Theorem 1. In Section 3 we review the DSA and obtain new relations useful for the analysis. In Section 4 these relations are used to derive Theorem 1.

In what follows, vectors are N-dimensional and matrices are N by N, unless indicated otherwise. The vector $\boldsymbol{e}_i$ stands for the ith unit vector, $i = 1, \ldots, N$. Each entry of the vector $\mathbf{1}$ equals 1. The matrix $\mathbf{0}$ stands for the null matrix and $\boldsymbol{I}$ stands for the identity matrix. Let the norm of a matrix $\boldsymbol{A} = (a_{i,j})$ be defined as $\|\boldsymbol{A}\| := \max_{i,j} |a_{i,j}|$. I_F stands for the indicator function on the event F. Indices corresponding to queue numbers are cyclic: index $i+1$ should be read as $(i+1)$ modulo N.

3. The descendant set approach: review and new results

Let X_k be the number of customers at Q_k at a polling instant at Q_k, when the system is in equilibrium. The mean waiting time at Q_k can be expressed in terms of the first two moments of X_k as follows (see [20]):

$$E[W_k] = \frac{E[(X_k)^2] - E[X_k]}{2\lambda_k E[X_k]} (1+\rho_k), \quad k \in G, \tag{4}$$

and

$$E[W_k] = \frac{E[(X_k)^2] - E[X_k]}{2\lambda_k E[X_k]} + \frac{\lambda_k b_k^{(2)}}{2(1-\rho_k)}, \quad k \in E. \tag{5}$$

The quantities $E[X_k]$ can be derived in closed form. To this end, note that simple balancing arguments yield the following relations,

$$\begin{aligned} E[X_k] &= \lambda_k r + \lambda_k \sum_{j \neq k} \tau_j E[X_j] + E[X_k]\rho_k, \quad k \in G, \\ E[X_k] &= \lambda_k r + \lambda_k \sum_{j \neq k} \tau_j E[X_j], \quad k \in E, \end{aligned} \tag{6}$$

where $\tau_j := b_j$ for $j \in G$ and $\tau_j := b_j/(1-\rho_j)$ for $j \in E$. The reader may verify that (6) is uniquely solved by $E[X_k] = \lambda_k r/(1-\rho)$ $(k \in G)$, and $E[X_k] = \lambda_k(1-\rho_k)r/(1-\rho)$ $(k \in E)$.

However, the variables $E[(X_k)^2]$ cannot be obtained generally in closed form. In the literature, there are various techniques available to determine $E[(X_k)^2]$. We will focus on the recently developed Descendant Set Approach (DSA). The DSA provides a means to compute the quantities $E[(X_k)^2]$ very efficiently, and moreover, will appear to be particularly useful for obtaining ω_k, our main performance measures of interest.

In Sections 3.1–3.3 we review the basic principles of the DSA; the reader is referred to [12] for a more detailed discussion. New results are provided in Section 3.4.

3.1. Terminology

All customers of a polling system can be classified into two classes: (1) *originators* and (2) *non-originators*. An originator is a customer who 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 for a fixed Q_k. To this end, the DSA concentrates on the determination of $X_k(P)$, defined as the number of customers at Q_k present at an arbitrary fixed polling instant P at Q_k. P is referred to as the *reference point* at Q_k. The main idea is the observation that each of these $X_k(P)$ customers belongs to the descendant set of *exactly one* originator. Therefore, the DSA concentrates on an arbitrary tagged customer T that arrived at Q_i in the past and on calculating the number of type-k descendants it has at P. Summing up these numbers over all past originators yields $X_k(P)$ and hence X_k, 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 (see Figure 1). Let $P_{N,0}$ be an arbitrary polling instant at Q_N, and for $i = N-1, \ldots, 1$, let $P_{i,0}$ be recursively defined as the first 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 $P_{i,c-1}$, $i = 1, \ldots, N$. We 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 $A_{(i,c),k}$ to be the number of type-k descendants it has at $P_{k,0}$. In this way, $A_{(i,c),k}$ can be viewed as the *contribution* of T to $X_k(P_{k,0})$. Define $\alpha_{(i,c),k} = E[A_{(i,c),k}]$.

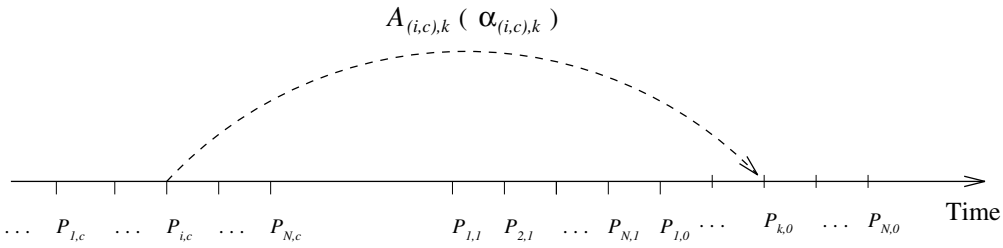

FIGURE 1: Contribution of an (i, c)-customer to $X_k(P_{k,0})$.

3.2. Relation to expected waiting times

The expected waiting times can be expressed in terms of the coefficients $\alpha_{(i,c),k}$ as follows (cf. [19]):

$$E[W_k] = \frac{1+\rho_k}{2}\frac{r}{1-\rho} + \frac{1+\rho_k}{2\lambda_k^2}\sum_{i=1}^{N}\frac{\psi_{i,k}}{\rho_i^2}\left(\lambda_i b_i^{(2)} + (r_i^{(2)} - r_i^2)\frac{1-\rho}{r}\right), \quad k \in G, \tag{7}$$

and

$$E[W_k] = \frac{1-\rho_k}{2}\frac{r}{1-\rho} + \frac{1}{2(1-\rho_k)}\Bigg[\lambda_k b_k^{(2)} + (r_{k-1}^{(2)} - r_{k-1}^2)\frac{1-\rho}{r} + \frac{1}{\lambda_k^2}\sum_{i=1}^{N}\frac{\psi_{i,k}}{\rho_i^2}\left(\lambda_i b_i^{(2)} + (r_{i-1}^{(2)} - r_{i-1}^2)\frac{1-\rho}{r}\right)\Bigg], \quad k \in E, \tag{8}$$

where

$$\psi_{i,k} := \begin{cases} \lambda_i^2 \displaystyle\sum_{c=1}^{\infty} {\alpha_{(i,c),k}}^2, & i = 1, \ldots, k-1, \\ \lambda_i^2 \displaystyle\sum_{c=1}^{\infty} {\alpha_{(i,c-1),k}}^2, & i = k, \ldots, N. \end{cases} \tag{9}$$

From (7)–(9), it follows that $E[W_k]$ can be expressed *only* in terms of $\alpha_{(i,c),k}$, i.e. the *first moments* of the numbers of descendants of the originators. This observation is generally not true for other service disciplines for which the DSA is applicable: in general, $E[W_k]$ must be expressed in terms of the the first *and second* moments of $A_{(i,c),k}$. However, it has been shown in [12] that for the special case of exhaustive service at all queues (or gated service at all queues), the second moments of $A_{(i,c),k}$ may be *eliminated*. Following the analysis in [12], one may verify that the second moments can also be eliminated for the model under our consideration, i.e. with mixtures of exhaustive and gated service.

3.3. The DSA original analysis: recursion via immediate children of predecessor

The DSA is based on recursive relations between the variables $\alpha_{(i,c),k}$ which we review next. 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 $X_k(P_{k,0})$. Let us first assume that Q_i is served according to the gated service policy. It is readily seen that the contribution of $T_i(P_{i,c})$ to $X_k(P_{k,0})$ is equal to the total contribution of all its *immediate children*, i.e. the customers which arrive

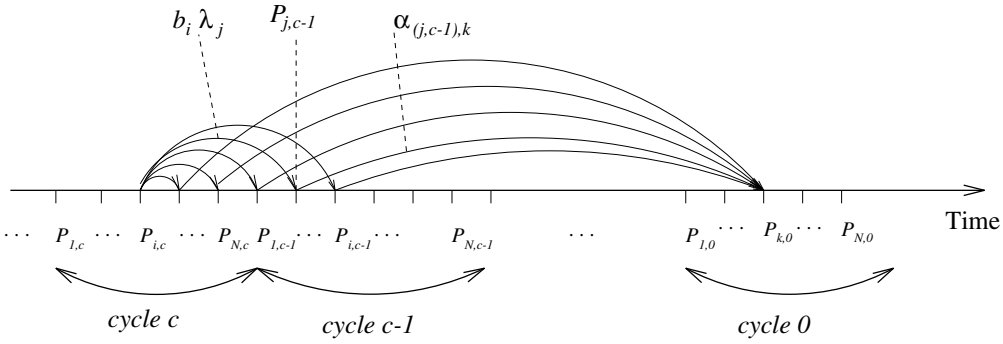

FIGURE 2: Original DSA recursion: via the immediate children of predecessor.

during the service of $T_i(P_{i,c})$, to $X_k(P_{k,0})$. The DSA is based on computing $\alpha_{(i,c),k}$ from the contribution of the children of $T_i(P_{i,c})$. The expected number of type-j children that $T_i(P_{i,c})$ has, is equal to $b_i\lambda_j$, $j = 1, \ldots, N$. Hence, the expected number of type-k descendants at $P_{k,0}$ is given by the following expression (see [12]). For $c = 0, 1, \ldots$,

$$\alpha_{(i,c),k} = b_i\left[\sum_{j=i+1}^{N} \lambda_j\alpha_{(j,c),k} + \sum_{j=1}^{i} \lambda_j\alpha_{(j,c-1),k}\right], \quad i \in G. \tag{10}$$

A similar expression can be obtained for the case where Q_i is served exhaustively. For $c = 0, 1, \ldots$,

$$\alpha_{(i,c),k} = \frac{b_i}{1-\rho_i}\left[\sum_{j=i+1}^{N} \lambda_j\alpha_{(j,c),k} + \sum_{j=1}^{i-1} \lambda_j\alpha_{(j,c-1),k}\right], \quad i \in E. \tag{11}$$

Note that $b_i/(1-\rho_i)$ takes into consideration the contribution of all type-i customers which arrive at Q_i during the sub-busy period of $T_i(P_{i,c})$.
The approach of this recursion is depicted in Figure 2. To carry out the recursion, one needs to set the initial values of the variables $\alpha_{(i,c),k}$. However, these initial values do not play an essential role in our analysis, and are given as follows just for the sake of completeness: $\alpha_{(k,0),k} := 1$; $\alpha_{(i,0),k} := 0$ $(i = k+1, \ldots, N)$; $\alpha_{(i,-1),k} := 0$ $(i = 1, \ldots, k-1)$. Starting with these initial values, all coefficients $\alpha_{(i,c),k}$ can be recursively determined according to (10)–(11). To be precise, relations (10) and (11) are only defined for $c = 0, 1, \ldots$, for $i = 1, \ldots, k-1$, and for $c = 1, 2, \ldots$, $i = k, \ldots, N$. In this way, for fixed k the variables $\alpha_{(i,c),k}$ can be computed in the order $\alpha_{(i,0),k}$, for $i = k-1, k-2, \ldots, 1$, followed by $\alpha_{(i,c),k}$, $i = N, N-1, \ldots, 1$, for $c = 1, 2, \ldots$.

3.4. New DSA analysis: recursion via immediate parents of descendants

The recursive equations in (10)–(11) relate the variables $\alpha_{(i,c),k}$ for a fixed k, thus leading to the derivation of the expected delay in a single queue. However, for our analysis we need the *interaction* between the queues and thus it requires relations between the variables $\alpha_{(i,c),k}$ and $\alpha_{(i,c),l}$ for $l \neq k$. The derivation of such relations requires that we conduct the recursion in a different way, i.e. rather than carrying it out via the children of the predecessor (see Figure 2), we carry it out via the parents of the descendants (see Figure 3).

We consider a tagged (i, c)-customer $T_i(P_{i,c})$, a type-i customer present at Q_i at $P_{i,c}$. We want to find the contribution of $T_i(P_{i,c})$ to $X_k(P_{k,0})$. To do so, we consider the most recent

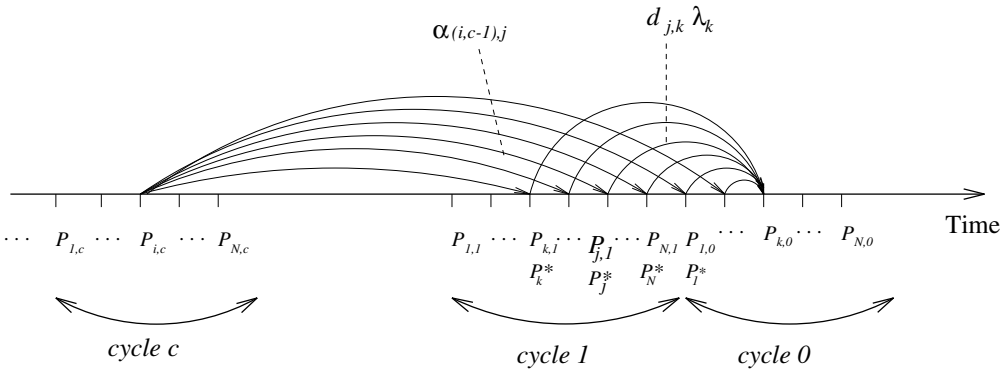

FIGURE 3: New recursion: via the immediate parents of descendants.

polling instants of Q_j, $j = 1, \ldots, N$, prior to $P_{k,0}$, denoted by P_j^* (see Figure 3). We will recursively derive the contribution of $T_i(P_{i,c})$ to $X_k(P_{0,k})$ as a function of the contribution of $T_i(P_{i,c})$ to $X_j(P_j^*)$. A crucial observation is that the distribution of the contribution of $T_i(P_{i,c})$ to $X_j(P_j^*)$ is identical to that of $A_{(i,c),j}$ for $j = 1, \ldots, k-1$, and to $A_{(i,c-1),j}$ for $j = k, \ldots, N$. This observation can be shown by a simple shift of indices and results from stationarity.

It remains to relate the number of descendants of $T_i(P_{i,c})$ at P_j^* to the number of type-k descendants of $T_i(P_{i,c})$ at $P_{k,0}$. To this end, we observe that each of the type-k customers at $P_{k,0}$ has arrived during the service of exactly one customer present at P_j^*, $j = 1, \ldots, N$ (for exhaustive service we add the requirement $j \neq k$), referred to as the *immediate parent*. (For conveniently accommodating the exhaustive service under this definition, we define in this case the service of a customer to be its sub-busy period, namely the service of the customer itself as well as the service of all the customers who recursively arrive to the same queue during the service of this customer.) The expected number of type-k customers at $P_{k,0}$ whose immediate parent is of type j, is given by $\alpha_{(i,c),j} d_{j,k}$ ($j = 1, \ldots, k-1$) and by $\alpha_{(i,c-1),j} d_{j,k}$ ($j = k, \ldots, N$), where $d_{j,k} = b_j$ if $j \in G$, and $d_{j,k} = I_{\{j \neq k\}} b_j/(1-\rho_j)$ if $j \in E$. These observations lead to the following relation: for $c = 0, 1, \ldots$,

$$\alpha_{(i,c),k} = \lambda_k \left[\sum_{j=1}^{k-1} \alpha_{(i,c),j} d_{j,k} + \sum_{j=k}^{N} \alpha_{(i,c-1),j} d_{j,k} \right]. \tag{12}$$

The initial values of this recursion can be defined similarly to those in Section 3.3; however, these values are not essential for our analysis and thus are omitted. It will be useful to express (12) in matrix notation. To this end, let $\alpha_{(i,c),\cdot}$ be the vector whose kth element is $\alpha_{(i,c),k}$ for $k = 1, \ldots, i$, and $\alpha_{(i,c-1),k}$ for $k = i+1, \ldots, N$ ($i = 1, \ldots, N$, $c = 0, 1, \ldots, \infty$). Further, let $\hat{\boldsymbol{M}}_k$ be the matrix whose (i, j)th element equals $I_{\{i=j\}}$ for $i \neq k$ and $\lambda_k d_{j,k}$ for $i = k$. Define $\boldsymbol{M}_i = \hat{\boldsymbol{M}}_i \cdots \hat{\boldsymbol{M}}_1 \hat{\boldsymbol{M}}_N \cdots \hat{\boldsymbol{M}}_{i+1}$ ($i = 1, \ldots, N$). Using these definitions, the relations (12) and the initial condition (see Section 3.3) can be expressed as follows: for $c = 1, 2, \ldots$, for $i = 1, \ldots, N$,

$$\alpha_{(i,0),\cdot} = \boldsymbol{e}_i, \quad \alpha_{(i,c),\cdot} = \boldsymbol{M}_i \alpha_{(i,c-1),\cdot} = (\boldsymbol{M}_i)^c \boldsymbol{e}_i, \tag{13}$$

where the third equality is a result of iterating the second equality c times.

Similarly, the relations (10) and (11) can be expressed in matrix notation. To this end, define the matrix $\boldsymbol{N}_k = \hat{\boldsymbol{N}}_k \cdots \hat{\boldsymbol{N}}_N \hat{\boldsymbol{N}}_1 \cdots \hat{\boldsymbol{N}}_{k-1}$, where $\hat{\boldsymbol{N}}_i$ is the matrix whose (j, k)th element equals

$I_{\{j=k\}}$ for $j \neq i$, and for $j = i$, $\lambda_k b_i$ if $i \in G$, $I_{\{i\neq k\}}\lambda_k b_i/(1-\rho_i)$ if $i \in E$. With these definitions, (10)–(11) can be denoted as $\alpha_{(\cdot,0),k} = \boldsymbol{e}_k$; $\alpha_{(\cdot,c),k} = \mathcal{N}_k\alpha_{(\cdot,c-1),k}$, $c = 1, 2, \ldots$. In the remainder, relations (13) will be more useful in obtaining the main result.

4. Analysis

The aim of the analysis is to obtain a proof for Theorem 1, giving closed-form expressions for ω_k, $k = 1, \ldots, N$. To show how they are related to the variables $\alpha_{(i,c),k}$, one may verify the following, by multiplying both sides of (7) and (8) by $(1-\rho)$ and letting $\rho \uparrow 1$, i.e.

$$\omega_k = \frac{r(1+\rho_k)}{2} + \frac{(1+\rho_k)}{2\lambda_k^2}\sum_{i=1}^{N}\frac{\tilde{\psi}_{i,k}}{\rho_i^2}\lambda_i b_i^{(2)}, \quad k \in G, \tag{14}$$

and

$$\omega_k = \frac{r(1-\rho_k)}{2} + \frac{1}{2(1-\rho_k)}\frac{1}{\lambda_k^2}\sum_{i=1}^{N}\frac{\tilde{\psi}_{i,k}}{\rho_i^2}\lambda_i b_i^{(2)}, \quad k \in E; \tag{15}$$

where the variables $\tilde{\psi}_{i,k}$, are defined by

$$\tilde{\psi}_{i,k} := \lim_{\rho\uparrow 1}\ (1-\rho)\psi_{i,k} \ \ (i,k = 1, \ldots, N). \tag{16}$$

The observation that $E[W_k]$ has a first-order pole at $\rho = 1$ for all $k = 1, \ldots, N$, implies that $\psi_{i,k}$ also has a first order pole at $\rho = 1$, for $i, k = 1, \ldots, N$. This ensures that the limits in (16) are well-defined.

Relations (14) and (15) indicate that in order to relate ω_k, $k = 1, \ldots, N$, one has to relate the variables $\tilde{\psi}_{i,k}$, $i, k = 1, \ldots, N$, for different i and k. Recall from (9) that $\psi_{i,k}$ is defined as an *infinite sum* of the sequence $\{\alpha_{(i,c),k}{}^2, c = 0, 1, \ldots\}$ and therefore the rate at which it tends to infinity is determined by the *tail behaviour* of the sequences $\{\alpha_{(i,c),k}, c = 0, 1, \ldots\}$.

4.1. Analysis of the descendant set variables

The variables $\alpha_{(i,c),k}$ are fully determined by the set of relations (13). More precisely, (13) constitutes a set of homogeneous difference equations of the first order. From the literature on difference equations it is well-known that the variables $\alpha_{(i,c),k}$ can be solved directly if the eigenvalues and eigenvectors of $\boldsymbol{M}_i$ are known. In general, however, the eigenvalues and eigenvectors of $\boldsymbol{M}_i$ are unknown for $\rho < 1$.

In this section we will derive some properties of $\boldsymbol{M}_i$ which, in turn, provide properties of the tail behaviour of the sequences $\{\alpha_{(i,c),k}, c = 0, 1, \ldots\}$. In Section 4.1.1, it is shown that $\boldsymbol{M}_k$ is decomposable into two parts, one of which becomes dominant in $(\boldsymbol{M}_i)^c$ for large c. Based on this property, we show in Section 4.1.2 that the sequence $\{\alpha_{(i,c),k}, c = 0, 1, \ldots\}$ converges for all $i, k = 1, \ldots, N$. In Section 4.1.3 we derive some properties of the limiting values of the sequence.

4.1.1. *Decomposition.* The following lemma decomposes $(\boldsymbol{M}_i)^c$ into two parts.

Lemma 1 (Decomposition.) *$\boldsymbol{M}_i$ has a maximal eigenvalue $\gamma_{\max,i}$ which is positive, has multiplicity 1, and has non-negative associated right and left eigenvectors $\boldsymbol{u}_i = (u_{i,1}, \ldots, u_{i,N})$ and $\boldsymbol{v}_i = (v_{i,1}, \ldots, v_{i,N})$. If these are normalized so that $\boldsymbol{u}_i^\top \boldsymbol{v}_i = \boldsymbol{u}_i^\top \boldsymbol{1} = 1$, then*

$$(\boldsymbol{M}_i)^c = \gamma_{\max,i}{}^c \boldsymbol{Q}_i + (\boldsymbol{R}_i)^c, \tag{17}$$

where $\boldsymbol{Q}_i = \boldsymbol{u}_i \boldsymbol{v}_i^\top$, $\boldsymbol{Q}_i \boldsymbol{R}_i = \boldsymbol{R}_i \boldsymbol{Q}_i = \boldsymbol{0}$, *and there exist* $K < \infty$ *and* γ $(0 < \gamma < \gamma_{\max,i})$ *such that* $\|\boldsymbol{R}_i{}^c\| \leq K\gamma^c$.

Proof. See Appendix.

Lemma 1 decomposes into two parts. The first part of $\boldsymbol{M}_i$, consisting of the maximal eigenvalue (and the left and right eigenvectors corresponding to this eigenvalue), dominates the second part for higher powers of $\boldsymbol{M}_i$, i.e. $(\boldsymbol{M}_i)^c$ for large c. Note that (13) implies that these higher powers of $\boldsymbol{M}_i$ determine the tail behaviour of the sequences of variables $\alpha_{(i,c),k}$.

4.1.2. *Convergence.* We will now show that the sequence $\{\alpha_{(i,c),k}, c = 0, 1, \ldots\}$ converges for all $i, k = 1, \ldots, N$. To this end, the following property is essential.

Lemma 2 (Maximal eigenvalues.) *(1)* $\gamma_{\max,i} < 1, i = 1, \ldots, N$, *if and only if* $\rho < 1$*;*
(2) $\gamma_{\max,i} = 1, i = 1, \ldots, N$, *if and only if* $\rho = 1$.

Proof. See Appendix.

Lemma 3 (Convergence.) *For* $i, k = 1, \ldots, N$*:*
(1) for $\rho < 1$*, the sequence* $\{\alpha_{(i,c),k}, c = 0, 1, \ldots\}$ *converges to 0;*
(2) for $\rho = 1$*, the sequence* $\{\alpha_{(i,c),k}, c = 0, 1, \ldots\}$ *converges to a finite value.*

Proof. From (13) and Lemma 1 it follows that we can express $\alpha_{(i,c),k}$ as follows:

$$\alpha_{(i,c),k} = \boldsymbol{e}_k^\top (\boldsymbol{M}_i)^c \boldsymbol{e}_i = \gamma_{\max,i}^c u_{i,k} v_{i,i} + r_{i,k}^{(c)}, \quad \text{for } i, k = 1, \ldots, N, \quad c = 0, 1, \ldots, \tag{18}$$

with $r_{i,k}^{(c)} < K\gamma^c$ for some $K < \infty$ and $0 < \gamma < \gamma_{\max,i}$. The results follow then directly from Lemma 2.

In the theory of branching processes, the case $\rho < 1$ is referred to as the 'subcritical case', and it is known that in this case all the mean numbers of the descendants of a customer at successive generations tend to 'die out' on the long run. The case $\rho = 1$ is commonly referred to as the 'critical case', and it is known that the mean number of descendants of a customer at successive generations tends to some constant (cf. [1]).

Throughout the following, the limiting values of the sequences $\{\alpha_{(i,c),k}, c = 0, 1, \ldots\}$ are denoted by

$$\alpha_{(i,\infty),k} := \lim_{c \to \infty} \alpha_{(i,c),k}, \quad i, k = 1, \ldots, N. \tag{19}$$

4.1.3. *Limiting values.* Recall that for $\rho < 1$ the eigenvalues and eigenvectors of $\boldsymbol{M}_i$ are generally unknown, so that the values of $\alpha_{(i,c),k}$ cannot be solved explicitly. However, the following lemma states that for the analysis of the model in heavy traffic, we can restrict ourselves to the case $\rho = 1$.

Lemma 4 (Simplification.) *For* $i, k, l = 1, \ldots, N$,

$$\frac{\tilde{\psi}_{i,k}}{\tilde{\psi}_{i,l}} = \left(\frac{\alpha_{(i,\infty),k}}{\alpha_{(i,\infty),l}} \right)^2, \tag{20}$$

where $\alpha_{(i,\infty),k}$ *and* $\alpha_{(i,\infty),l}$ *are both evaluated at* $\rho = 1$.

Proof. See Appendix.

Lemma 4 implies that for the analysis of the sequences $\{\alpha_{(i,c),k}, c = 0, 1, \ldots\}$, in the limiting case $\rho \uparrow 1$, we do not have to consider the behaviour of the system for arbitrary $\rho < 1$ and then let ρ tend to 1. Instead, we can restrict ourselves to the analysis of the sequences $\{\alpha_{(i,c),k}, c = 0, 1, \ldots\}$ for the case $\rho = 1$.

In order to derive closed-form expressions for the ω_k, we will first derive closed-form expressions for the ratios between ω_k and ω_l ($l \neq k$). To this end, the ratios between the variables $\tilde{\psi}_{i,k}$ and $\tilde{\psi}_{i,l}$ have to be determined (see (14) and (15)). Lemma 4 states that the ratios between the variables $\tilde{\psi}_{i,k}$ and $\tilde{\psi}_{i,l}$ can be simply derived by obtaining the ratios between the limiting values of the sequences $\{\alpha_{(i,c),k}, c = 0, 1, \ldots\}$, and $\alpha_{(i,c),l}, c = 0, 1, \ldots\}$, for $\rho = 1$. The following theorem expresses the ratios between these limiting values for $\rho = 1$.

Theorem 2 (Ratios between limiting values.) *For $i, k, l = 1, \ldots, N$, and $\rho = 1$,*

$$\frac{\alpha_{(i,\infty),k}}{\alpha_{(i,\infty),l}} = \frac{\lambda_k}{\lambda_l} \frac{(1 - \rho_k I_{\{k \in E\}})}{(1 - \rho_l I_{\{l \in E\}})}; \tag{21}$$

but for $k, l, m = 1, \ldots, N$, and $\rho = 1$,

$$\frac{\alpha_{(l,\infty),k}}{\alpha_{(m,\infty),k}} = \frac{b_l}{b_m}. \tag{22}$$

Proof. See Appendix.

Note the striking difference between (21) and (22), where the former depends on the service policy and the latter does not. In explaining (21), consider service at all queues, and consider a tagged type-i customer T. It is then obvious that the numbers of descendants of T at the various queues in the infinite future are proportional to the arrival rates. The sensitivity of (21) to the service discipline can be explained as follows. If $k \in E$, then only a fraction $1 - \rho_k$ of the type-k customers that are served in the system are actually present at some polling instant of Q_k. This is because each (single) customer present at a polling instant at Q_k generates an $M/G/1$ busy period during which, on average, $1/(1 - \rho_k)$ customers are served. This explains why, in the case of exhaustive service, the factor $(1 - \rho_k)$ occurs in (21). In order to give an intuitive explanation for (22), consider a tagged type-l customer T. On average, T has $\lambda_i b_l$ type-i children, $i = 1, \ldots, N$, regardless of the service discipline at Q_l. (Note that it does not matter whether its children at Q_l are served during the same visit, as in the exhaustive case; or in future visits, as in the gated case.) Each of these children, in turn, will have on average $\alpha_{(i,\infty),k}$ type-k descendants at some polling instants of Q_k in the infinite future. Hence, the mean number of descendants of T is proportional to b_l.

4.2. Derivation of the main result

We are now ready to derive the main result (Theorem 1). The derivation proceeds along two steps. First, we derive simple expressions for the ratios between the scaled expected waiting times. Second, we determine the unknown scaling factor.

Theorem 3 (Ratios between scaled expected waiting times.) *For $k, l = 1, \ldots, N$,*

$$\frac{\omega_k}{\omega_l} = \frac{(1 + \rho_k) I_{\{k \in G\}} + (1 - \rho_k) I_{\{k \in E\}}}{(1 + \rho_l) I_{\{l \in G\}} + (1 - \rho_l) I_{\{l \in E\}}}. \tag{23}$$

Proof. The result is obtained by combining (20), (21), and substituting the resulting ratios into (14) and (15).

Bsed on Theorem 3, the scaled expected waiting times are known up to some unknown scaling factor. This scaling factor can easily be obtained by using the pseudo-conservation law for the model under consideration (see [4]): for $\rho < 1$,

$$\sum_{i=1}^{N} \rho_i E[W_i] = \frac{\rho}{1-\rho}\frac{b^{(2)}}{2b} + \rho\frac{r^{(2)}}{2r} + \frac{r}{2(1-\rho)}\left[\rho^2 - \sum_{i\in E}\rho_i^2 + \sum_{i\in G}\rho_i^2\right]. \tag{24}$$

Multiplying both sides by $(1-\rho)$ and letting $\rho \uparrow 1$ yields the following relation,

$$\sum_{i=1}^{N} \rho_i \omega_i = \frac{b^{(2)}}{2b} + \frac{r}{2}\left[1 - \sum_{i\in E}\rho_i^2 + \sum_{i\in G}\rho_i^2\right]. \tag{25}$$

We are now ready to obtain the main result (Theorem 1), i.e.

$$\omega_i = \frac{1+\rho_i}{\sum_{j\in G}\rho_j(1+\rho_j) + \sum_{j\in E}\rho_j(1-\rho_j)}\frac{b^{(2)}}{2b} + \tfrac{1}{2}r(1+\rho_i), \quad i \in G, \tag{26}$$

$$\omega_i = \frac{1-\rho_i}{\sum_{j\in G}\rho_j(1+\rho_j) + \sum_{j\in E}\rho_j(1-\rho_j)}\frac{b^{(2)}}{2b} + \tfrac{1}{2}r(1-\rho_i), \quad i \in E. \tag{27}$$

Proof. The result follows directly from Theorem 3 and (25).

For the special case of exhaustive service at all queues (i.e. $G = \emptyset$, $E = \{1, \ldots, N\}$) the same results have been obtained by Blanc and Van der Mei [3] by using the buffer-occupancy method [20], and by Coffman et al. [6] by using diffusion approximations.

The following properties follow directly from (26) and (27) and provide new insights into the behaviour of polling systems in heavy traffic.

Corollary 1. *For $k = 1, \ldots, N$;*
(1) ω_k depends on the service-time distributions only through b and $b^{(2)}$, i.e. the first two moments of the service time of an arbitrary customer;
(2) ω_k depends on the switch-over time distributions only through r, i.e. the first moment of the total switch-over time per cycle of the server along the queues;
(3) if the service discipline at Q_k is modified from gated to exhaustive service, then ω_k decreases, while ω_j increases for all $j \neq k$.

Note that the properties stated in Corollary 1 are not generally valid for the expected waiting times in stable systems (i.e. $\rho < 1$).

5. Approximation

Theorem 1 implies that under heavy load the expected waiting times can be approximated by (3). That is equivalent to approximating the ratios between the expected waiting times as follows:

$$\frac{E[W_i]}{E[W_j]} \approx \frac{(1+\rho_i)I_{\{i\in G\}} + (1-\rho_i)I_{\{i\in E\}}}{(1+\rho_j)I_{\{j\in G\}} + (1-\rho_j)I_{\{j\in E\}}}. \tag{28}$$

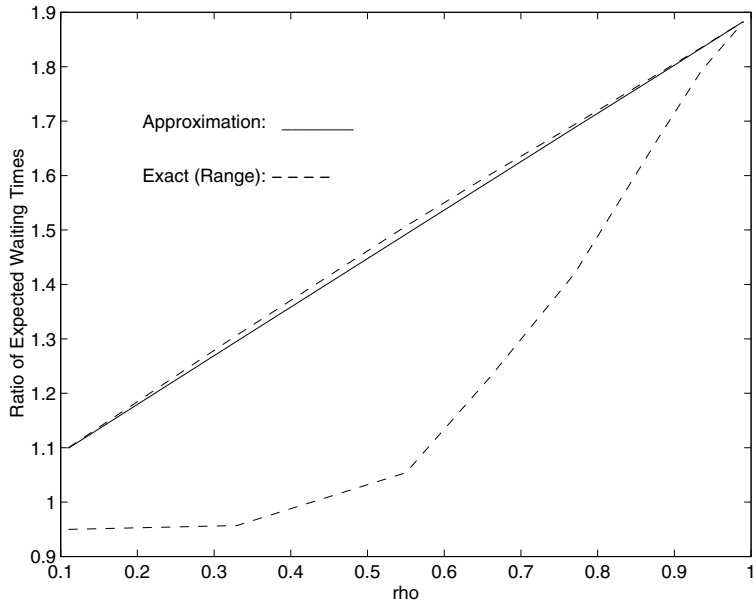

FIGURE 4: Quality of approximation as function of load.

The approximation in (3) was previously proposed by Groenendijk [8] for *arbitrary loads*. Our results suggest that the approximation is *indeed* good for *heavy traffic*.

One may question the practicality of heavy traffic, in the sense of 'how heavy should the traffic be' to make the approximation work well. Another interesting question is how the quality of the approximation varies as function of the load.

In order to demonstrate these issues, we consider one example in which we evaluate the quality of the approximation in (28) as function of the system load. We consider a system consisting of 11 queues, in which the service times in all queues are identical (exponential with mean 1) while their arrival rates vary, i.e. $\lambda_i = 0.001x$, $i = 1, \ldots, 10$ and $\lambda_{11} = 0.1x$ where x varies between 1 and 9. Thus, for every traffic load ρ, we have $\rho_i/\rho = 0.009$, $i = 1, \ldots, 10$ and $\rho_{11}/\rho = 0.909$. The switch-over periods are all of mean 1 and second moment 4, except for the sixth switch-over period which has mean 200 and second moment 180000. The system is therefore, very asymmetric in its arrival rates as well as in the switch-over periods. The service policy of all queues is gated. We use numerical procedures to evaluate the expected waiting times in all queues and to compute the exact ratio $E[W_{11}]/E[W_i]$, $i = 1, \ldots, 10$. We then compare it to (28). In Figure 4 we provide this comparison where the solid line depicts (28) and the dashed lines depict the upper and lower values obtained for this ratio in the exact analysis (note that in practice the expected waiting times of the first 10 queues differ from each other, thus yielding a range of ratios). These results are depicted as function of the total system load.

The figure demonstrates that for heavy traffic this approximation is indeed excellent. For $\rho = 0.93$ and above, the relative error in estimating the ratio is smaller than 2.3% and for $\rho = 0.88$ the error is smaller than 6.8%. The implication is that for most practical cases

the approximation can be used with confidence. This implication stems from the fact that, in practice, heavy load is the main region of interest. The figure also demonstrates that at lower load values the approximation quality degrades, reaching a 48% relative error at $\rho = 0.55$. We may conclude that the approximation is especially good for heavy traffic.

We have examined several other cases, all showing excellent results at heavy traffic. It should be noted that for many of the other cases (when the system is not very asymmetric) the approximation is significantly better than in the above example in both the heavy load region and the non-heavy load region.

6. Optimization

We now consider the general problem of minimizing an arbitrary weighted sum of the expected waiting times with respect to the disciplines at each of the queues. The optimization problem can be formulated as follows,

$$\min_{(G,E)} c(G, E) = \sum_{i=1}^{N} c_i \omega_i, \tag{29}$$

where the weights c_i are strictly positive. Thus, the problem is to find a partitioning (G^*, E^*) of $\{1, \ldots, N\}$ that minimizes $c(G, E)$ over all partitionings of $\{1, \ldots, N\}$. The solution of (29) is believed to be a good estimate for the optimal static assignment of service disciplines to the queues in heavily loaded systems. Without loss of generality, it is assumed that $c_1/\rho_1 \geq \cdots \geq c_N/\rho_N$, and that $\sum_{i=1}^{N} c_i = 1$ and $\sum_{i=1}^{N} \rho_i = 1$.

6.1. Zero switch-over times

Using (26) and (27), one may verify the following (we refer to [23] for details on the proof).

Corollary 2. *The optimal assignment* (G^*, E^*) *has the structure*

$$E^* = \{1, \ldots, K\}, \quad G^* = \{K + 1, \ldots, N\}, \quad \textit{for some } K (1 \leq K \leq N - 1). \tag{30}$$

Thus, the queues with the K highest c_i/ρ_i ratios should be served exhaustively, while the queues with the $N - K$ lowest c_i/ρ_i ratios should be served according to the gated policy. The optimal value of K can be determined according to the following numerical procedure. Starting with $G = \{1, \ldots, N\}$, $E = \emptyset$, for increasing values of k, the procedure successively checks if the transition of k from G to E leads to a decrease in the cost function. If so, k is removed from G and added to E; otherwise the optimum is found. Corollary 2 guarantees that this numerical procedure leads to an optimal assignment of the service disciplines.

$E := \emptyset; \quad x := \sum_{i=1}^{N} c_i \rho_i; \quad y := \sum_{i=1}^{N} \rho_i^2; \quad k := 1;$

repeat

$x_{\text{temp}} := x - 2c_k \rho_k; \quad y_{\text{temp}} := y - 2\rho_k^2;$

if $x_{\text{temp}}/y_{\text{temp}} < x/y$ do

begin $E := E \cup \{k\}; \quad x := x_{\text{temp}}; \quad y := y_{\text{temp}}; \quad k := k + 1;$ end;

until $x_{\text{temp}}/y_{\text{temp}} \geq x/y;$

$E^* := E; \quad G^* := \{1, \ldots, N\} - E^*.$

From Theorem 4, one may verify that all queues for which $c_i/\rho_i = \max_{j=1}^{N} \{c_j/\rho_j\}$ should be served exhaustively. Moreover, similar arguments imply that all queues for which $c_i/\rho_i = \min_{j=1}^{N} \{c_j/\rho_j\}$, should be served according to the gated policy. These observations indicate that both G^* and E^* are generally non-empty. The only exception is made for the case $c_i = \rho_i$ for all $i = 1, \ldots, N$. In that case, it follows from (25) that for $r = 0$, the cost function is independent of the service disciplines at the queues, so that every assignment (G, E) is optimal.

6.2. Non-zero switch-over times

For the case of non-zero switch-over times, the optimal structure of the partioning problem does not generally satisfy (30). This observation is caused by the second term in (26) and (27). The following property gives a partial solution to the optimization problem (see [23] for details on the proof).

Corollary 3. *If* $c_i/\rho_i = \max_j \{c_j/\rho_j\}$, *then* $i \in E^*$.

Thus, all queues with the highest c_i/ρ_i ratio should be served exhaustively.

In the special case $c_i = \rho_i$ for all i, Corollary 3 states that all queues should be served exhaustively. Note that in this case the cost function (29) corresponds to the left-hand side of (25), and can be interpreted as the scaled expected mean total amount of waiting work in the system. From (25) it is directly seen that it is shown that if $r > 0$, then all queues should be served exhaustively, i.e. $(G^*, E^*) = (\emptyset, \{1, \ldots, N\})$.

The problem of determining the service discipline for the queues which are not covered by (**??**) is not trivial. From (26) and (27), it is obvious that for r large enough, all queues should be served exhaustively. However, it may well happen that $i \in E_r^*$, $i \notin E_{\tilde{r}}^*$ and $i \in E_{\hat{r}}^*$ for some $\hat{r} > \tilde{r} > r$. For instance, consider a 3-queue model with $c_1 = 0.4, c_2 = 0.01, c_3 = 0.59$, $\rho_1 = 0.2, \rho_2 = 0.01, \rho_3 = 0.79$, and $b_i = b_i^{(2)} = 1, i = 1, 2, 3$. Then one may verify that for $r < 1.022$, $E^* = \{1, 2\}$, while for $1.022 < r < 2.091$, $E^* = \{1, 3\}$ and for $r > 2.091$, $E^* = \{1, 2, 3\}$. This examples illustrates that for increasing values of r, all kinds of possible 'swaps' of E^* should be taken into account. The problem seems to be a complicated combinatorical optimization problem, and remains open in this context.

7. Topics for further research

The model allows only exhaustive and gated service at each of the queues. Whether a similar analysis can be performed to derive closed-form expressions similar to (26) and (27) for models with other service disciplines is an open question. In fact, further analysis has shown that the results in the present paper can be generalized to a more general class of service disciplines satisfying a certain branching property [18]. The results will be presented in a forthcoming paper [24].

The DSA allows for the computation of higher moments of the (scaled) delay at the queues. In fact, for the case of zero switch-over times, closed-form expressions for the (scaled) moments of the delay at each of the queues are presented in [25].

The monotonicity property stated in Corollary 1 (Item 3) seems to be intuitively clear, and provides insight into the behaviour of the model in heavy traffic. A further question is whether a similar monotonicity property also holds for stable systems (i.e. with $\rho < 1$). Such qualitative results would contribute strongly to the understanding of the stochastic behaviour of polling systems. Analysis of the system for $\rho < 1$, however, would not only require properties of the tail behaviour, but also properties of the heading part of the sequences $\{\alpha_{(i,c),k}, c = 0, 1, \ldots\}$.

Appendix

Proof of Lemma 1. The proof is based on the so-called Frobenius Theorem (FT) for strictly positive matrices. A square matrix $\boldsymbol{A}$ with real-valued entries is called strictly positive if there exists n such that all entries of the matrix $\boldsymbol{A}^n$ are strictly positive. For strictly positive matrices, the FT guarantees the existence of a unique 'maximal eigenvalue', i.e. a real eigenvalue of multiplicity 1 and with strictly positive corresponding left and right eigenvectors. To apply the FT to $\boldsymbol{M}_i$, one may verify that $\boldsymbol{M}_i$ is strictly positive, except for the case where Q_{i+1} is served exhaustively. In that case, namely, the $(i+1)$th column of $\boldsymbol{M}_i$, and hence of $(\boldsymbol{M}_i)^c$, consists of zeros for all c. In that case, consider the $(N-1)\times(N-1)$ submatrix $\tilde{\boldsymbol{M}}_i$, which is obtained from $\boldsymbol{M}_i$ by omitting the $(i+1)$th row and column. The characteristic equation of $\boldsymbol{M}_i$ can be written as $0 = \det(\boldsymbol{M}_i - x\boldsymbol{I}) = x\det(\tilde{\boldsymbol{M}}_i - x\boldsymbol{I}_{N-1})$ (where $\boldsymbol{I}_{N-1}$ is the $(N-1)$-dimensional identity matrix). This implies that if γ_i is an eigenvalue of $\tilde{\boldsymbol{M}}_i$ with corresponding right eigenvector $\boldsymbol{w}_i = (w_{i,1}, \ldots, w_{i,N-1})$, then it is readily seen that $(w_{i,1}, \ldots, w_{i,i}, 0, w_{i,i+2}, \ldots, w_{i,N-1})$ is an eigenvector of $\boldsymbol{M}_i$ corresponding to eigenvalue γ_i, $i = 1, \ldots, N-1$. Moreover, it is readily seen that $\boldsymbol{M}_i$ has an additional eigenvalue 0, with corresponding eigenvector $\boldsymbol{e}_{i+1}$. Hence, $\tilde{\boldsymbol{M}}_i$ has the same 'maximal eigenvalue' as $\boldsymbol{M}_i$. The decomposition (13) can then be obtained along the same lines as in [10]. Note that a similar result for the matrix $\boldsymbol{N}_k$ (see end of Section 3) can be obtained by following similar arguments.

Proof of Lemma 2. For the matrix $\boldsymbol{N}_k$, similar properties of the maximal eigenvalue have been obtained in [18]. Following the same lines, but with λ^*, defined by the vector whose ith element is defined by $\lambda_i(1-\rho_i I_{\{i\in E\}})$, $i = 1, \ldots, N$, instead of $\boldsymbol{b}$, leads to the result.

Proof of Lemma 4. From the definition of $\tilde{\psi}_{i,k}$ (see (16)), using (9) and by some straightforward manipulations it follows that

$$\frac{\tilde{\psi}_{i,k}}{\tilde{\psi}_{i,l}} = \lim_{\rho\uparrow 1}\lim_{n\to\infty}\frac{n^{-1}\sum_{c=0}^{n}\alpha_{(i,c),k}{}^2}{n^{-1}\sum_{c=0}^{n}\alpha_{(i,c),l}{}^2} = \lim_{\rho\uparrow 1}\lim_{n\to\infty}\frac{n^{-1}\sum_{c=0}^{n}(\gamma_{\max,i}^{c}u_{i,k}v_{i,i}+r_{i,k}^{(c)})^2}{n^{-1}\sum_{c=0}^{n}(\gamma_{\max,i}^{c}u_{i,l}v_{i,i}+r_{i,l}^{(c)})^2}. \tag{31}$$

A number of straightforward algebraic manipulations can be used to determine the limit for $n\to\infty$ in (31). Using this, and denoting $u_{i,k}$ and $v_{i,i}$ as $u_{i,k}(\rho)$ and $v_{i,i}(\rho)$ (to emphasize the dependence of $\boldsymbol{u}$ and $\boldsymbol{v}$ on ρ), the right-hand side of (31) can be rewritten as

$$\lim_{\rho\uparrow 1}\left(\frac{u_{i,k}(\rho)}{u_{i,l}(\rho)}\right)^2 = \left(\frac{u_{i,k}(1)}{u_{i,l}(1)}\right)^2, \tag{32}$$

where the equality in (32) is based on the continuity of the eigenvectors of a matrix with respect to its entries (cf. [11]). The proof of Lemma 4 is completed by the observation that it follows from (18) that for $\rho = 1$ we have $\alpha_{(i,\infty),k} = u_{i,k}(1)v_{i,i}(1)$ and $\alpha_{(i,\infty),l} = u_{i,l}(1)v_{i,i}(1)$, for $k, l = 1, \ldots, N$.

Proof of Theorem 2. To prove equation (21), define $\lambda^* = (\lambda_1^*, \ldots, \lambda_N^*)$, with $\lambda_i^* = \lambda_i(1-\rho_i I_{\{i\in E\}})$, $i = 1, \ldots, N$. Then it is sufficient to show that both the vectors $\alpha_{(i,\infty),\cdot}$ and λ^* are right eigenvectors of $\boldsymbol{M}_i$, corresponding to $\gamma_{\max,i} = 1$ at $\rho = 1$ (see Lemma 2). To this end, assume $\rho = 1$. By definition, we have $\boldsymbol{M}_i\alpha_{(i,\infty),\cdot} = \alpha_{(i,\infty),\cdot}$, $i = 1, \ldots, N$. Hence, $\alpha_{(i,\infty),\cdot}$ is an eigenvector of $\boldsymbol{M}_i$ corresponding to eigenvalue 1. Further, we have

$$\lambda_k\sum_{j=1}^{N}\lambda_j^* d_{j,k} = \lambda_k\sum_{j\in G}\lambda_j b_j + \sum_{j\in E, j\neq k}\lambda_j(1-\rho_j)\frac{b_j}{1-\rho_j} = \lambda_k\sum_{j=1}^{N}\lambda_j b_j(I_{\{j\in E\}}I_{\{j\neq k\}}+I_{\{k\in G\}}),$$

which is equal to λ_k if $k \in G$ and $\lambda_k(1-\rho_k)$ if $k \in E$ (cf. (12)). This implies $\hat{\boldsymbol{M}}_k\lambda^* = \lambda^*$, $k = 1, \ldots, N$ and hence, $\boldsymbol{M}_i\lambda^* = \lambda^*$, $i = 1, \ldots, N$. This implies that λ^* is an eigenvector of $\boldsymbol{M}_i$ corresponding to eigenvalue 1.

Acknowledgement

This work was done while the authors were with RUTCOR, Rutgers University, New Brunswick, NJ. RDM was supported by the NATO Science Fellowship (grant N61-329).

References

[1] Athreya, K. B and Ney, P. E. (1971). *Branching Processes*. Springer, Berlin.
[2] Blanc, J. P. C. (1992). Performance evaluation of polling systems by means of the power-series algorithm. *Ann. Operat. Res.* **35**, 155–186.
[3] Blanc, J. P. C. and Van der Mei, R. D. (1995). Optimization of polling systems with Bernoulli schedules. *Perf. Eval.* **22**, 139–158.
[4] Boxma, O. J. and Groenendijk, W. P. (1987). Pseudo-conservation laws in cyclic service systems. *J. Appl. Prob.* **24**, 949–964.
[5] Choudhury, G. and Whitt, W. (1994). Computing transient and steady state distributions in polling models by numerical transform inversion. *Perf. Eval.* **25**, 267–292.
[6] 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.
[7] Coffman, E. G., Puhalskii, A. A. and Reiman, M. I. (1995). Polling systems in heavy-traffic: a Bessel process limit. To appear in *Math. Oper. Res.*
[8] 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, pp. 1434–1441.
[9] Fricker, C. and Jaïbi, M. R. (1994). Monotonicity and stability of periodic polling models. *Queueing Systems* **15**, 211–238.
[10] Karlin, S. (1966). *A First Course in Stochastic Processes*. Academic Press, New York.
[11] Kato, T. (1966). *Perturbation Theory for Linear Operators*. Springer, New York.
[12] Konheim, A. G., Levy, H. and Srinivasan, M. M. (1994). Descendant set: an efficient approach for the analysis of polling systems. *IEEE Trans. Comm.* **42**, 1245–1253.
[13] Leung, K. K. (1991). Cyclic service systems with probabilistically-limited service. *IEEE J. Sel. Areas Commun.* **9**, 185–193.
[14] Levy, H. and Sidi, M. (1991). Polling models: applications, modeling and optimization. *IEEE Trans. Commun.* **38**, 1750–1760.
[15] Levy, H., Sidi, M. and Boxma, O. J. (1990). Dominance relations in polling systems. *Queueing systems* **6**, 155–171.
[16] Markowitz, D. (1995). *Dynamic Scheduling of Single-Server Queues with Setups: A Heavy Traffic Approach.* Ph.D. Thesis, Operations research Center, MIT, Cambridge, MA.
[17] Reiman, M. I. and Wein, L. M. (1998). Dynamic scheduling of a two-class queue with setups. To appear in *Math. Oper. Res.*
[18] Resing, J. A. C. (1993). Polling systems and multitype branching processes. *Queueing Systems* **13**, 409–426.
[19] Srinivasan, M. M., Levy, H. and Konheim, A. G. (1996). The individual station technique for the analysis of cyclic polling systems. *Naval Research Logistics* **73**, 79–101.
[20] Takagi, H. (1986). *Analysis of Polling Systems*. The 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, pp. 267–318.
[22] Takagi, H. (1994). Queueing analysis of polling models: progress in 1990–1993. In *Frontiers in Queueing: Models, Methods and Problems*, ed. J. H. Dshalalow. CRC Press.
[23] Van der Mei, R. D. and Levy, H. (1996). Expected delay analysis of polling systems in heavy traffic. Technical Report RRR 17–96, RUTCOR, Rutgers University.
[24] Van der Mei, R. D. and Levy, H. (1997). Polling systems in heavy traffic: exhaustiveness of the service policies. *Queueing Systems and their applications* **27**, 227–250.
[25] Van der Mei, R. D. (1997). Polling systems in heavy traffic: higher moments of the delay. In *Teletraffic Contributions for the Information Age*, eds. V. Ramaswamy and P. E. Wirth. Elsevier, Amsterdam, pp. 275–284.