Polling systems in heavy traffic: Exhaustiveness of service policies *

R.D. van der Mei [a] and H. Levy [b]

[a] *AT&T Labs, P.O. Box 3030, Holmdel, NJ 07733, USA*
[b] *Department of Computer Science, Tel-Aviv University, Tel-Aviv, Israel*

Received 12 March 1997; revised 7 October 1997

We study the expected delay in cyclic polling models with general 'branching-type' service disciplines. For this class of models, which contains models with exhaustive and gated service as special cases, we obtain closed-form expressions for the expected delay under standard heavy-traffic scalings. We identify a single parameter associated with the service discipline at each queue, which we call the 'exhaustiveness'. We show that the scaled expected delay figures depend on the service policies at the queues only through the exhaustiveness of each of the service disciplines. This implies that the influence of different service disciplines, but with the same exhaustiveness, on the expected delays at the queues becomes the same when the system reaches saturation. This observation leads to a new classification of the service disciplines. In addition, we show monotonicity of the scaled expected delays with respect to the exhaustiveness of the service disciplines. This induces a complete ordering in terms of efficiency of the service disciplines. The results also lead to new rules for optimization of the system performance with respect to the service disciplines at the queues. Further, the exact asymptotic results suggest simple expected waiting-time approximations for polling models in heavy traffic. Numerical experiments show that the accuracy of the approximations is excellent for practical heavy-traffic scenarios.

Keywords: polling systems, heavy traffic, expected delay, exhaustiveness, monotonicity, service disciplines, classification

1. Introduction

The basic polling system consists of a number of queues and a single server that visits the queues in cyclic order to render service to the customers waiting at 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 [15] for an overview of the applicability of polling models, and to [25] for a review of the state-of-the-art in the

* The work of the first author was supported by the NATO Science Fellowship (grant N61-329). The work was done while the authors were with RUTCOR, Rutgers University, New Brunswick, NJ 08903, USA.

analysis of polling models. Some variations of polling models do not allow for an exact detailed analysis, and the others usually require the use of numerical techniques to determine performance measures of interest.

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 asymmetrical systems with a large number of queues. Moreover, numerical techniques can only contribute to the understanding of the behavior of the system to a limited extent. Exact closed-form expressions provide much more insight into the dependence of the performance measures with respect to the system parameters. These observations raise the importance of an exact asymptotic analysis of the performance of polling models in heavy traffic.

We will show that a general class of polling models allows for an exact analysis under heavy traffic assumptions. The results generalize those obtained in [20] for the special case of exhaustive and gated service at each of the queues. The analysis, however, requires several substantial extensions of the analysis in [20]. Moreover, the obtained results are much more general and provide new insights into the behavior of polling systems in heavy traffic.

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 [23], 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 several numerical algorithms have been proposed to determine the moments of the delay at the queues by solving sets of linear equations (cf., e.g., [24] for references). Recently, the efficiency of the numerical techniques has been considerably improved by the so-called descendant set approach (DSA). The DSA is an iterative technique which explores the MTBP-structure of the model by making use of the concept of so-called descendant sets (cf. [13]). Choudhury and Whitt [7] use numerical transform-inversion to extend the DSA for the determination of tail probabilities of the waiting times. The key element in the identification of the class of polling models in [23] is that the service policies at each of the queues should satisfy a certain 'branching property'. This property is satisfied by the classical exhaustive and gated service policies, but also by more flexible fractional service disciplines like binomial-gated [18], fractional-exhaustive [17], binomial-exhaustive [6] and Bernoulli-type [23]. Polling models with service disciplines that do not have a branching structure (e.g., limited-type service disciplines), are usually not exactly analyzable and generally require much more computational effort to obtain performance measures such as moments of the delay at each of the queues (cf. [2,14]).

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. [8]. 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, they prove that similar properties hold for systems with more than two queues. Moreover, in [9] they show 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 [8,9] also hold for non-exhaustive policies, Reiman and Wein [22] 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 [19] extends the results in [22] to the multi-class case.

We consider an asymmetrical polling model with general service disciplines that satisfy the branching property [23]. We study the scaled expected delay at each of the queues in heavy-traffic. All the queues become instable when the system load (denoted by ρ) approaches 1 (cf. [11]). More precisely, the expected delay (considered as function of ρ) at each of the queues possesses a first-order pole at $\rho = 1$. Therefore, the main performance measure of interest will be the scaled expected delay, i.e., the limit of $(1-\rho)$ 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. A key role in the derivation is played by a new view on the concept of descendant sets [20]. Based on the obtained expressions, we identify a *single* parameter associated with the service discipline at each queue, referred to as the exhaustiveness of the service policy at that queue. We show that the scaled expected delay at each of the queues depends on the service disciplines at the queues only through the exhaustiveness of the service disciplines at each of the queues. In other words, we show that the influence of *different* service policies, but with the same exhaustiveness, becomes the *same* when the system reaches saturation. This observation leads to a new classification of the service disciplines. Further, we derive *monotonicity* of the scaled expected delays: if queue i is served more exhaustively, then the scaled expected delay at queue i decreases, whereas the scaled expected delays at all other queues increase. The obtained monotonicity leads to a *complete ordering* in terms of efficiency of the general class of service disciplines under consideration. We also obtain new results on *optimization* of the system performance. It is shown that, in order to minimize an arbitrary weighted sum of the (scaled) expected waiting times with respect to the service disciplines at the queues, all queues with the highest weight/load ratio should be served exhaustively. In addition, the obtained expressions for the scaled expected delay suggest very simple *approximations* for the expected waiting times in stable systems. Numerical results are presented to show that these approximations are very accurate for practical heavy-traffic scenarios, for loads typically exceeding 80%.

The remainder of the paper is organized as follows. In section 2 the model is described in detail. In section 3 we review the principles of the DSA and discuss some preliminary results. In section 4 these results are used to derive closed-form expressions for the scaled expected waiting times. In section 5 we discuss the implications of the expressions obtained. In section 6 numerical examples are presented to illustrate the results. In section 7 we propose and test simple and fast-to-evaluate approximations for the expected delays at the queues in stable systems. Finally, in section 8 we address a number of topics for further research.

2. Model description and the basic result

We consider a system consisting of N infinite-buffer queues $Q_1, \ldots, Q_N$ and a single server. 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 $\Lambda = \sum_{i=1}^{N} \lambda_i$. At each of the queues the customers are served on a first-come-first-served basis. The service time of a type-i customer is a random variable B_i, with Laplace–Stieltjes transform (LST) $\beta_i(\cdot)$ and finite first and second moments b_i and $b_i^{(2)}$. Denote $\mathbf{b} = (b_1, \ldots, b_N)$. The first two moments of an arbitrary service time are denoted by

$$\beta_1 = \sum_{i=1}^{N} \frac{\lambda_i b_i}{\Lambda} \quad \text{and} \quad \beta_2 = \sum_{i=1}^{N} \frac{\lambda_i b_i^{(2)}}{\Lambda},$$

respectively. The load offered to Q_i is $\rho_i = \lambda_i b_i$, and the total offered load is equal to $\rho = \sum_{i=1}^{N} \rho_i$.

After completing service at Q_i the server proceeds to Q_{i+1}, typically incurring a *switch-over period* whose duration is an independent random variable R_i. The first two moments of R_i are finite and are denoted by r_i and $r_i^{(2)}$. Denote the first moment of the total switch-over time per cycle of the server along the queues 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 that $r > 0$.

The moments at which the server arrives at Q_i are referred to as the *polling instants* at Q_i. The periods during which the server is working at Q_i are called *service periods* at Q_i. The moments at which the server departs from Q_i are referred to as *departure instants* from Q_i.

The service disciplines at the queues are assumed to satisfy the following property (cf. [10,23]):

Branching property. If the server arrives at Q_i to find k_i customers present there, then during the course of the server's visit, each of these k_i customers will be effectively

replaced in an i.i.d. manner by a random population having probability generating function (pgf) $h_i(s_1, \ldots, s_N)$.

In the general context of [23], the pgf $h_i(s_1, \ldots, s_N)$ may be any N-dimensional pgf. For ease of the interpretation, we will only consider service disciplines which are *non-idling*, i.e., the server is not allowed to idle (rest) while visiting a queue. The visit period at Q_i starting with k_i customers C_j ($j = 1, \ldots, k_i$) present at Q_i is considered to consist of k_i i.i.d. sub-busy periods, each characterized by joint pgf-LST

$$\psi_i(u, v) = E\left[\mathrm{e}^{-uT_i} v^{L_i}\right], \tag{1}$$

where T_i is the length of the sub-busy period and where L_i is the so-called sub-busy period residue. To define the latter, consider a sub-busy period at Q_i initiated by a type-i customer C_j ($j = 1, \ldots, k_i$). If C_j is served during the current service period, L_i is the number of type-i children of C_j residing in Q_i at the end of the sub-busy period (for a precise definition of the notion of 'children', see section 3.1); otherwise, $L_i = 1$ (note that in this case the length of the sub-busy period equals 0). The assumption of non-idling service implies the following relation between $h_i(\cdot)$ and $\psi_i(\cdot, \cdot)$ (cf. [23]):

$$h_i(s_1, \ldots, s_N) = \psi_i\left(\sum_{j \neq i} \lambda_j(1 - s_j), s_i\right) \quad (i = 1, \ldots, N). \tag{2}$$

For instance, in the case of gated service,

$$h_i(s_1, \ldots, s_N) = \beta_i\left(\sum_{i=1}^{N} \lambda_j(1 - s_j)\right).$$

In the case of exhaustive service,

$$h_i(s_1, \ldots, s_N) = \theta_i\left(\sum_{j \neq i} \lambda_j(1 - s_j)\right),$$

where $\theta_i(\cdot)$ stands for the LST of the duration of a busy period in an $M/G/1$ system with arrival rate λ_i and service-time LST $\beta_i(\cdot)$. Under binomial-gated service [18], where each of the type-i customers present at a polling instant at Q_i is served with probability p_i ($0 < p_i \leqslant 1$), we have

$$h_i(s_1, \ldots, s_N) = p_i\beta_i\left(\sum_{i=1}^{N} \lambda_j(1 - s_j)\right) + (1 - p_i)s_i.$$

In the case of binomial-exhaustive service [6], where each of the type-i customers present at a polling instant at Q_i generates an $M/G/1$ busy period with probability q_i ($0 < q_i \leqslant 1$), we have

$$h_i(s_1, \ldots, s_N) = q_i\theta_i\left(\sum_{j \neq i} \lambda_j(1 - s_j)\right) + (1 - q_i)s_i.$$

Define the *exhaustiveness* of the service discipline at Q_i by

$$f_i := 1 - E[L_i] \quad (i = 1, \ldots, N). \tag{3}$$

For the cases of gated, exhaustive, binomial-gated and binomial-exhaustive service, the exhaustiveness is given by $1 - \rho_i$, 1, $p_i(1 - \rho_i)$ and q_i, respectively.

Based on the assumption that the service disciplines are non-idling, the following relation between the expected values of T_i and L_i holds:

$$E[T_i] = \big(1 - E[L_i]\big)\frac{b_i}{1 - \rho_i} \quad (i = 1, \ldots, N). \tag{4}$$

To see this, consider a sub-busy period at Q_i initiated by a type-i customer C_j. Then it is readily verified that $E[L_i]$ equals 1, *plus* the expected number of type-i customers arriving at Q_i during the sub-busy period, *minus* the expected number of type-i customers served during the sub-busy period. Evidently, the mean values of the latter two quantities are equal to $\lambda_i E[T_i]$ and $E[T_i]/b_i$, respectively. These observations yield $E[L_i] = 1 + \lambda_i E[T_i] - E[T_i]/b_i$, which directly implies the validity of (4).

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

A necessary and sufficient condition for the stability of the system is $\rho < 1$ (cf. [11]). In the sequel, 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 behavior of $E[W_k]$, the expected delay at Q_k, in heavy traffic. Throughout, $E[W_k]$ will be considered as function of ρ. To be specific, we assume that the arrival rates are parametrized as $\lambda_i = a_i\rho$, where the 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. [11]). More precisely, $E[W_k]$ has a first-order pole at $\rho = 1$. Therefore, we may write

$$E[W_k] = \frac{\omega_k}{1 - \rho} + \mathrm{o}\big((1 - \rho)^{-1}\big) \quad (k = 1, \ldots, N), \tag{5}$$

where $\mathrm{o}((1 - \rho)^{-1})$ stands for a function of ρ which becomes negligible compared to $(1 - \rho)^{-1}$ when $\rho \uparrow 1$. Based on (5), the analysis will be oriented towards the determination of

$$\omega_k = \lim_{\rho\uparrow 1}(1 - \rho)\ E[W_k] \quad (k = 1, \ldots, N), \tag{6}$$

referred to as the *scaled expected delay* at Q_k. In words, ω_k indicates the *rate* at which $E[W_k]$ tends to infinity as $\rho \uparrow 1$.

The basic result of the paper is the following closed-form expression for the scaled expected delay at each of the queues.

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

$$\omega_i = \frac{(1-\rho_i)((2/f_i)-1)}{\sum_{j=1}^{N} \rho_j(1-\rho_j)((2/f_j)-1)} \frac{\beta_2}{2\beta_1} + \frac{1}{2} r(1-\rho_i)\left(\frac{2}{f_i} - 1\right), \tag{7}$$

where the right-hand side is evaluated at $\rho = 1$.

Sections 3 and 4 will be focused on the derivation of theorem 1.

Finally, we introduce some notation. All vectors are N-dimensional, and all matrices are N by N, unless indicated otherwise. A vector $\mathbf{v}$ consists of components $(v_1, \ldots, v_N)$. Define

$$|\mathbf{v}| := \sum_{i=1}^{N} v_i.$$

The vector $\mathbf{e}_i$ stands for the ith unit vector, $i = 1, \ldots, N$. Each entry of the vector $\mathbf{1}$ equals 1. Let the norm of a matrix $\mathbf{A} = (a_{i,j})$ be defined as $\|\mathbf{A}\| := \max_{i,j} |a_{i,j}|$. Indices corresponding to queue numbers are cyclic: index i should be read as $((i-1) \bmod N) + 1$ notation. Denote by I_E the indicator function on the event E.

3. The Descendant Set Approach (DSA)

Let X_k be the number of type-k customers present in the system at a polling instant at Q_i, when the system is in equilibrium.

Then the expected waiting time at Q_k can be expressed in terms of the first two moments of X_k as follows (cf. [3]): for $k = 1, \ldots, N$,

$$\begin{aligned} E[W_k] = {} & \frac{\lambda_k b_k^{(2)}}{2(1-\rho_k)} - \frac{E[L_k(L_k-1)]}{2\lambda_k(1-E[L_k])} \\ & + \frac{\mathrm{Var}[X_k] + (E[X_k])^2 - E[X_k]}{2\lambda_k E[X_k]}\big(1 + E[L_k]\big). \end{aligned} \tag{8}$$

The quantities $E[X_k]$ can be derived in closed form. To this end, it is readily verified that

$$E[X_k] = \lambda_k r + \lambda_k \sum_{j \neq k} E[X_j]E[T_j] + E[X_k]E[L_k].$$

This implies

$$\frac{E[T_i]}{\rho_i} E[X_i] = \frac{E[T_j]}{\rho_j} E[X_j] \quad \text{for all } i, j = 1, \ldots, N,$$

so that $E[X_k]$ can be expressed in closed form as follows: for $k = 1, \ldots, N$,

$$E[X_k] = \frac{r}{1-\rho} \frac{\rho_k}{E[T_k]}. \tag{9}$$

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

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 which arrives at the system during a switch-over period. A non-originator is a customer who arrives at the system during the service of another customer. For a customer C, let the *children set* be the set of customers arriving during the service of C; the *descendant set* of C is recursively defined to consist of C, its children and the descendants of its children.

The DSA is focused on the determination of the moments of the delay 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. To this end, 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*

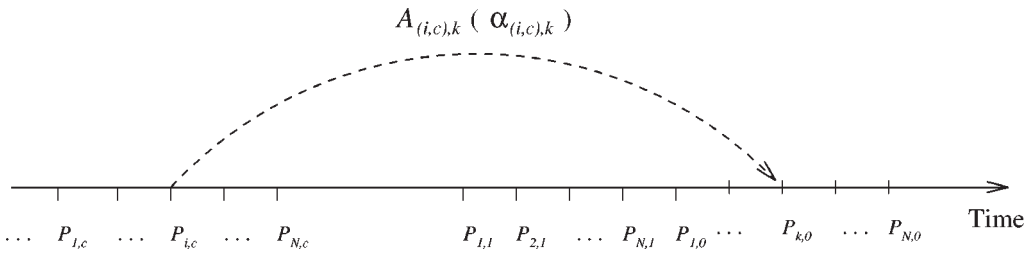

Figure 1. Conribution of an (i, c)-customer to $X_k(P_{k,0})$.

of T to $X_k(P_{k,0})$. Denote by $A_{(i,c),k}(z)$ the probability generating function (pgf) of $A_{(i,c),k}$, and define the first two (factorial) moments of $A_{(i,c),k}$ by

$$\alpha_{(i,c),k} = E[A_{(i,c),k}] \quad \text{and} \quad \alpha^{(2)}_{(i,c),k} = E\big[A_{(i,c),k}(A_{(i,c),k}-1)\big].$$

The variables $\alpha_{(i,c),k}$ and $\alpha^{(2)}_{(i,c),k}$ play a key role in DSA. The mean and the variance of X_k can be expressed in terms of the variables $\alpha_{(i,c),k}$ and $\alpha^{(2)}_{(i,c),k}$ as follows (cf. [13]): for $k = 1, \ldots, N$,

$$E[X_k] = \sum_{i=1}^{N} r_i \sum_{c=0}^{\infty} \left[\sum_{j=i+1}^{N} \lambda_j \alpha_{(j,c),k} + \sum_{j=1}^{i} \lambda_j \alpha_{(j,c-1),k} \right], \tag{10}$$

and

$$\begin{aligned} \mathrm{Var}[X_k] = &\sum_{i=1}^{N} \big(r_i^{(2)} - r_i^2\big) \sum_{c=0}^{\infty} \left[\sum_{j=i+1}^{N} \lambda_j \alpha_{(j,c),k} + \sum_{j=1}^{i} \lambda_j \alpha_{(j,c-1),k} \right]^2 \\ &+ \sum_{i=1}^{N} r_i \sum_{c=0}^{\infty} \left[\sum_{j=i+1}^{N} \lambda_j \alpha^{(2)}_{(j,c),k} + \sum_{j=1}^{i} \lambda_j \alpha^{(2)}_{(j,c-1),k} \right]. \end{aligned} \tag{11}$$

Based on relations (10) and (11), $E[W_k]$ can be expressed in terms of the variables $\alpha_{(i,c),k}$ and $\alpha^{(2)}_{(i,c),k}$. The remainder of section 3 we discuss how the variables $\alpha_{(i,c),k}$ and $\alpha^{(2)}_{(i,c),k}$ can be computed recursively.

3.2. Recursion via immediate children of predecessor

The DSA is based on recursive relations between the variables $\alpha_{(i,c),k}$ and $\alpha^{(2)}_{(i,c),k}$ which we review next.

To this end, 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 $X_k(P_{k,0})$. We observe that this contribution is equal to the total contribution to $X_k(P_{k,0})$ of all *immediate children* of $T_i(P_{i,c})$, i.e., the customers which arrive during the service of $T_i(P_{i,c})$. To be precise, if $T_i(P_{i,c})$ has not yet been served at $P_{k,0}$, then $T_i(P_{i,c})$ contributes 1 to $X_k(P_{k,0})$ if $i = k$ and 0 if $i \neq k$. The DSA is based on computing $\alpha_{(i,c),k}$ and $\alpha^{(2)}_{(i,c),k}$ from the contribution of the children of $T_i(P_{i,c})$. The type-j ($j \neq i$) children of $T_i(P_{i,c})$ are exactly the customers which arrived at Q_j during the sub-busy period generated by $T_i(P_{i,c})$; the type-i children of $T_i(P_{i,c})$ are the residue of the sub-busy period initiated by $T_i(P_{i,c})$. It is easily verified that we obtain the following expression for $A_{i,c}(z)$: for $i, k = 1, \ldots, N$, $c = 0, 1, \ldots$,

$$A_{(i,c),k}(z) = \psi_i\Bigg(\sum_{j=i+1}^{N} \lambda_j \big(1 - A_{(j,c),k}(z)\big) + \sum_{j=1}^{i-1} \lambda_j \big(1 - A_{(j,c-1),k}(z)\big), A_{(i,c-1),k}(z)\Bigg). \tag{12}$$

Differentiating once with respect to z and substituting $z = 1$ leads to the following equation: for $i, k = 1, \ldots, N$, $c = 0, 1, \ldots$,

$$\alpha_{(i,c),k} = E[T_i]\left[\sum_{j=i+1}^{N} \lambda_j \alpha_{(j,c),k} + \sum_{j=1}^{i-1} \lambda_j \alpha_{(j,c-1),k} \right] + E[L_i]\alpha_{(i,c-1),k}, \tag{13}$$

and differentiating two times with respect to z and substituting $z = 1$ yields: for $i, k = 1, \ldots, N$, $c = 0, 1, \ldots$,

$$\alpha^{(2)}_{(i,c),k} = E[T_i]\left[\sum_{j=i+1}^{N} \lambda_j \alpha^{(2)}_{(j,c),k} + \sum_{j=1}^{i-1} \lambda_j \alpha^{(2)}_{(j,c-1),k} \right] + E[L_i]\alpha^{(2)}_{(i,c-1),k} + \xi_{(i,c),k}, \tag{14}$$

where

$$\xi_{(i,c),k} := E[T_i^2]\left[\sum_{j=i+1}^{N} \lambda_j \alpha_{(j,c),k} + \sum_{j=1}^{i-1} \lambda_j \alpha_{(j,c-1),k} \right]^2 + E\big[L_i(L_i - 1)\big]\alpha^2_{(i,c-1),k} + 2E[L_i T_i]\alpha_{(i,c),k}\left[\sum_{j=i+1}^{N} \lambda_j \alpha_{(j,c),k} + \sum_{j=1}^{i-1} \lambda_j \alpha_{(j,c-1),k} \right]. \tag{15}$$

The initial conditions are as follows (cf. [13]): $A_{(k,0),k}(z) := z$; $A_{(i,0),k}(z) := 1$ $(i = k+1, \ldots, N)$; $A_{(i,-1),k}(z) := 1$ $(i = 1, \ldots, k-1)$. Differentiating once and twice and substituting $z = 1$ gives the initial conditions for the variables $\alpha_{(i,c),k}$ and $\alpha^{(2)}_{(i,c),k}$, respectively: for $k = 1, \ldots, N$,

$$\alpha_{(k,0),k} := 1; \qquad \alpha_{(i,0),k} := 0 \quad (i = k+1, \ldots, N);$$
$$\alpha_{(i,-1),k} := 0 \quad (i = 1, \ldots, k-1),$$

and

$$\alpha^{(2)}_{(k,0),k} := 1; \qquad \alpha^{(2)}_{(i,0),k} := 0 \quad (i = k+1, \ldots, N);$$
$$\alpha^{(2)}_{(i,-1),k} := 0 \quad (i = 1, \ldots, k-1).$$

Starting with these initial values, all coefficients $\alpha_{(i,c),k}$ and $\alpha^{(2)}_{(i,c),k}$ can be recursively determined according to (13) and (14), respectively.[1] In this way, for *fixed* k the

[1] To be precise, relations (13) and (14) are only defined for $c = 0, 1, \ldots$, for $i = 1, \ldots, k-1$, and for $c = 1, 2, \ldots$, for $i = k, \ldots, N$.

variables $\alpha_{(i,c),k}$ (and $\alpha^{(2)}_{(i,c),k}$) can be computed in the order $\alpha_{(i,0),k}$ ($\alpha^{(2)}_{(i,0),k}$), for $i = k-1, k-2, \ldots, 1$, followed by $\alpha_{(i,c),k}$ ($\alpha^{(2)}_{(i,c),k}$), $i = N, N-1, \ldots, 1$, for $c = 1, 2, \ldots$.

3.3. Recursion via immediate parents of descendants

The recursive equations in (13)–(14) relate the variables $\alpha_{(i,c),k}$ and $\alpha^{(2)}_{(i,c),k}$ 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 $\alpha_{(i,c),k}$ ($\alpha^{(2)}_{(i,c),k}$) and $\alpha_{(i,c),l}$ ($\alpha^{(2)}_{(i,c),l}$) for $l \neq k$. The derivation of such relations requires to conduct the recursion in a different way: 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 $P_{i,c}$, and try to find the contribution of $T_i(P_{i,c})$ to $X_k(P_{k,0})$. To this end, we consider the most recent polling instants of Q_j ($j = 1, \ldots, N$), prior to $P_{k,0}$, denoted by P^*_j (see figure 3), and derive the contribution of $T_i(P_{i,c})$ to $X_k(P_{0,k})$ by conditioning on the contribution of $T_i(P_{i,c})$ to $X_j(P^*_j)$. A crucial observation is that the distribution of the

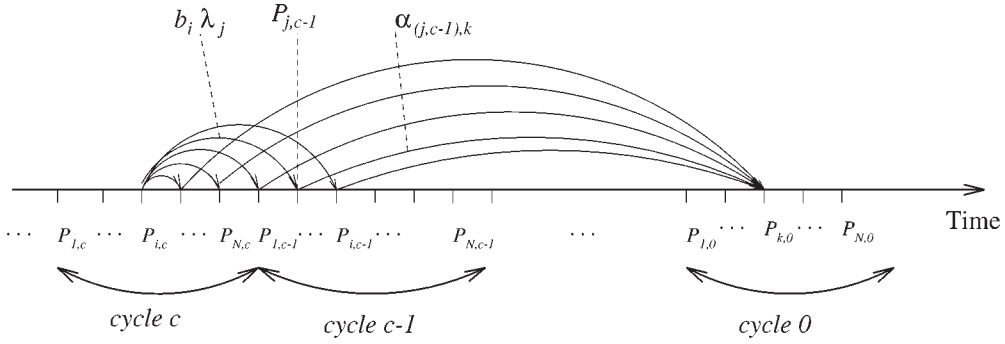

Figure 2. Recursion via immediate children of predecessor.

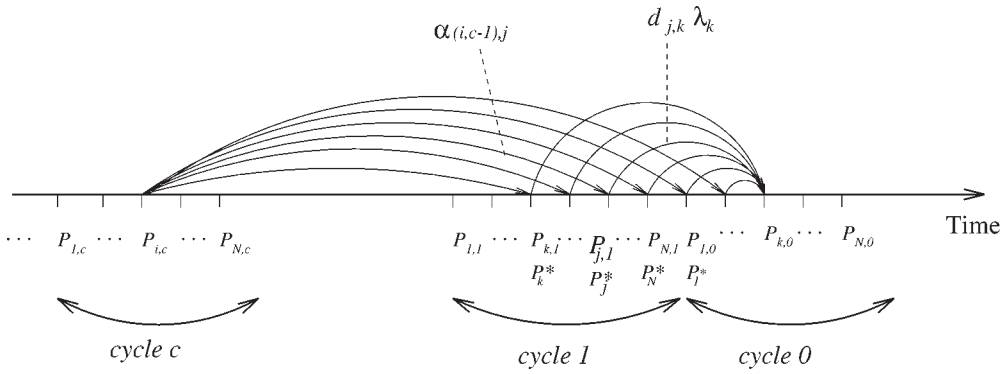

Figure 3. Recursion via immediate parents of descendants.

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 $j = k, \ldots, N$. Moreover, we observe that each descendant of $T_i(P_{i,c})$ at $P_{k,0}$ has arrived during the service period generated by exactly one customer present at P_j^* ($j = 1, \ldots, N$), referred to as the *immediate parent*. The following result follows then directly from the fact that the expected number of type-k customers at $P_{k,0}$ whose immediate parent is of type-j, is given by $\lambda_k E[T_j]$ if $j \neq k$, and by $E[L_k]$ if $j = k$ (cf. [20]): for $i, k = 1, \ldots, N$, $c = 0, 1, \ldots$,

$$\alpha_{(i,c),k} = \lambda_k \left[\sum_{j=k+1}^{N} E[T_j] \alpha_{(i,c),j} + \sum_{j=1}^{k-1} E[T_j] \alpha_{(i,c-1),j} \right] + E[L_k] \alpha_{(i,c-1),k}. \qquad (16)$$

The initial conditions are (see also section 3.2): $\alpha_{(k,0),k} := 1$; $\alpha_{(i,0),k} := 0$ ($k = 1, \ldots, i$); $\alpha_{(i,-1),k} := 0$ ($k = i+1, \ldots, N$).[2] Using these initial conditions, for fixed i, the variables $\alpha_{(i,c),k}$ can be computed recursively in the order $\alpha_{(i,0),k}$ for $k = i+1, \ldots, N$, followed by $\alpha_{(i,c),k}$ ($\alpha^{(2)}_{(i,c),k}$) for $c = 1, 2, \ldots$, for $k = 1, \ldots, N$. We emphasize that for computational purposes the recursive scheme discussed in section 3.2 is preferred to the scheme given here, because the former allows for the computation of individual expected waiting times (i.e., for specific queues). However, relation (16) will play a key role in the derivation of theorem 1.

It will be useful to express relations (13), (14) and (16) in matrix notation. To this end, define by $\boldsymbol{\alpha}_{(\cdot,c),k}$ ($\boldsymbol{\alpha}^{(2)}_{(\cdot,c),k}$) to be the vector whose ith component equals $\alpha_{(i,c),k}$ ($\alpha^{(2)}_{(i,c),k}$), $c = 0, 1, \ldots$. Moreover, let $\widehat{\mathbf{M}}_i$ be the matrix whose (j,k)th element equals $I_{\{j=k\}}$ for $j \neq i$, and whose (i,k)th element equals $\lambda_k E[T_i]$ for $k \neq i$ and $E[L_i]$ for $k = i$. Define $\mathbf{M} := \widehat{\mathbf{M}}_1 \widehat{\mathbf{M}}_2 \cdots \widehat{\mathbf{M}}_N$, and let $\boldsymbol{\psi}_k$ be the vector whose ith component is $\alpha_{(i,0),k}$ ($i, k = 1, \ldots, N$). It will be convenient to write equations (13) in the following form:

$$\boldsymbol{\alpha}_{(\cdot,0),k} = \boldsymbol{\psi}_k; \qquad \boldsymbol{\alpha}_{(\cdot,c),k} = \mathbf{M} \boldsymbol{\alpha}_{(\cdot,c-1),k} = \mathbf{M}^c \boldsymbol{\psi}_k$$
$$(k = 1, \ldots, N, \ c = 1, 2, \ldots). \qquad (17)$$

To rewrite (14) in matrix notation, define $\boldsymbol{\psi}^{(2)}_k$ to be the vector whose ith component equals $\alpha^{(2)}_{(i,0),k}$ ($i, k = 1, \ldots, N$). This leads to the following set of equations:

$$\boldsymbol{\alpha}^{(2)}_{(\cdot,0),k} = \boldsymbol{\psi}^{(2)}_k; \qquad \boldsymbol{\alpha}^{(2)}_{(\cdot,c),k} = \mathbf{M} \boldsymbol{\alpha}^{(2)}_{(\cdot,c-1),k} + \boldsymbol{\chi}_{c,k} = \mathbf{M}^c \boldsymbol{\psi}^{(2)}_k + \sum_{j=0}^{c-1} \mathbf{M}^j \boldsymbol{\chi}_{c-j,k}$$
$$(k = 1, \ldots, N, \ c = 1, 2, \ldots), \qquad (18)$$

[2] To be precise, relations (16) are only defined for $c = 0, 1, \ldots$, for $k = i+1, \ldots, N$, and for $c = 1, 2, \ldots$, for $k = 1, \ldots, i$.

where

$$\boldsymbol{\chi}_{c,k} := \sum_{j=1}^{N} \xi_{(j,c),k} \widehat{\mathbf{M}}_1 \cdots \widehat{\mathbf{M}}_{j-1} \mathbf{e}_j \quad (k = 1, \ldots, N,\ c = 1, 2, \ldots). \tag{19}$$

Finally, to rewrite (16) in matrix notation, define by $\boldsymbol{\alpha}_{(i,c),\cdot}$ the vector whose kth component equals $\alpha_{(i,c),k}$ $(i, k = 1, \ldots, N,\ c = 0, 1, \ldots)$. Moreover, let $\widehat{\mathbf{N}}_k$ be the matrix whose (i, j)th element equals $I_{\{i=j\}}$ for $i \neq k$, and whose (k, j)th element equals $\lambda_k E[T_j]$ for $j \neq k$ and $E[L_k]$ for $j = k$, and let $\mathbf{N} = \widehat{\mathbf{N}}_N \cdots \widehat{\mathbf{N}}_1$. If we define $\widehat{\boldsymbol{\psi}}_i$ to be the vector whose kth component equals $\alpha_{(i,0),k}$ $(i, k = 1, \ldots, N)$, then equations (16) can be expressed as follows:

$$\boldsymbol{\alpha}_{(i,0),\cdot} = \widehat{\boldsymbol{\psi}}_i; \qquad \boldsymbol{\alpha}_{(i,c),\cdot} = \mathbf{N} \boldsymbol{\alpha}_{(i,c-1),\cdot} = \mathbf{N}^c \widehat{\boldsymbol{\psi}}_i \quad (i = 1, \ldots, N,\ c = 1, 2, \ldots). \tag{20}$$

4. Analysis

In this section we use the concept of descendant sets to derive theorem 1. From relations (8) and (9) it remains to find an expression for Var$[X_k]$ under heavy-traffic assumptions. Recall that Var$[X_k]$ is related to the descendant set variables $\alpha_{(i,c),k}$ and $\alpha^{(2)}_{(i,c),k}$ according to relation (11). Since Var$[X_k]$ tends to infinity for $\rho \uparrow 1$, the rate of tendency to infinity of Var$[X_k]$ is determined by the *tail* behavior of the sequences $\{\alpha_{(i,c),k},\ c = 0, 1, \ldots\}$ and $\{\alpha^{(2)}_{(i,c),k},\ c = 0, 1, \ldots\}$.

The derivation of theorem 1 consists of two parts. First, by using the concept of descendant sets we obtain an expression for the ratios between the values of Var$[X_k]$ for $\rho \uparrow 1$. Then, the expression for the scaled expected delay is obtained by substituting the ratios into the so-called pseudo conservation law, i.e., a known closed-form expression for a specific weighted sum of the expected waiting times.

The variables $\alpha_{(i,c),k}$ are completely determined by the recursive schemes (17) or (20). Both recursive schemes (which determine the same values of $\alpha_{(i,c),k}$, but in a different order) constitute a set of first-order homogeneous difference equations. From the theory of difference equations it is known that the recursive equations can be solved explicitly if the eigenvalues and eigenvectors of $\mathbf{M}$ or $\mathbf{N}$ are known. However, the eigenvalues and the eigenvectors are generally unknown for $\rho < 1$. Nevertheless, to analyze the system behavior in heavy traffic we do not need to solve the eigenvalues and eigenvectors explicitly for general $\rho < 1$ and then letting $\rho \uparrow 1$. Instead, it will be sufficient to obtain the eigenvectors of $\mathbf{M}$ and $\mathbf{N}$ *at* $\rho = 1$, which will be shown to be rather simple.

The following lemma states that $\mathbf{M}^c$ and $\mathbf{N}^c$ can be decomposed into two parts, one of which becomes dominant when c gets large.

Lemma 1 (decomposition). The matrix $\mathbf{M}$ has a unique, strictly positive eigenvalue $\gamma_{\max}$, with multiplicity 1, with associated right and left eigenvectors $\mathbf{u}$ and $\mathbf{v}$. If $\mathbf{u}$ and $\mathbf{v}$ are normalized such that $\mathbf{u}^{\mathrm{T}}\mathbf{1} = \mathbf{u}^{\mathrm{T}}\mathbf{v} = 1$, then: for $c = 0, 1, \ldots,$

$$\mathbf{M}^c = \gamma_{\max}^c \mathbf{u}\mathbf{v}^{\mathrm{T}} + \mathbf{R}^c, \tag{21}$$

where $\|\mathbf{R}^c\| < K\gamma^c$ for some $K < \infty$ and γ ($0 < \gamma < \gamma_{\max}$). Similarly, $\mathbf{N}^c$ can be decomposed as follows: for $c = 0, 1, \ldots,$

$$\mathbf{N}^c = \hat{\gamma}_{\max}^c \hat{\mathbf{u}}\hat{\mathbf{v}}^{\mathrm{T}} + \widehat{\mathbf{R}}^c, \tag{22}$$

where $\|\widehat{\mathbf{R}}^c\| < \widehat{K}\hat{\gamma}^c$ for some $\widehat{K} < \infty$ and $\hat{\gamma}$ ($0 < \hat{\gamma} < \hat{\gamma}_{\max}$).

Proof. The results follow from the Frobenius theorem for strictly positive matrices (cf., e.g., [1]). We refer to [20] for more details. □

From (17) and (21) it follows that $\alpha_{(i,c),k}$ can be rewritten as follows: for $i, k = 1, \ldots, N$, $c = 0, 1, \ldots,$

$$\alpha_{(i,c),k} = \mathbf{e}_i^{\mathrm{T}} \mathbf{M}^c \boldsymbol{\psi}_k = \gamma_{\max}^c u_i \mathbf{v}^{\mathrm{T}} \boldsymbol{\psi}_k + r_{i,k}^{(c)} \tag{23}$$

with $r_{i,k}^{(c)} < K\gamma^c$ for some $K < \infty$ and γ ($0 < \gamma < \gamma_{\max}$). Alternatively, using (22) we may write: for $i, k = 1, \ldots, N$, $c = 0, 1, \ldots,$

$$\alpha_{(i,c),k} = \mathbf{e}_k^{\mathrm{T}} \mathbf{N}^c \widehat{\boldsymbol{\psi}}_k = \hat{\gamma}_{\max}^c \hat{u}_k \hat{\mathbf{v}}^{\mathrm{T}} \widehat{\boldsymbol{\psi}}_i + \hat{r}_{i,k}^{(c)}, \tag{24}$$

with $\hat{r}_{i,k}^{(c)} < \widehat{K}\hat{\gamma}^c$ for some $\widehat{K} < \infty$ and $\hat{\gamma}$ ($0 < \hat{\gamma} < \hat{\gamma}_{\max}$). Relations (23) and (24) will be useful in the remainder of this section.

The eigenvalues $\gamma_{\max}$ and $\hat{\gamma}_{\max}$ are referred to as the maximal eigenvalues of the matrices **M** and **N**, respectively. The following lemma gives properties of the maximal eigenvalues.

Lemma 2 (maximal eigenvalues).

(1) If $\rho < 1$ then $\gamma_{\max}, \hat{\gamma}_{\max} < 1$.

(2) If $\rho = 1$ then $\gamma_{\max}, \hat{\gamma}_{\max} = 1$.

(3) If $\rho = 1$ then $\mathbf{u} = |\mathbf{b}|^{-1}\mathbf{b}$.

(4) If $\rho = 1$ then $\hat{\mathbf{u}} = |\mathbf{w}|^{-1}\mathbf{w}$, where $w_i := \rho_i / E[T_i]$ ($i = 1, \ldots, N$).

(5) $\lim_{\rho\uparrow 1} \gamma_{\max} = \lim_{\rho\uparrow 1} \hat{\gamma}_{\max} = 1$; $\lim_{\rho\uparrow 1} \mathbf{u} = |\mathbf{b}|^{-1}\mathbf{b}$; $\lim_{\rho\uparrow 1} \hat{\mathbf{u}} = |\mathbf{w}|^{-1}\mathbf{w}$.

Proof. The validity of parts 1 and 2 for $\gamma_{\max}$ is shown in [23]. The validity of parts 1 and 2 for $\hat{\gamma}_{\max}$ follows then directly from (23) and (24). To proof part 3, note that for $\rho = 1$ we have

$$E[T_i] \sum_{j=1, j\neq i}^{N} \lambda_j b_j + E[L_i] b_i = E[T_i](1 - \rho_i) + E[L_i] b_i = b_i$$

(where the latter equality follows from (4)), which implies $\widehat{\mathbf{M}}_i\mathbf{b} = \mathbf{b}$ $(i = 1, \ldots, N)$ and hence, $\mathbf{Mb} = \mathbf{b}$. Similarly, part 4 follows from the fact that for $\rho = 1$ we have

$$\lambda_k \sum_{j \neq k} E[T_j]w_j + E[L_k]w_k = \lambda_k(1 - \rho_k) + \frac{E[L_k]\rho_k}{E[T_k]} = w_k$$

(where the latter equality follows from (4)), so that $\widehat{\mathbf{N}}_k\mathbf{w} = \mathbf{w}$ for all $k = 1, \ldots, N$, and hence, $\mathbf{Nw} = \mathbf{w}$. Part 5 follows from the continuity of the eigenvectors and eigenvalues in the entries of $\mathbf{M}$ and $\mathbf{N}$ (cf. [12]) which, in turn, are continuous in ρ. □

Remark 4.1. Lemma 2 implies that the sequence $\{\alpha_{(i,c),k},\ c = 0, 1, \ldots\}$ converges to 0 for $\rho < 1$ and converges to a constant for $\rho = 1$. Moreover, from (23) and (24) and lemma 2 it follows that for $i, j, k, l = 1, \ldots, N$,

$$\lim_{\rho \uparrow 1} \lim_{c \to \infty} \frac{\alpha_{(i,c),k}}{\alpha_{(j,c),l}} = \frac{b_i}{b_j} \frac{\rho_k / E[T_k]}{\rho_l / E[T_l]}, \tag{25}$$

where the right-hand side is evaluated at $\rho = 1$. To give an intuitive interpretation for (25), it is clear that for $\rho = 1$ the expected total number of type-k descendants of each of the original customers on the long run is proportional to λ_k, regardless of the type of the original customer and of the service disciplines at the queues. However, for $\rho = 1$, during the service-period generated by a type-k customer which is present at a polling instant at Q_k, on the average $E[T_k]/b_k$ customers are served, while none of other customers served the service period was present as the polling instant. Hence, on the average a fraction $b_k/E[T_k]$ of the customers is actually present at a polling instant at Q_k. These observations indicate that $\alpha_{(i,c),k}$ becomes proportional to $\lambda_k b_k / E[T_k]$ for $c \to \infty$. To explain why $\alpha_{(i,c),k}$ is also proportional to b_i (in the limiting case $\rho = 1$) we observe that each type-i customer has on the average $\lambda_j b_i$ type-j children, regardless of the service discipline at Q_i. To see this, note that it does not matter whether its children at Q_i are served during the same visit or not, because the numbers of type-k descendants in the infinite future of each of these children tends to a constant.

Remark 4.2. From (4) one may verify that $\sum_{c=0}^{\infty} \alpha_{(i,c),k}$ possesses a first-order pole at $\rho = 1$. Moreover, it is easy to verify by using (23) and (24) that: for $i, j, k, l = 1, \ldots, N$,

$$\lim_{\rho \uparrow 1} \frac{\sum_{c=0}^{\infty} \alpha_{(i,c),k}}{\sum_{c=0}^{\infty} \alpha_{(j,c),l}} = \frac{b_i}{b_j} \frac{\rho_k / E[T_k]}{\rho_l / E[T_l]}, \tag{26}$$

where the right-hand side is evaluated at $\rho = 1$.

Lemma 3. For $k, l = 1, \ldots, N$,

$$\lim_{\rho \uparrow 1} \frac{\mathrm{Var}[X_k]}{\mathrm{Var}[X_l]} = \lim_{\rho \uparrow 1} \frac{\sum_{i=1}^{N} r_i \sum_{c=0}^{\infty} \left[\sum_{j=i+1}^{N} \lambda_j \alpha^{(2)}_{(j,c),k} + \sum_{j=1}^{i} \lambda_j \alpha^{(2)}_{(j,c-1),k} \right]}{\sum_{i=1}^{N} r_i \sum_{c=0}^{\infty} \left[\sum_{j=i+1}^{N} \lambda_j \alpha^{(2)}_{(j,c),l} + \sum_{j=1}^{i} \lambda_j \alpha^{(2)}_{(j,c-1),l} \right]}. \tag{27}$$

Proof. First, we observe that $\mathrm{Var}[X_k]$ has a second-order pole at $\rho = 1$. This is an immediate consequence of equations (8) and (9) and the fact that $E[W_k]$ has a first-order pole at $\rho = 1$. Substituting (23) into (10), it follows from (9) that $\sum_{c=0}^{\infty}[\sum_{j=i+1}^{N} \lambda_j \alpha_{(j,c),k} + \sum_{j=1}^{i} \lambda_j \alpha_{(j,c-1),k}]$ possesses a first-order pole at $\rho = 1$ for $i, k = 1, \ldots, N$. Using (23), this implies that the first summation in (11) also possesses a first-order pole at $\rho = 1$. To this end, note that the term $(1 - \gamma_{\max})^{-1}$ corresponds to a first-order pole at $\rho = 1$, which implies that the term $(1 - \gamma_{\max}^2)^{-1}$ (which would occur by substituting (23) into the first summation in (11)), which equals $(1 - \gamma_{\max})^{-1}(1 + \gamma_{\max})^{-1}$, *also* possesses a first-order pole at $\rho = 1$.

Hence, the second summation in (11) must possess a second-order pole at $\rho = 1$. This motivates why the second summation dominates the first summation in (11) in the limiting case $\rho \uparrow 1$. □

Based on lemma 3, we focus on the limiting behavior of the sequence $\{\alpha^{(2)}_{(i,c),k},\ c = 0, 1, \ldots\}$. To this end, note that it follows from (28) and (21) that we may write: for $i, k = 1, \ldots, N,\ c = 0, 1, \ldots,$

$$\alpha^{(2)}_{(i,c),k} = \gamma^{c}_{\max} u_i \mathbf{v}^{\mathrm{T}} \boldsymbol{\psi}^{(2)}_k + \mathbf{e}_i^{\mathrm{T}} \mathbf{R}^c \boldsymbol{\psi}^{(2)}_k + \sum_{j=0}^{c-1} \left[\gamma^{j}_{\max} u_i \mathbf{v}^{\mathrm{T}} \boldsymbol{\chi}_{c-j,k} + \mathbf{e}_i^{\mathrm{T}} \mathbf{R}^j \boldsymbol{\chi}_{c-j,k}\right]. \quad (28)$$

Note that from lemma 2 and (28) it follows that the sequence $\{\alpha^{(2)}_{(i,c),k},\ c = 0, 1, \ldots\}$ tends to 0 for $\rho < 1$ and tends to a linearly increasing sequence for $\rho = 1$. Moreover, one may verify that the series $\sum_{c=0}^{\infty} \alpha^{(2)}_{(i,c),k}$ converges for $\rho < 1$, and diverges for $\rho = 1$.

Lemma 4. For $i, k, l = 1, \ldots, N$,

$$\lim_{\rho\uparrow 1} \frac{\sum_{c=0}^{\infty} \alpha^{(2)}_{(i,c),k}}{\sum_{c=0}^{\infty} \alpha^{(2)}_{(i,c),l}} = \lim_{\rho\uparrow 1} \frac{\sum_{c=0}^{\infty} \mathbf{v}^{\mathrm{T}} \boldsymbol{\chi}_{c,k}}{\sum_{c=0}^{\infty} \mathbf{v}^{\mathrm{T}} \boldsymbol{\chi}_{c,l}} = \lim_{\rho\uparrow 1} \left(\frac{\hat{u}_k}{\hat{u}_l}\right)^2 = \left(\frac{\rho_k / E[T_k]}{\rho_l / E[T_l]}\right)^2, \quad (29)$$

where the right-hand side is evaluated at $\rho = 1$.

Proof. The first equality follows from (28) and using lemmas 1 and 2. The second equality follows from definitions (19) and (15), and relation (24). The third equality follows from lemma 2. □

Theorem 2 (ratios between variances). For $k, l = 1, \ldots, N$,

$$\lim_{\rho\uparrow 1} \frac{\mathrm{Var}[X_k]}{\mathrm{Var}[X_l]} = \left(\frac{\rho_k / E[T_k]}{\rho_l / E[T_l]}\right)^2, \quad (30)$$

where the right-hand side is evaluated at $\rho = 1$.

Proof. The result follows directly from equation (27) and lemma 4. □

We are now ready to derive expressions for the ratios between the scaled expected delays.

Theorem 3 (ratios between scaled expected delays). For $k, l = 1, \ldots, N$,

$$\frac{\omega_k}{\omega_l} = \frac{(1-\rho_k)((2/f_k)-1)}{(1-\rho_l)((2/f_l)-1)}, \tag{31}$$

where the right-hand side is evaluated at $\rho = 1$.

Proof. The result is obtained by substituting (8) and (9) into (30) and some straightforward manipulations. □

From theorem 3, ω_k is known up to a yet unknown scaling factor.

The latter can be obtained from the following expression (cf. [5]): for $\rho < 1$,

$$\sum_{i=1}^{N} \rho_i E[W_i] = \frac{\rho}{1-\rho}\frac{\beta_2}{2\beta_1} + \rho\frac{r^{(2)}}{2r} + \frac{r}{2(1-\rho)}\left[\rho^2 - \sum_{i=1}^{N}\rho_i^2\right] + \sum_{i=1}^{N} E[M_i], \tag{32}$$

where M_i is the amount of work at Q_i at a departure instant at Q_i. Multiplying both sides by $(1-\rho)$ and letting $\rho \uparrow 1$ yields the following relation between the ω_k's:

$$\begin{aligned}\sum_{i=1}^{N} \rho_i\omega_i &= \frac{\beta_2}{2\beta_1} + \frac{r}{2}\left[1 - \sum_{i=1}^{N}\rho_i^2\right] + r\sum_{i=1}^{N}\rho_i(1-\rho_i)\frac{1-f_i}{f_i} \\ &= \frac{\beta_2}{2\beta_1} + \frac{r}{2}\sum_{i=1}^{N}\rho_i(1-\rho_i)\left(\frac{2}{f_i} - 1\right).\end{aligned} \tag{33}$$

To verify this, note that it follows from (9) and (3) that

$$E[M_i] = b_i E[X_i]E[L_i] = r\rho_i(1-\rho_i)(1-f_i)/f_i(1-\rho).$$

Combining (31) and (33) directly yields the basic result (theorem 1):

$$\omega_i = \frac{(1-\rho_i)((2/f_i)-1)}{\sum_{j=1}^{N}\rho_j(1-\rho_j)((2/f_j)-1)}\frac{\beta_2}{2\beta_1} + \frac{1}{2}r(1-\rho_i)\left(\frac{2}{f_i}-1\right), \quad i = 1, \ldots, N. \tag{34}$$

5. Discussion and implications of the results

Equation (34) forms a closed-form expression for the scaled expected delay in a polling system with a cyclic server and arbitrary branching-type service policy. Several properties of these systems can be concluded from this equation.

5.1. Equivalence and classification of policies via exhaustiveness factor

The scaled expected delay at Q_i depends on the service policy of all queues only via their exhaustiveness factor f_j, $j = 1, \ldots, N$. Thus, systems which differ in their service policies but equal in their exhaustiveness factors will have identical performance (in terms of scaled expected delay). This implies equivalence of the service policies in the sense that one service policy can be used to imitate another policy by a proper adaptation of policy parameters to match the exhaustiveness factors. For example, in a system with binomial gated service [18] we have $f_i^{\text{bin}-\text{gate}} = p_i^{\text{bin}-\text{gate}}(1-\rho_i)$ and in a system with binomial exhaustive service [6] we have $f_i^{\text{bin}-\text{exh}} = p_i^{\text{bin}-\text{exh}}$. Thus, the binomial exhaustive system can imitate the performance of the binomial gated system by setting $p_i^{\text{bin}-\text{exh}} = p_i^{\text{bin}-\text{gate}}(1-\rho_i)$.

For this reason one may classify and order various policies simply by their exhaustiveness factors. Two policies will be equivalent (in terms of their scaled delays) if their exhaustiveness factors are identical, regardless of the specific service policies implemented at the stations. This holds for systems with identical service policies in all stations as well as systems with mixed services. The exhaustiveness factor can be used as a single parameter to classify and compare policies (instead of conducting the comparison by a full analysis of the policies).

5.2. Monotonicity of the performance in the exhaustiveness parameters

Differentiation of (34) shows the following properties: for $i = 1, \ldots, N$,

(1) ω_i monotonically decreases in f_i,

(2) ω_i monotonically increases in f_j, $j \neq i$.

In other words, the greater the exhaustiveness of the service discipline at Q_i, the lower the delay experienced by its customers and the higher the delay experienced by the customers of the other queues.

Note that despite being quite intuitive such monotonicity property has not been proved (to the knowledge of the authors) in other settings of polling systems. For similar monotonicity properties see: (1) Borst et al. [4], where a semi-conjectured result is obtained regarding the optimal setting of the k-limits in a polling system with limited-k service, and (2) Levy et al. [16], where path-wise monotonicity of the *total work* in the exhaustiveness of the service policies is established. However, none of those works established the monotonicity of the (expected) delay of an individual queue in the parameter settings of any of the queues.

5.3. Sensitivity and insensitivity to station specific parameters

The scaled expected delay depends on the system parameters only through specific system parameters and is insensitive to others. The following properties follow directly from (34):

1. ω_k depends on the specific parameters of the stations only through their utilizations, $\rho_i = \lambda_i b_i$ $(i = 1, \ldots, N)$. It depends on the service time distributions at the queues only through the first two moments of the service time of an *arbitrary* customer, rather than on the (first two) moments of the individual service times.

2. ω_k depends on the switch-over times only through the first moment of the *total* switch-over time per cycle of the server along the queues.

3. ω_k is independent of the service order of the stations.

Note that these observed insensitivities are generally not valid for general $\rho < 1$. Apparently, these dependencies 'die out' when the system reaches saturation.

5.4. Optimization of operation

A reasonable performance measure which encapsulates the performance of the system as a whole is the weighted scaled delay:

$$C = \sum_{i=1}^{N} c_i \omega_i, \tag{35}$$

where c_i are arbitrary strictly positive cost parameters. Consider the problem of minimizing (35) with respect to the service disciplines. The solution of the problem is believed to be a good estimate for the optimal static assignment of service disciplines to the queues in heavily-loaded systems. Theorem 1 implies that the problem is equivalent to the problem of finding a vector $(f_1^*, \ldots, f_N^*)$ which minimizes (35). Note that optimization with respect to service order is irrelevant due to insensitivity to service order. Theorem 4 below gives a explicit partial solution to the optimization problem.

Theorem 4. If $c_k/\rho_k = \max_{j=1}^{N}\{c_j/\rho_j\}$, then $f_k = 1$.[3]

Proof. The following equations show that the cost function (35) is decreasing in f_k for all $(f_1, \ldots, f_N)$:

$$\frac{\partial}{\partial f_k}\sum_{j=1}^{N} c_j\omega_j = \frac{c_k}{\rho_k}\frac{\partial}{\partial f_k}\rho_k\omega_k + \sum_{j\neq k}\frac{c_j}{\rho_j}\frac{\partial}{\partial f_k}\rho_j\omega_j \leqslant \frac{c_k}{\rho_k}\frac{\partial}{\partial f_k}\rho_k\omega_k + \frac{c_k}{\rho_k}\sum_{j\neq k}\frac{\partial}{\partial f_k}\rho_j\omega_j$$

$$= \frac{c_k}{\rho_k}\frac{\partial}{\partial f_k}\sum_{j=1}^{N}\rho_j\omega_j \leqslant 0. \tag{36}$$

The first inequality follows from the definition of k (in theorem 4), and the second inequality follows from (33). □

[3] In case the (fully) exhaustive service discipline is technically infeasible, Q_k should be served 'as exhaustively as possible'.

The problem of finding the values of f_i^* for queues which are not covered by theorem 4 is more involved, and generally requires the use of numerical techniques. We leave this as a topic for future research.

6. Illustration

Equation (34) suggests that the scaled expected delay figures depend on the service disciplines of the queues only through their exhaustiveness factors, f_j ($j = 1, \ldots, N$). Thus, two different systems with the same sets of exhaustiveness parameters should have identical scaled expected delays.

To demonstrate this property, we consider an example of a polling system with 10 queues, whose parameters are: $b_1 = 5.5$, $b_2 = \cdots = b_{10} = 0.5$, $b_1^{(2)} = 100$, $b_2^{(2)} = \cdots = b_{10}^{(2)} = 1$, $r_i = 5$, $r_i^{(2)} = 25$ ($i \neq 6$ and $i \neq 8$), $r_6^{(2)} = 1025$, $r_8^{(2)} = 225$. The arrival rate of all queues is the same; we vary the arrival rate as to affect the total load and to examine $(1-\rho)E[W_i]$ as function of the load. As can be seen the system is very asymmetric.

We examine this system under two different service disciplines, the gated service and the fractional-exhaustive service [17]. In the latter system we select the fractional exhaustive parameters (p_i) as to match the exhaustiveness of the gated system. This is done by noting that under gated service we have $f_i = 1 - \rho_i$ and under fractional-exhaustive service we have $f_i = p_i(1-\rho_i)/(1-p_i\rho_i)$ and thus selection $p_i = 1/(1+\rho_i)$ for the fractional exhaustive system yields identical exhaustiveness systems.

In figure 4 we depict the scaled expected delay $(1-\rho)E[W_i]$ for Q_1 and Q_6 in these systems as function of ρ. The figure demonstrates that the expected delays under the two systems are very close to each other and converge towards each other as ρ

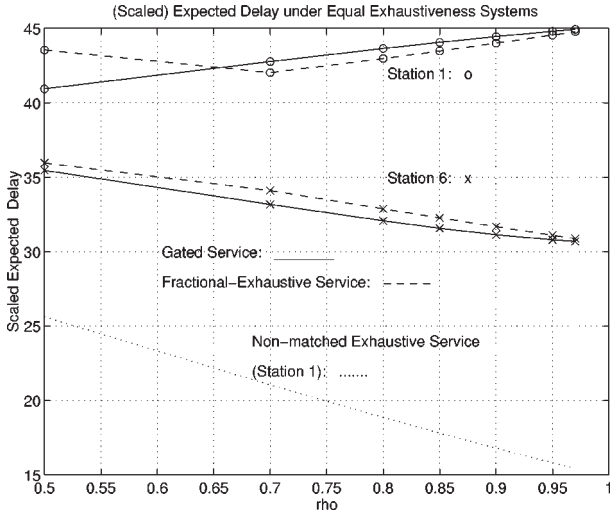

Figure 4. Scaled expected delay and exhaustiveness.

approaches 1. For comparison we also plot the expected delay of Q_1 in the fractional-exhaustive system where the p_i parameters are *not selected* to match the gated system (rather, they are all set to 1); note that this delay is significantly different from that in the gated system.

7. Approximation

Equation (34) suggests the following approximation for the expected waiting time at Q_i in stable systems (i.e., for $\rho < 1$):

$$E[W_i] \approx \frac{1}{1-\rho}\left[\frac{(1-\rho_i)((2/f_i)-1)}{\sum_{j=1}^{N}\rho_j(1-\rho_j)((2/f_j)-1)}\,\frac{\beta_2}{2\beta_1} + \frac{1}{2}\,r(1-\rho_i)\left(\frac{2}{f_i}-1\right)\right], \quad i = 1,\ldots,N. \tag{37}$$

This approximation provides a closed-form approximation for the expected delays in stable systems as function of the system parameters and the exhaustiveness of the service disciplines at the queues. Note that the analysis provided in this paper implies that this expression is exact when $\rho \uparrow 1$.

To evaluate the quality of this proposed approximation, we numerically examine several typical cases. First, we consider a system with binomial gated service (recall that the exhaustiveness factor of this policy is given by $f_i = p_i(1-\rho_i)$). The system consists of 10 queues whose parameters are: $b_1 = 5.5$, $b_2 = \cdots = b_{10} = 0.5$, $b_1^{(2)} = 100$, $b_2^{(2)} = \cdots = b_{10}^{(2)} = 1$, $r_i = 5$, $r_i^{(2)} = 25$. The binomial probabilities are: $p_1 = p_3 = p_5 = p_7 = \cdots = p_{10} = 1.0$, $p_2 = 0.5$, $p_4 = 0.3$, $p_6 = 0.1$. The arrival rate of all queues is the same; we vary the arrival rate as to affect the total load and to examine the approximation quality as function of the load. As can be seen the system is very asymmetric.

We examine the quality of the approximation (37) as function of the system load. Further, since the second moment of the switch-over periods does not serve as a parameter in equation (37), we examine the effect of its value on the quality of the approximation. This is done by considering two cases: (1) $r_i^{(2)} = 25$ ($i = 1,\ldots,10$), and (2) $r_i^{(2)} = 25$, $i \in \{1,2,3,4,5,7,9,10\}$, $r_6^{(2)} = 1025$, $r_8^{(2)} = 225$. To assess the quality of the approximation, we evaluate the expected delay in each of the 10 queues both by the approximation (denoted by z_i(app)) and exactly (denoted by z_i(exact)). For each of the 10 queues the *relative error* of the approximation of the expected waiting time at Q_i is defined as the *absolute value of*

$$\frac{z_i(\text{app}) - z_i(\text{exact})}{z_i(\text{exact})}. \tag{38}$$

Figure 5 below depicts the *maximal* relative error (over the 10 queues) of the approximation as function of the system load. Figure 5 demonstrates, as expected, that as we approach $\rho = 1$ the quality of the approximation improves. Further, it demonstrates that for most practical cases of high load (range of $\rho > 0.8$) the quality

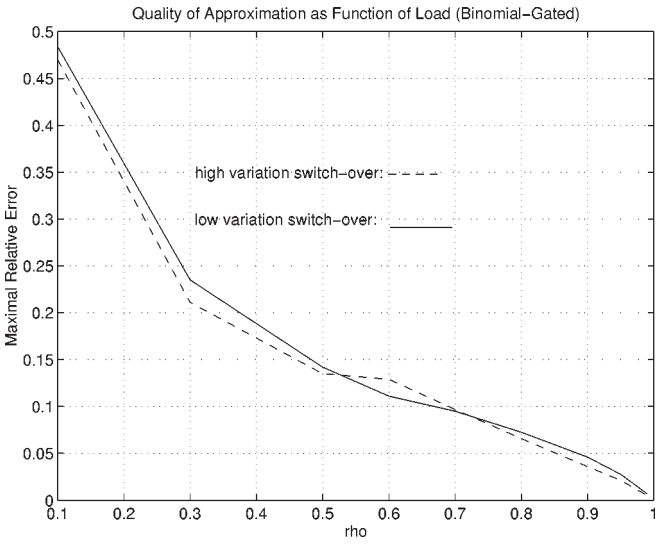

Figure 5. Quality of approximation: binomial gated service.

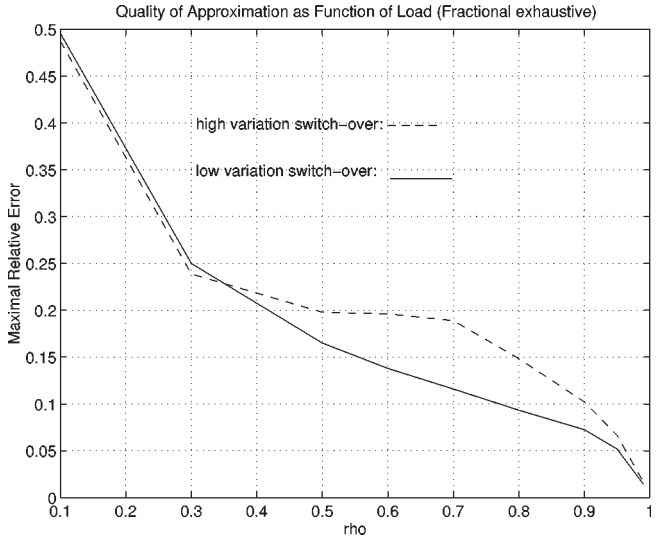

Figure 6. Quality of approximation: fractional exhaustive service.

of the approximation is very good (relative error less than 10%). This is the case for systems with either low or high value of switch-over variation (the approximation quality is somewhat better in the low variation case).

Next we conduct a similar comparison for a system with fractional exhaustive service (in all queues). Recall that the exhaustiveness factor of this policy is given by $f_i = p_i(1-\rho_i)/(1-p_i\rho_i)$. We consider the system under the same parameter sets chosen for the binomial-gated case. Evaluation of this system is depicted in figure 6.

We observe that, although the expected waiting times differ between the two systems drastically, the quality of the approximation in this case is quite similar to that of the binomial-gated system.

8. Topics for further research

The monotonicity property stated in section 5.2 provides new insights into the behavior of polling systems heavy traffic. One may investigate whether a similar monotonicity property also holds for stable, medium and lightly-loaded, systems. Such qualitative results would contribute strongly to the understanding of the stochastic behavior of polling systems. Analysis of the system for $\rho < 1$, however, would not only require properties of the tail behavior, but also of the heading part of the sequences $\{\alpha_{(i,c),k},\ c = 0, 1, \ldots\}$ and $\{\alpha^{(2)}_{(i,c),k},\ c = 0, 1, \ldots\}$. In general, for $\rho < 1$ the expected delay no longer depends on the service disciplines only through their exhaustiveness. Instead, the second moments of L_i and T_i should also be taken into account.

The analysis might be extended to obtain higher moments of the scaled delay. For the special case of mixtures of gated and exhaustive service at all queues, and with zero switch-over times, expressions for the kth moment of the delay at each of the queues have been obtained in [21].

The closed-form expressions obtained in this paper open possibilities for optimization purposes. For instance, in systems in which the load can be (statically) balanced, equations (34) may be useful for solving load-balancing problems in a heavy-traffic environment. Moreover, as indicated in section 5.4, the problem of minimizing (35) with respect to the service disciplines at the queues, should be pursued further.

References

[1] K.B. Athreya and P.E. Ney, *Branching Processes* (Springer, Berlin, 1971).
[2] J.P.C. Blanc, Performance evaluation of polling systems by means of the power-series algorithm, Ann. Oper. Res. 35 (1992) 155–186.
[3] S.C. Borst and O.J. Boxma, Polling models with and without switch-over times, Oper. Res. 45 (1997) 536–543.
[4] S.C. Borst, O.J. Boxma and H. Levy, The use of service limits for efficient operation of multi-station single-medium communication systems, IEEE ACM Trans. Networking 3 (1995) 602–612.
[5] O.J. Boxma and W.P. Groenendijk, Pseudo-conservation laws in cyclic service systems, J. Appl. Probab. 24 (1988) 949–964.
[6] O.J. Boxma, Workloads and waiting times in single-server systems with multiple customer classes, Queueing Systems 5 (1989) 185–214.
[7] G. Choudhury and W. Whitt, Computing transient and steady state distributions in poling models by numerical transform inversion, Performance Evaluation 25 (1996) 267–292.
[8] E.G. Coffman, A.A. Puhalskii and M.I. Reiman, Polling systems with zero switch-over times: a heavy-traffic principle, Ann. Appl. Probab. 5 (1995) 681–719.
[9] E.G. Coffman, A.A. Puhalskii and M.I. Reiman, Polling systems in heavy-traffic: a Bessel process limit, Preprint (1995).

[10] S.W. Fuhrmann, A decomposition result for a class of polling models, Queueing Systems 11 (1992) 109–120.
[11] C. Fricker and M.R. Jaïbi, Monotonicity and stability of periodic polling models, Queueing Systems 15 (1994) 211–238.
[12] T. Kato, *Perturbation Theory for Linear Operators* (Springer, New York, 1966).
[13] A.G. Konheim, H. Levy and M.M. Srinivasan, Descendant set: an efficient approach for the analysis of polling systems, IEEE Trans. Commun. 42 (1994) 1245–1253.
[14] K.K. Leung, Cyclic service systems with probabilistically-limited, IEEE J. Selected Areas Commun. 9 (1991) 185–193.
[15] H. Levy and M. Sidi, Polling models: applications, modeling and optimization, IEEE Trans. Commun. 38 (1991) 1750–1760.
[16] H. Levy, M. Sidi and O.J. Boxma, Dominance relations in polling systems, Queueing Systems 6 (1990) 155–171.
[17] H. Levy, Optimization of polling systems: the fractional exhaustive service method, Report, Tel-Aviv University (1988).
[18] H. Levy, Binomial-gated service: a method for effective operation and optimization of polling systems, IEEE Trans. Commun. 39 (1991) 1341–1350.
[19] D. Markowitz, Dynamic scheduling of single-server queues with setups: A heavy traffic approach, Ph.D. thesis, Operations Research Center, MIT, Cambridge, MA (1995).
[20] R.D. Van der Mei and H. Levy, Expected delay analysis of polling systems in heavy traffic, to appear in Adv. in Appl. Probab. (1996).
[21] R.D. Van der Mei, 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, 1997) pp. 275–284.
[22] M.I. Reiman and L.M. Wein, Dynamic scheduling of a two-class queue with setups, Technical Report, Sloan School of Management, MIT, Cambridge, MA (1994).
[23] J.A.C. Resing, Polling systems and multitype branching processes, Queueing Systems 13 (1993) 409–426.
[24] H. Takagi, Queueing analysis of polling models: an update, in: *Stochastic Analysis of Computer and Communication Systems*, ed. H. Takagi (North-Holland, Amsterdam, 1990) pp. 267–318.
[25] H. Takagi, Queueing analysis of polling models: progress in 1990–1993, in: *Frontiers in Queueing: Models, Methods and Problems*, ed. J.H. Dshalalow (CRC Press, 1994).