COMMUN. STATIST.—STOCHASTIC MODELS, 15(2), 273-297 (1999)

Polling Systems with Periodic Server Routing in Heavy Traffic

R.D. van der Mei
AT&T Labs, P.O. Box 3030, Holmdel, NJ 07733, USA

Abstract

We consider polling systems with mixtures of exhaustive and gated service, and in which the server visits the queues according to a general polling table, in heavy traffic. We derive exact expressions for the expected delay at each of the queues (under heavy-traffic scalings), requiring the solution of a set of $M - N$ linear equations, where M is the length of the polling table and N is the number of queues. The results lead to closed-form expressions for the scaled expected delay under the commonly used star and elevator routing schemes, in a general parameter setting. In addition, the results reveal several insensitivity properties of the scaled expected delay with respect to the system parameters, providing new insights into the behavior of periodic polling systems in heavy traffic. The results also suggest simple and fast-to-evaluate approximations for the expected delay at each of the queues in stable polling systems. Numerical examples show that the approximations are very accurate in practical heavy-traffic scenarios.

Keywords: polling system, polling table, periodic routing, waiting time, heavy traffic.

1 INTRODUCTION

The basic polling system consists of a number of queues attended by a single server that visits the queues in cyclic order to provide service to the customers waiting at the queues. Polling models find applications in the areas of computer-communication systems, maintenance, manufacturing and production, amongst others (cf. [16, 21] for overviews). A natural extension of the cyclic order is periodic server routing, in which the order in which the server visits the queues is prescribed by a so-called polling table of finite length.

Only a few papers in the literature have been devoted to polling systems in heavy traffic. We refer to [8, 9, 14, 23, 25, 26] and references therein for results on (cyclic) polling systems in heavy traffic, and to [22] for periodic polling models with multiple servers. Exact analysis of polling models is only possible in some cases, and even in those cases numerical techniques have to be used to obtain performance metrics of interest, like the expected delay at the queues. For periodic polling models with gated and exhaustive service at each queue, the expected delay can be determined by solving (generally large) sets of linear equations [10, 1, 2], or by using iterative techniques [13, 7]. Systems with limited-type service disciplines require more computationally intensive techniques [3, 15]. However, a common disadvantage of all these numerical techniques is that their efficiency degrades significantly for heavily-loaded, large and highly asymmetric systems, while for these systems the proper operation of the system is particularly critical. In addition, numerical analysis in itself can only provide limited insight into the system behavior. These observations raise the importance an exact analysis of the delay in polling models in heavy traffic.

In this paper we study the expected delay in periodic polling models under heavy-traffic assumptions. We express the expected delay (under proper scalings) as the solution of a set of $M - N$ linear equations. The results lead to closed-from expressions for the scaled expected delay under polling schemes commonly used in industry, like the star and elevator polling schemes, in a general parameter setting. Moreover, the results show that the scaled expected delay figures depend on the system parameters only through a few aggregated system parameters. In addition, the results suggest simple

and fast-to-evaluate approximations for the expected delay at each of the queues in stable polling systems. Numerical experiments show that the approximations are very accurate in practical heavy-traffic scenarios.

In Section 2 the model is described and some notation is introduced. In Section 3 we give some preliminary results. In Section 4 we obtain exact expressions for the expected delay under heavy-traffic scalings. In Section 5 the implications of the results are discussed and illustrated by numerical examples. In Section 6 we propose and test approximations of the mean waiting times in stable polling systems. Section 7 contains some concluding remarks and addresses a number of topics for further research.

2 MODEL DESCRIPTION

Consider a system consisting of N infinite-buffer queues, $Q_1, \ldots, Q_N$. Customers arrive at Q_i according to a Poisson arrival process with rate λ_i. The total arrival rate is denoted by $\Lambda = \sum_{i=1}^{N} \lambda_i$. The first two moments of the service times at Q_i are denoted by b_i and $b_i^{(2)}$. Denote $\underline{b} = (b_1, \ldots, b_N)$ and $\underline{b}^{(2)} = (b_1^{(2)}, \ldots, b_N^{(2)})$. The first two moments of an arbitrary service time are denoted by $b = \sum_{i=1}^{N} \lambda_i b_i / \Lambda$ and $b^{(2)} = \sum_{i=1}^{N} \lambda_i b_i^{(2)} / \Lambda$. The load offered to Q_i is $\rho_i = \lambda_i b_i$, and the total offered load is equal to $\rho = \sum_{i=1}^{N} \rho_i$. A single server inspects the queues periodically according to a general polling table of length M, described by a mapping $T(\cdot)$, which is used such that the server visits the queues periodically in the order $T(1), T(2), \ldots, T(M), T(1), T(2), \ldots$. Following the approach in [2], a unique pseudo-queue is associated with each entry in the polling table. Denote by PQ_k the pseudo-queue associated with the k-th entry in the polling table; its corresponding queue has index $T(k)$. Customers which arrive at $Q_{T(k)}$ and are served at PQ_k are referred to as type-k customers. The moments at which the server arrives at PQ_k are referred to as the polling instants at PQ_k. Define a service period at PQ_k as the time between a polling instant at PQ_k and its successive departure from PQ_k. Define a cycle as the time interval between successive visits of the server to PQ_1. The service at each pseudo-queue is either according to the gated policy or the exhaustive policy. For ease of the discussion, we assume that pseudo-queues corresponding to the same queue have the same service strategy. Define

$E := \{i : Q_i \text{ is served exhaustively}\}$ and $G := \{i : Q_i \text{ is served according to the gated policy}\}$. At each queue the customers are served on a FIFO basis. After completing service at PQ_k the server proceeds to PQ_{k+1}, incurring a switch-over period whose first two moments are r_k and $r_k^{(2)}$. The first two moments of the total switch-over time per cycle are denoted by r and $r^{(2)}$. It is assumed throughout that $r > 0$. Denote $\underline{r} = (r_1, \ldots, r_M)$.

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 $\rho < 1$ [12]. Throughout, we assume that this condition is satisfied, and that the system is in steady state, unless indicated otherwise.

Denote by W_i the delay incurred by an arbitrary customer at Q_i. Our main interest is in the behavior of $E[W_i]$, the expected delay at Q_i, in heavy traffic. Throughout, $E[W_i]$ is considered as function of ρ. We assume that the arrival rates are parametrized as $\lambda_i = a_i\rho$, where relative arrival rates a_i remain fixed. It is known that $E[W_i]$, considered as a function of ρ, has a first-order pole at $\rho = 1$ (see Remark 4.1). Therefore, the analysis is oriented towards the determination of

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

the scaled expected delay at Q_i. The expected delay and the scaled expected delay at PQ_k are denoted by $E[W_k^{(PQ)}]$ and $\omega_k^{(PQ)}$, respectively, for $k = 1, \ldots, M$.

Finally we introduce some notation. Denote $\varphi_i := b_i$ for $i \in G$ and $\varphi_i := b_i/(1-\rho_i)$ for $i \in E$. Let τ_{ij} be the entry in the polling table corresponding to the next visit to Q_j after a departure from PQ_i, and let σ_{ij} be the entry corresponding to the last visit to Q_j prior to an arrival of the server at PQ_i $(i = 1, \ldots, M,\ j = 1, \ldots, N)$. I_E stands for the indicator function on the event E. Let $\underline{1}$ be the vector whose i-th components equals 1 for all i. Indices i corresponding to queues and pseudo-queues should be read as $[(i-1)\ mod\ N]+1$ and $[(i-1)\ mod\ M]+1$, respectively.

3 PRELIMINARIES

Let X_k be the number of customers at PQ_k at a polling instant at PQ_k. For a customer served at PQ_k, we define the waiting time at PQ_k to be

the time between its arrival into the system and the moment at which the customer starts service at PQ_k. The expected waiting time at PQ_k can be expressed in terms of the first two moments of X_k as follows (cf. [20]): For $T(k) \in G$,

$$E[W_k^{(PQ)}] = \frac{Var[X_k] + (E[X_k])^2 - E[X_k]}{2\lambda_{T(k)}E[X_k]}(1 + \rho_{T(k)}), \tag{2}$$

and for $T(k) \in E$,

$$E[W_k^{(PQ)}] = \frac{Var[X_k] + (E[X_k])^2 - E[X_k]}{2\lambda_{T(k)}E[X_k]} + \frac{\lambda_{T(k)}b_{T(k)}^{(2)}}{2(1 - \rho_{T(k)})}. \tag{3}$$

Hence, to obtain expressions for $E[W_k^{(PQ)}]$, we need to quantify $E[X_k]$ and $Var[X_k]$.

To obtain expressions for $E[X_k]$, it is convenient to relate $E[X_k]$ to $E[V_k]$, where V_k stands for the duration of a service period of the server to PQ_k ($k = 1, \ldots, M$). To this end, it is readily verfied that, for $k = 1, \ldots, M$,

$$E[V_k] = \varphi_{T(k)}E[X_k]. \tag{4}$$

The variables $E[V_k]$ can be obtained by solving the following set of linear equations (cf. also [4]): For $k = 1, \ldots, M$,

$$E[V_k] = \lambda_{T(k)}\varphi_{T(k)}\left[\sum_{j=l_k+1}^{k-1}(r_j + E[V_j]) + r_{l_k} + E[V_{l_k}]I_{\{T(k)\in G\}}\right], \tag{5}$$

with $l_k := \sigma_{k\ T(k)}$, supplemented with the balancing equations

$$\sum_{m:T(m)=i} E[V_m] = \rho_i\frac{r}{1-\rho} \quad (i = 1, \ldots, N). \tag{6}$$

Notice that both $E[V_k]$ and $E[X_k]$ possess a first-order pole at $\rho = 1$. Therefore, we define, for $k = 1, \ldots, M$,

$$v_k := \lim_{\rho\uparrow 1}(1-\rho)E[V_k], \quad x_k := \lim_{\rho\uparrow 1}(1-\rho)E[X_k]. \tag{7}$$

Using equations (4)-(7) it follows that the variables v_k, and hence x_k, $k = 1, \ldots, M$, are (uniquely) determined by the following set of equations: For $k = 1, \ldots, M$, $i = 1, \ldots, N$,

$$v_k = \lambda_{T(k)}\varphi_{T(k)}\left[\sum_{j=l_k+1}^{k-1} v_j + v_{l_k}I_{\{T(k)\in G\}}\right], \quad \sum_{m:T(m)=i} v_m = \rho_i r, \tag{8}$$

where the parameter values in (8) are evaluated at $\rho = 1$. It is readily verified from equations (4)-(8) that the variables $E[V_k]$ and $E[X_k]$, and hence also x_k and v_k, can be effectively determined (by eliminating redundant equations) by solving a set of $M - N$ linear equations.

The derivation of expressions for $Var[X_k]$ is more involved. To this end, we use the Descendant Set Approach (DSA). The DSA considers an arbitrary fixed polling instant at PQ_k, called the reference point P_k^*. The idea is to consider an arbitrary customer $C_{i,c}$ that was served at PQ_i c cycles ago and to obtain recursive relations for $A_{(i,c),k}$, defined as the number of type-k "descendants" $C_{i,c}$ has at P_k^*. By conditioning on the number of so-called immediate children of $C_{i,c}$, we obtain recursive relations for the distribution of $A_{(i,c),k}$. These relations lead to the following recursive relations for $\alpha_{(i,c),k}$ and $\alpha^{(2)}_{(i,c),k}$, i.e., the first two factorial moments of $A_{(i,c),k}$ (cf. [13] for an extensive discussion of the use of the DSA for the present model): For $T(i) \in G$, $k = 1, \ldots, M$, $c = 0, 1, \ldots$,

$$\alpha_{(i,c),k} = b_{T(i)} \left[\sum_{j:\tau_{ij}>i} \lambda_j \alpha_{(\tau_{ij},c),k} + \sum_{j:\tau_{ij}\leq i} \lambda_j \alpha_{(\tau_{ij},c-1),k} \right], \tag{9}$$

$$\alpha^{(2)}_{(i,c),k} = \frac{b^{(2)}_{T(i)}}{b^2_{T(i)}} \alpha^2_{(i,c),k} + b_{T(i)} \left[\sum_{j:\tau_{ij}>i} \lambda_j \alpha^{(2)}_{(\tau_{ij},c),k} + \sum_{j:\tau_{ij}\leq i} \lambda_j \alpha^{(2)}_{(\tau_{ij},c-1),k} \right]. \tag{10}$$

Recursive relations for the case $T(i) \in E$ can be obtained similarly. The initial conditions are $\alpha_{(k,0),k} := 1$, $\alpha_{(i,0),k} := 0$ $(i = k+1, \ldots, M)$, $\alpha_{(i,-1),k} := 0$ $(i = 1, \ldots, k-1)$, and $\alpha^{(2)}_{(k,0),k} := 1$, $\alpha^{(2)}_{(i,0),k} := 0$ $(i = k+1, \ldots, M)$, $\alpha^{(2)}_{(i,-1),k} := 0$ $(i = 1, \ldots, k-1)$.

The variables $E[X_k]$ and $Var[X_k]$ can be expressed in terms of the variables $\alpha_{(i,c),k}$ and $\alpha^{(2)}_{(i,c),k}$ as follows (cf. [13]): For $k = 1, \ldots, M$,

$$E[X_k] = \sum_{i=1}^{M} r_i \sum_{c=0}^{\infty} \left[\sum_{j:\tau_{ij}>i} \lambda_j \alpha_{(\tau_{ij},c),k} + \sum_{j:\tau_{ij}\leq i} \lambda_j \alpha_{(\tau_{ij},c-1),k} \right], \tag{11}$$

and

$$Var[X_k] = \sum_{i=1}^{M} r_i \sum_{c=0}^{\infty} \left[\sum_{j:\tau_{ij}>i} \lambda_j \alpha^{(2)}_{(\tau_{ij},c),k} + \sum_{j:\tau_{ij}\leq i} \lambda_j \alpha^{(2)}_{(\tau_{ij},c-1),k} \right] \tag{12}$$

$$+\sum_{i=1}^{M}(r_i^{(2)}-r_i^2)\sum_{c=0}^{\infty}\left[\sum_{j:\tau_{ij}>i}\lambda_j\alpha_{(\tau_{ij},c),k}+\sum_{j:\tau_{ij}\le i}\lambda_j\alpha_{(\tau_{ij},c-1),k}\right]^2. \tag{13}$$

For the analysis, we also need to conduct the recursion in a different way. To this end, for $k=1,\ldots,M$, let $d_{jk}:=b_j$ for $j\in G$ and $d_{jk}:=I_{\{T(k)\neq j\}}b_j/(1-\rho_j)$ for $j\in E$. Then by conditioning on the number of customers present at the beginnings of the service periods since the last visit to $Q_{T(k)}$ prior to P_k^*, we obtain the following relations between the first moments (relations for the second moments are not needed here) of $A_{(i,c),j}$ and $A_{(i,c),k}$ with $j\neq k$ (cf. [23] for details): For $i,k=1,\ldots,M$, $c=0,1,\ldots$,

$$\alpha_{(i,c),k}=\lambda_{T(k)}\sum_{j=\sigma_{k\,T(k)}}^{k-1}\left[d_{jk}\alpha_{(i,c),j}I_{\{j<k\}}+d_{jk}\alpha_{(i,c-1),j}I_{\{j\ge k\}}\right]. \tag{14}$$

It is useful to express relations (9) and (14) in matrix notation. Let $\underline{\alpha}_{(\cdot,c),k}$ be the vector whose i-th element is $\alpha_{(i,c),k}$, for $i=1,\ldots,M$. Moreover, let $\mathbf{P}_i$ be the M by M matrix whose (j,k)-th element equals $I_{\{j=k\}}$ for $j\neq i$, while the (i,τ_{ij})-th element equals $b_{T(i)}\lambda_j$ if $T(i)\in G$, and $I_{\{j\neq T(i)\}}\lambda_j b_{T(i)}/(1-\rho_{T(i)})$ if $T(i)\in E$, and all other components of the i-th row are 0. Define $\mathbf{M}:=\mathbf{P}_1\cdots\mathbf{P}_M$, and let $\underline{q}_k$ be the vector whose i-th element is $\alpha_{(i,0),k}$, for $i,k=1,\ldots,M$. Then it is readily verified that $\underline{q}_k=\mathbf{P}_1\cdots\mathbf{P}_{k-1}\underline{e}_k$, and that the recursive relations (9) can be expressed as follows: For $k=1,\ldots,M$,

$$\underline{\alpha}_{(\cdot,0),k}=\underline{q}_k,\quad \underline{\alpha}_{(\cdot,c),k}=\mathbf{M}\underline{\alpha}_{(\cdot,c-1),k}=\mathbf{M}^c\underline{q}_k,\quad c=1,2,\ldots. \tag{15}$$

To express (14) in matrix notation, let $\underline{\alpha}_{(i,c),\cdot}$ be the vector whose k-th element is $\alpha_{(i,c),k}$, and let $\hat{\mathbf{P}}_k$ be the matrix whose (i,j)-th element equals $I_{\{i=j\}}$ for $i\neq k$, while the (k,j)-th element of $\hat{\mathbf{P}}_k$ is given by $\lambda_{T(k)}d_{jk}$ for $j=\sigma_{k\,T(k)},\ldots,k-1$, and all other components of the k-th row are 0. Define $\hat{\mathbf{M}}:=\hat{\mathbf{P}}_M\cdots\hat{\mathbf{P}}_1$, and let $\hat{\underline{q}}_i$ be the vector whose k-th element is $\alpha_{(i,0),k}$. Then $\hat{\underline{q}}_i=\hat{\mathbf{P}}_M\cdots\hat{\mathbf{P}}_{i+1}\underline{e}_i$, and relations (14) can be expressed as follows: For $i=1,\ldots,M$,

$$\underline{\alpha}_{(i,0),\cdot}=\hat{\underline{q}}_i,\quad \underline{\alpha}_{(i,c),\cdot}=\hat{\mathbf{M}}\underline{\alpha}_{(i,c-1),\cdot}=\hat{\mathbf{M}}^c\hat{\underline{q}}_i,\quad c=1,2,\ldots. \tag{16}$$

4 ANALYSIS

The variables $\alpha_{(i,c),k}$ are determined by both sets of relations (15) and (16), which constitute a set of first-order homogeneous difference equations. The

theory of difference equations shows that the variables $\alpha_{(i,c),k}$ can be solved explicitly if the eigenvalues and eigenvectors of $\mathbf{M}$ ($\hat{\mathbf{M}}$) are known. In general, however, the eigenvalues and eigenvectors of $\mathbf{M}$ ($\hat{\mathbf{M}}$) are unknown for $\rho < 1$. Nevertheless, to analyze the system behavior in the limiting case $\rho \uparrow 1$, there is no need to obtain all eigenvalues and eigenvectors of $\mathbf{M}$ ($\hat{\mathbf{M}}$) and then let $\rho \uparrow 1$. More precisely, since $E[X_k]$ and $Var[X_k]$ are known to tend to infinity when $\rho \uparrow 1$, it follows from (11)-(13) that the heavy-traffic behavior of $E[X_k]$ and $Var[X_k]$ is determined by the "dominant" behavior of the sequences $\{\alpha_{(i,c),k}, c = 0, 1, \ldots\}$ and $\{\alpha^{(2)}_{(i,c),k}, c = 0, 1, \ldots\}$, which appears to be relatively easy to analyze.

Lemmas 4.1 and 4.2 below are useful in the analysis. The proofs, which proceed along similar lines as discussed in [23], are omitted for compactness of the presentation.

Lemma 4.1

The matrices $\mathbf{M}$ *and* $\hat{\mathbf{M}}$ *have respective maximal eigenvalues* γ *and* $\hat{\gamma}$ *which are real-valued, positive, have multiplicity 1, and have associated right and left eigenvectors* $\underline{u}$, $\underline{w}$, *and* $\underline{\hat{u}}$, $\underline{\hat{w}}$, *respectively. If these are normalized so that* $\underline{u}^\top \underline{w} = \underline{u}^\top \underline{1} = 1$, $\underline{\hat{u}}^\top \underline{\hat{w}} = \underline{\hat{u}}^\top \underline{1} = 1$, *then*

$$\mathbf{M}^c = \gamma^c \underline{u}\, \underline{w}^\top + \mathbf{R}^c, \quad \hat{\mathbf{M}}^c = \hat{\gamma}^c \underline{\hat{u}}\, \underline{\hat{w}}^\top + \hat{\mathbf{R}}^c, \tag{17}$$

where there exist $K < \infty$ *and* $\bar{\gamma}$ $(0 < \bar{\gamma} < \gamma, \hat{\gamma})$, *such that all entries of* $\mathbf{R}^c$ *and* $\hat{\mathbf{R}}^c$ *are strictly smaller than* $K\bar{\gamma}^c$.

Lemma 4.2

(1) *If* $\rho < 1$, *then* $\gamma, \hat{\gamma} < 1$, *and if* $\rho = 1$ *then* $\gamma = \hat{\gamma} = 1$.
(2) *If* $\rho = 1$, *then* $\underline{u}$ *is proportional to* $(b_{T(1)}, \ldots, b_{T(M)})$.
(3) *If* $\rho = 1$, *then* $\underline{\hat{u}}$ *is proportional to* $(x_1, \ldots, x_M)$.

Remark 4.1

From equation (9) it is readily verified that the series $\sum_{c=0}^{\infty} \alpha_{(i,c),k}$, considered as a function of ρ, possesses a first-order pole at $\rho = 1$. Equation (11) then implies that $E[X_k]$ has a first-order pole at $\rho = 1$, which also follows from equations (4)-(6). Then, equation (15) and Lemmas 4.1 and 4.2 imply that the series $\sum_{c=0}^{\infty} \alpha^2_{(i,c),k}$ also has a first-order pole at $\rho = 1$. Similarly,

the series $\sum_{c=0}^{\infty} \alpha^{(2)}_{(i,c),k}$ possesses a second-order pole at $\rho = 1$ (see equations (34)-(37) in the Appendix). These observations imply that $Var[X_k]$ has a second-order pole at $\rho = 1$, see equations (12)-(13). This implies, by using (2) and (3), that $E[W_k^{(PQ)}]$ $(k = 1, \ldots, M)$ has a first-order pole at $\rho = 1$, which, in turn, implies that $E[W_i]$ $(i = 1, \ldots, N)$ also has a first-order pole at $\rho = 1$.

Theorem 1

For $i, j, k, l = 1, \ldots, M$,

$$(1) \lim_{\rho \uparrow 1} \frac{\sum_{c=0}^{\infty} \alpha_{(i,c),k}}{\sum_{c=0}^{\infty} \alpha_{(j,c),l}} = \frac{b_{T(i)}}{b_{T(j)}} \frac{x_k}{x_l}; \quad (2) \lim_{\rho \uparrow 1} \frac{\sum_{c=0}^{\infty} \alpha^{(2)}_{(i,c),k}}{\sum_{c=0}^{\infty} \alpha^{(2)}_{(j,c),l}} = \frac{b_{T(i)}}{b_{T(j)}} \frac{x_k^2}{x_l^2}. \tag{18}$$

Proof: Part 1 follows from Lemmas 4.1 and 4.2 and using the continuity of the eigenvalues and eigenvectors at $\rho = 1$. The proof of Part 2 is given in the Appendix.

Theorem 2

For $k, l = 1, \ldots, M$,

$$(1) \lim_{\rho \uparrow 1} \frac{E[X_k]}{E[X_l]} = \frac{x_k}{x_l}; \quad (2) \lim_{\rho \uparrow 1} \frac{Var[X_k]}{Var[X_l]} = \frac{x_k^2}{x_l^2}. \tag{19}$$

Proof: Part 1 follows directly from the first part of Theorem 1 and equation (11). Part 2 follows from the second part of Theorem 1, and the observation that (13) is dominated by (12) in the limiting case (see Remark 4.1).

Theorem 3

For $k, l = 1, \ldots, M$,

$$\frac{\omega_k^{(PQ)}}{\omega_l^{(PQ)}} = \frac{v_k(1 + \rho_{T(k)} I_{\{T(k) \in G\}})/\lambda_{T(k)}\varphi_{T(k)}}{v_l(1 + \rho_{T(l)} I_{\{T(l) \in G\}})/\lambda_{T(l)}\varphi_{T(l)}}. \tag{20}$$

Proof: The result follows from Theorem 2 and equations (2) and (3).

The waiting times at the pseudo-queues can be related to the waiting times at the queues by conditioning on π_k, defined as the fraction of customers which arrive at $Q_{T(k)}$ that are served at PQ_k $(k = 1, \ldots, M)$. Using this definition, it is readily seen that, for $k = 1, \ldots, M$,

$$\pi_k = \frac{v_k}{\sum_{j:T(j)=T(k)} v_j} = \frac{v_k}{\rho_{T(k)} r}, \tag{21}$$

where the second equality follows from (8). Hence, for $i = 1, \ldots, N$,

$$\omega_i = \sum_{k:T(k)=i} \pi_k \omega_k^{(PQ)} = \frac{1}{\rho_i r} \sum_{k:T(k)=i} v_k \omega_k^{(PQ)}. \tag{22}$$

The following result gives an expression for the ratios between scaled expected waiting times at the queues.

Theorem 4

For $i, j = 1, \ldots, N$,

$$\frac{\omega_i}{\omega_j} = \frac{\eta_i/\rho_i^2}{\eta_j/\rho_j^2} \frac{\sum_{k:T(k)=i} v_k^2}{\sum_{k:T(k)=j} v_k^2}, \tag{23}$$

where $\eta_i := 1 + \rho_i$ for $i \in G$ and $\eta_i := 1 - \rho_i$ for $i \in E$.

Proof: The results follows directly from Theorem 3 and relation (22).

From Theorem 4, the scaled expected delays are known up to some scaling factor. This factor can be obtained by using the so-called pseudo-conservation law (PCL), i.e., an exact expression for $\sum_{i=1}^N \rho_i E[W_i] = E[U]$, where U stands for the total amount of waiting work in the system (cf. [4]). Denoting $u := \lim_{\rho \uparrow 1} (1-\rho) E[U]$, the PCL leads to an exact expression for $\sum_{i=1}^N \rho_i \omega_i = u$, requiring only the solution of the variables v_k, $k = 1, \ldots, M$.

Theorem 5

$$\sum_{i=1}^N \rho_i \omega_i = \frac{b^{(2)}}{2b} + \sum_{k=1}^N \left(\sum_{m=\sigma_{1k}+1}^M \rho_k v_m + \rho_k v_{\sigma_{1k}} I_{\{k \in G\}} \right). \tag{24}$$

Proof: Follows from [4], definition (1) and several straightforward arguments.

Theorem 6 (Main result)

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

$$\omega_i = \frac{(\eta_i/\rho_i^2) \sum_{k:T(k)=i} v_k^2}{\sum_{j=1}^N (\eta_j/\rho_j) \sum_{k:T(k)=j} v_k^2} \tag{25}$$

$$\times \left[\frac{b^{(2)}}{2b} + \sum_{k=1}^{N} \left(\sum_{m=\sigma_{1k}+1}^{M} \rho_k v_m + \rho_k v_{\sigma_{1k}} I_{\{k \in G\}} \right) \right]. \tag{26}$$

Proof: The result follows directly from Theorems 4 and 5.

Recall that the determination of the expression in Theorem 6 (for all i simultaneously) generally requires the solution of a set of $M - N$ linear equations to determine the variables v_k, from equation (8). We emphasize that all terms in Theorems 1-6 corresponding to arrival rates and loads at the queues have to be evaluated at $\rho = 1$.

Remark 4.2

The results presented in this paper generalize the results in [23], where we considered the special case of cyclic polling, i.e., $T = (1, 2, \ldots, N)$. In that case, the balance equations (5)-(8) can be solved explicitly; it is readily verified that $E[V_k] = \rho_k r/(1-\rho)$ and $E[X_k] = \lambda_k(1-\rho_k I_{\{k \in E\}})r/(1-\rho)$, which directly implies $v_k = \rho_k r$ and $x_k = \lambda_k(1-\rho_k I_{\{k \in E\}})$. Hence, for cyclic polling models the scaled expected delay can be expressed in closed form, whereas for non-cyclic polling models a set of $M - N$ equations has to solved to obtain the variables v_k. This observation addresses a fundamental difference between cyclic and non-cyclic periodic polling models: in the case of cyclic polling all customers that arrive at Q_i are served at the *same* pseudo-queue (namely, Q_i itself), whereas in non-cyclic polling customers arriving at the same queue (say Q_i) may be served at *different* pseudo-queues (namely, any PQ_k for which $T(k) = i$), depending on the position of the server upon the arrival of a customer. In other words, in cyclic polling the arrival process at each (pseudo-)queue is independent of the position of the server, whereas in non-cyclic polling the arrival process at a pseudo-queue generally does depend on the position of the server. This dependence *inherently* leads to a set of linear equations for $E[V_k]$ (and $E[X_k]$, x_k and v_k) that can not be solved in a general closed-form expression.

Remark 4.3

Another key difference between cyclic and non-cyclic polling is the influence of the individual switch-over times on the performance metrics. For

cyclic polling, $E[V_k]$ and $E[X_k]$ depend on the individual switch-over time distributions only through r, i.e., the total expected switch-over time per cycle, whereas equation (5) implies that for general polling tables $E[V_k]$ and $E[X_k]$ generally do depend on the individual mean switch-over times. To explain this, recall that for cyclic polling all customers arriving at a queue are served at the same (pseudo-)queue, whereas for non-cyclic polling customers arriving at a queue (say Q_i) during some switch-over time are served at the pseudo-queue (say PQ_k, with $T(k) = i$) corresponding to the first visit to Q_i after that switch-over time. Hence, the mean number of customers served at a particular pseudo-queue generally depends on the mean individual switch-over times.

In this perspective, Theorem 6 implies that the dependence of $E[V_k]$ and $E[X_k]$ on the individual mean switch-over times, for general polling tables, vanishes in when the system reaches saturation: v_k and x_k depend on the switch-over times only through r, which follows directly from equations (5)-(8). In this way, the impact of the individual mean switch-over times on $E[V_k]$ and $E[X_k]$, and also on the expected delay, is of "lower order" in the limiting case. Insensitivity properties with respect to the system parameters in heavy traffic are discussed in more detail in the next section.

Remark 4.4

The difference between cyclic and non-cyclic polling models also manifests itself when the switch-over times become negligible. For stable non-cyclic polling models, the expected waiting times generally depend on *how* the individual switch-over time distributions converge to 0. In fact, in the case of zero switch-over times the expected delay may not even be uniquely determined (cf. [6, 11]), whereas the case of cyclic polling models similar complications do not occur. However, in the limiting case $\rho \uparrow 1$, we observe that since v_k is linear in r, Theorem 6 implies that ω_i converges to a unique value when the switch-over times vanish, regardless of how the individual switch-over time distributions tend to 0. Apparently, the differences in the expected delays caused by different limiting regimes (for $r \to 0$) vanish when the system reaches saturation.

5 IMPLICATIONS OF THE RESULTS

In this section we discuss a number of implications of the results discussed in Section 4. First, it is shown that the results lead to explicit expressions the scaled expected delay under the star and elevator polling schemes. Second, the scaled expected delays appear to depend on the system parameter only through a few aggregated system parameters. Numerical examples are given to demonstrate the validity of these observations.

5.1 Star and elevator polling

In the case of star polling, the server visits the queues in the order $T = (1, 2, 1, 3, \ldots, 1, N)$.

Corollary 5.1 (Star polling)
For the star polling configuration, ω_i $(i = 1, \ldots, N)$ *is given by equation* (25), *with* $\sigma_{11} = 2N - 1$, $\sigma_{1k} = 2k$ $(k = 2, \ldots, N)$, *and*
(1) *if* k *even, then* $v_k = \rho_{T(k)} r$;
(2) *if* k *odd, then*

$$v_k = \frac{\rho_{T(k-1)} \rho_1 r}{1 - \rho_1} \quad (1 \in E), \qquad v_k = \frac{r \sum_{j=1}^{N-1} \rho_1^j \rho_{k-j}}{1 - \rho_1^{N-1}} \quad (1 \in G). \tag{27}$$

In the case of elevator polling, the queues are visited in the order $T = (1, 2, \ldots, N, N, \ldots, 2, 1)$.

Corollary 5.2 (Elevator polling)
For the elevator polling order, ω_i $(i = 1, \ldots, N)$ *is given by equation* (25), *with* $\sigma_{1k} = 2N - k + 1$ $(k = 1, \ldots, N)$, *and*

$$v_k = \frac{\rho_k r \sum_{i=1}^{k-1} \rho_i}{1 - \rho_k}, \qquad v_{2N-k+1} = \frac{\rho_k r \sum_{i=k+1}^{N} \rho_i}{1 - \rho_k} \quad (k \in E), \tag{28}$$

and

$$v_k = \frac{\rho_k r \sum_{i=1}^{k} \rho_i}{1 + \rho_k}, \qquad v_{2N-k+1} = \frac{\rho_k r \sum_{i=k}^{N} \rho_i}{1 + \rho_k} \quad (k \in G). \tag{29}$$

We emphasize that the results presented here do not pose any additional symmetry restriction on the parameters, making the results applicable in a

general parameter setting. To the best of the author's knowledge, similar closed-form expressions have not been presented in the literature in this general context.

5.2 Insensitivity properties

Theorem 6 provides new insights into how the expected waiting times depend on the system parameters. In this section we discuss these properties in more detail and provide numerical examples to illustrate the validity of the results.

Corollary 5.3

ω_i $(i = 1, \ldots, N)$ depends on the higher moments individual service-time distributions only through $b^{(2)}$, i.e., the second moment of the service time of an arbitrary customer (weighted proportionally to the arrival rates).

Notice that in stable polling systems the expected delay generally depends on the first two moments of each of the individual service-time distributions, even when $b^{(2)}$ is kept fixed. Therefore, Corollary 5.3 implies that the dependence of the expected delay on the second moments of the individual service times, for given $b^{(2)}$, vanishes when ρ tends to 1.

To illustrate this, we consider the model with the following parameters (referred to as model 1): $N = 4$; $M = 6$; $T = (1, 2, 1, 3, 1, 4)$; $G = \{2, 3, 4\}$; $E = \{1\}$; all arrival rates are equal; all switch-over times are exponentially distributed with mean 0.05; the service times at Q_1 have a gamma distribution with $b_1 = 1$ and $b_1^{(2)} = 11$, and the service times at all other queues are deterministic with mean 1. For comparison, we also consider model 2, which is identical to model 1, except that the service times at Q_4 are gamma-distributed with first two moments 1 and 11, while the service time at queues 1, 2 and 3 are deterministic with mean 1. Note that for both models we have $b^{(2)} = 3.5$. Hence, according to Theorem 6, the expected waiting times at each of the queues should tend to infinity at the same rate when $\rho \uparrow 1$. Denoting the expected waiting time at Q_i in model m by $E[W_i^{(m)}]$, we define

$$\Delta_{m/n} = \Delta = \max_{i=1,\ldots,N} \text{abs}\left(\frac{E[W_i^{(m)}] - E[W_i^{(n)}]}{E[W_i^{(n)}]}\right) \times 100\%, \tag{30}$$

Table 5.1. Expected waiting times for different values of the load: influence of second moments of the individual service times.

ρ	$\underline{b}^{(2)} = (11,1,1,1)$				$\underline{b}^{(2)} = (1,1,1,11)$				Δ
0.70	2.1	5.4	5.4	5.4	2.1	4.9	5.3	6.1	11.1
0.80	3.0	9.3	9.3	9.3	3.0	8.5	9.2	10.3	10.8
0.90	5.5	21.2	21.2	21.2	5.5	20.0	21.0	22.5	6.1
0.95	10.3	44.9	44.9	44.9	10.3	43.5	44.7	46.5	3.6
0.98	24.6	116.1	116.1	116.1	26.6	114.6	115.9	118.0	1.6
0.99	48.4	234.9	234.9	234.9	48.4	233.3	234.6	236.8	0.8

i.e., the relative difference between $E[W_i^{(m)}]$ and $E[W_i^{(n)}]$ (where the maximum is taken over all $i = 1, \ldots, N$). Table 5.1 shows the expected waiting time at each of the queues for different values of ρ for models 1 and 2, and their relative differences.

Table 5.1 illustrates that the relative difference between the corresponding expected waiting times in models 1 and 2 indeed vanishes when $\rho \uparrow 1$, as expected on the basis of Theorem 6. This illustrates that the dependencies of the second moments of the individual service times (i.e., $b_i^{(2)}, i = 1, \ldots, N$), with given $b^{(2)}$, indeed vanish when the system reaches saturation.

Corollary 5.4

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

(1) *ω_i depends on the individual switch-over time distributions only through r, the expected total switch-over time per cycle.*

(2) *ω_i, considered as a function of r, is the sum of a constant and a linear function of r.*

Corollary 5.4 is known to be not generally true for systems with $\rho < 1$, for which the mean delay generally depends on the first two moments of each of the individual switch-over times. In this context, an interesting observation is made by Srinivasan et al. [19], who show that for cyclic polling models with deterministic switch-over times, the (complete) waiting-time distributions depend on the switch-over times only through r, even for sta-

Table 5.2. Expected waiting times for different values of the load: influence of the individual switch-over times.

ρ	$\underline{r} = (1,1,1,1,1,1)$ (exp)				$\underline{r} = (6,0,0,0,0,0)$ (Erlang)				Δ
0.70	4.1	15.1	15.1	15.1	4.9	15.0	15.0	15.2	19.5
0.80	5.8	23.4	23.4	23.4	6.3	23.4	23.3	23.5	8.6
0.90	10.9	48.4	48.4	48.4	11.1	48.4	48.4	48.4	1.8
0.95	20.9	98.4	98.4	98.4	21.0	98.4	98.4	98.4	0.5
0.98	50.9	248.4	248.4	248.4	50.9	248.4	248.4	248.4	0.0
0.99	100.9	498.4	498.4	498.4	100.9	498.4	498.4	498.4	0.0

ble polling systems. However, similar insensitivity properties are not generally true for non-cyclic polling models and for cyclic polling models with non-deterministic switch-over times.

In this context, Corollary 5.4 shows that the sensitivity of the expected delay with respect to the individual switch-over time distributions vanishes when the system reaches saturation, and as such can be considered to be of "lower order" in the limiting case.

To illustrate this, we consider several variants of model 1, with exponential service at all queues, and which only differ in their switch-over time distributions. For model 3, all switch-over times are exponentially distributed with mean 1. For model 4, all switch-over times are 0, except for the switch-over time from PQ_1 to PQ_2, which consists of 6 independent exponential phases with mean 1 and is therefore Erlang distributed with mean $r_1 = 6$ and $r_1^{(2)} = 42$. Note that the total switch-over time per cycle in models 3 and 4 are identically distributed. Table 5.2 shows the expected waiting times for models 3 and 4 for different values of ρ. The maximal relative difference between the corresponding expected waiting times in the different models is determined according to (30).

Table 5.2 illustrates that the relative difference between the corresponding expected waiting times in models 3 and 4 tends to 0 when $\rho \uparrow 1$, which we expected on the basis of Theorem 6. This illustrates that the dependencies of the expected waiting times with respect to the individual switch-over

Table 5.3. Expected waiting times for different values of the load: influence of the second moments of the switch-over times.

ρ	$\underline{r} = (6,0,0,0,0,0)$ (Erlang)				$\underline{r} = (6,0,0,0,0,0)$ (exp)				Δ
0.70	4.9	15.0	15.0	15.2	6.4	17.4	17.8	18.5	30.6
0.80	6.3	23.4	23.3	23.5	7.6	25.9	26.2	26.8	20.6
0.90	11.1	48.4	48.4	48.4	12.1	51.2	51.3	51.7	9.0
0.95	21.0	98.4	98.4	98.4	21.8	101.3	101.4	101.6	3.8
0.98	50.9	248.4	248.4	248.4	51.6	251.4	251.5	251.6	1.4
0.99	100.9	498.4	498.4	498.4	101.6	501.5	501.5	501.5	0.7

times, with given total switch-over time distribution per cycle, vanish in heavy traffic.

Corollary 5.4 also suggests that the influence of the seconds moments of the switch-over time distributions disappear when ρ tends to 1. To illustrate this, we compare the expected waiting times in model 4 with those in model 5, which is similar to model 4, with the exception that the switch-over times from PQ_1 to PQ_2 are exponentially (instead of Erlang) distributed with mean 6. In this way, models 4 and 5 have the same mean total switch-over times per cycle ($r = 6$), but the variability of the switch-over times in model 5 (where $r^{(2)} = 72$) is larger than those in model 4 (where $r^{(2)} = 42$). Table 5.3 shows the expected waiting times in models 4 and 5 for different values of ρ. The relative differences are computed according to (30).
Table 5.3 illustrates that the relative difference in the corresponding expected waiting times in identical models which only differ in the variability of the switch-over times vanishes when the system reaches saturation.

Notice that in all cases the expected waiting times in model 5 are larger than those in model 4. This is due to the fact that the variability of the total switch-over time per cycle in model 5 is larger than in model 4, which motivates why in al cases the expected waiting times in model 5 exceed those in model 4.

Table 6.1. Exact and approximated expected waiting times for different values of the load.

ρ	$E[W_1]$	$E[W_1^{(app)}]$	$err_1\%$	$E[W_2]$	$E[W_2^{(app)}]$	$err_2\%$
0.70	4.1	3.3	18.7	15.1	16.7	10.4
0.80	5.8	5.0	13.8	23.4	25.0	6.8
0.90	10.9	10.0	8.2	48.4	50.0	3.3
0.95	20.9	20.0	4.3	98.4	100.0	1.6
0.98	50.9	50.0	1.8	248.4	250.0	0.6
0.99	100.9	100.0	0.9	498.4	500.0	0.3

6 APPROXIMATION

Theorem 6 suggests the following approximation for the expected waiting times at each of the queues in stable systems: For $\rho < 1$, $i = 1, \ldots, N$,

$$E[W_i^{(app)}] = \frac{1}{1-\rho} \times \tag{31}$$

$$\frac{(\eta_i/\rho_i^2)\sum_{k:T(k)=i} v_k^2}{\sum_{j=1}^N (\eta_j/\rho_j)\sum_{k:T(k)=j} v_k^2} \left[\frac{b^{(2)}}{2b} + \sum_{k=1}^{N} \left(\sum_{m=\sigma_{1k}+1}^{M} \rho_k v_m + \rho_k v_{\sigma_{1k}} I_{\{k \in G\}} \right) \right], \tag{32}$$

where all parameters in (32) are evaluated at $\rho = 1$. In general, the approximation requires the solution of the set of $M - N$ linear equations to determine the variables v_k. Based on Theorem 6, the approximation of the expected waiting times is known to be asymptotically exact for $\rho \uparrow 1$, in the sense that $\lim_{\rho \uparrow 1} E[W_i^{(app)}]/E[W_i] = 1$ for all i. We have performed numerical experiments to investigate the accuracy of the approximations for different values of the system load. The results are outlined below. Define the relative error of the approximated mean waiting time at Q_i as follows: For $i = 1, \ldots, N$,

$$err_i\% = \text{abs}\left(\frac{E[W_i^{(app)}] - E[W_i]}{E[W_i]} \right) \times 100\%. \tag{33}$$

Table 6.1 shows the exact and approximated expected waiting times at Q_1 and Q_2 for model 3 (defined in section 5) for different values of ρ.

Table 6.1 shows that the relative error decreases when the system load increases, as expected. Moreover, we observe that the approximations are accurate when the system load is 80-90% or more. Notice that the model considered has a star polling configuration, so that from Corollary 5.1 the approximations are given in closed form.

To test the accuracy of the approximation for a highly asymmetric system, consider the following model: $N = 4$; $M = 8$; $T = (1, 2, 1, 3, 4, 2, 1, 3)$; $G = \{1, 4\}$; $E = \{2, 3\}$; the ratio between the arrival rates is 5:1:1:1; the service times are exponentially distributed with means $\underline{b} = (1, 1, 5, 1)$; the switch-over times are exponentially distributed with $r_8 = 5.0$, whereas all other switch-over times have mean 0.5. The system is therefore highly asymmetric in the arrival rates, service times and switch-over times. Table 6.2 below shows the mean waiting times, the approximations and the relative error, for Q_1 and Q_4.
The results in Table 6.2 show that the accuracy of the approximations may decrease when the system is highy asymmetric. In Table 6.2, Q_4 represents the "worst case" in the sense that the relative error of the approximation of the mean waiting time at Q_4 is the maximum relative error over all queues (for all considered values of ρ). However, we observe that the results may still be considered acceptable when the load is 90% or more.

The results in Tables 6.1 and 6.2 imply that for most practical cases the approximation can be used with good confidence. This implication stems from the fact that in practice heavy load is the main region of interest. We emphasize that the example presented here is highly asymmetric in the arrival rates, the service rates and the switch-over times, and that the approximations are considerably more accurate in most cases.

The expected waiting times at Q_i for the different values of ρ were obtained according to the following steps: (a) calculate the variables $\alpha_{(j,c),k}$ and $\alpha^{(2)}_{(j,c),k}$ according to the recursive relations (9)-(10), for $j = 1, \ldots, M$, $c = 0, 1, \ldots, C$ (where C is a sufficiently large integer), and for all k for which $T(k) = i$, (b) determine $E[X_k]$ $(k = 1, \ldots, M)$ according to (11), (c) determine $Var[X_k]$ according to (12)-(13), for all k for which $T(k) = i$, (d) determine $E[W_k^{(PQ)}]$ according to (2)-(3), for all k for which $T(k) = i$, (e) determine $E[V_k]$ $(k = 1, \ldots, M)$ according to (4), (f) determine π_k $(k =$

Table 6.2. Exact and approximated expected waiting times for different values of the load.

ρ	$E[W_1]$	$E[W_1^{(app)}]$	$err_1\%$	$E[W_4]$	$E[W_4^{(app)}]$	$err_4\%$
0.70	15.2	16.9	11.2	27.5	36.2	31.5
0.80	24.0	25.4	6.0	44.7	54.3	21.5
0.90	49.7	50.8	2.2	97.9	108.5	10.8
0.95	100.7	101.6	1.0	205.9	217.0	5.4
0.98	253.1	253.9	0.3	531.2	542.6	2.2
0.99	507.1	507.8	0.1	1073.6	1085.2	1.1

$1, \ldots, M$), according to (21), where v_k is replaced by $E[V_k]$, (g) determine $E[W_i]$ according to the right-hand side of (22), where v_k and $\omega_k^{(PQ)}$ are replaced by $E[V_k]$ and $E[W_k^{(PQ)}]$, respectively.

We emphasize that the computation time required to obtain the results for the higher values of ρ in Tables 6.1 and 6.2 is on the order of minutes, whereas the computation times required to obtain the approximations are negligible.

7 CONCLUDING REMARKS AND TOPICS FOR FURTHER RESEARCH

This paper extends the analysis of cyclic polling models in [23] to the case of general periodic polling models. Although the derivation of the results proceeds along similar lines as those in [23], the added value of this paper is significant. First, the obtained expressions for the expected delay are new and specific for periodic polling. Second, we obtain insensitivity results which provide new insights into the behavior of general periodic polling systems under heavy load. Third, the results lead to closed-form expressions for the expected delay for the star and elevator polling schemes, and potentially other polling schemes with a specific structure. Fourth, the results suggest new, fast-to-evaluate and accurate approximations of the expected delay in general (not necessarily cyclic) periodic polling systems.

Finally, we address a number of topics for further research.

The ultimate goal of performance modeling is to obtain the "best" system performance, while the proper operation of the system is particularly critical in heavy-load scenarios. The results presented in this paper provide new insights into how the expected delay figures depend on the routing scheme and other system parameters, and also suggest sharp approximations for the expected delay. This opens possibilities for optimization of the system performance, e.g., with respect to the order in which the queues are visited (cf. [5]). Optimization of the system performance is a challenging topic for further research.

In many cases it is important to have knowledge about more than the expected delay only. To this end, it is useful to extend the results presented here to the higher moments of the delay. For the special case of cyclic polling and zero switch-over times, explicit expressions for the delay moments are presented in [26]. Extension to models general non-cyclic periodic polling models is a topic for further research.

Acknowledgments

I thank G.L. Choudhury for providing the software used to perform the numerical experiments. The results in this paper extend joint work with H. Levy. A partial and preliminary version of this paper appeared in [24].

References

[1] Alford, M. and Muntz, R.R. (1975). Queueing models for polled multi-queues. In: *Fourth Texas Conf. Comput. Syst.*, 5B-2.1-2.5.

[2] Baker, J.E. and Rubin, I.R. (1987). Polling with a general-service order table. *IEEE Trans. Commun.* **35**, 283-288.

[3] Blanc, J.P.C. (1993). Performance analysis and optimization with the power-series algorithm. In: *Performance Evaluation of Computer and Communication Systems*, eds. L. Donatiello and R. Nelson (Springer-Verlag, Berlin), 53-80.

[4] Boxma, O.J., Groenendijk, W.P. and Weststrate, J.A. (1990). A pseu-

doconservation law for service systems with a polling table. *IEEE Trans. Commun.* **38**, 1865-1870.

[5] Boxma, O.J., Levy, H. and Weststrate, J.A. (1991). Efficient visit frequencies for polling tables: minimization of the waiting cost. *Queueing Systems* **9**, 133-162.

[6] Choudhury, G.L. (1990) Polling with a general service order table: gated service. In: *Proc. INFOCOM '90*, 268-276.

[7] Choudhury, G. and Whitt, W. (1994). Computing transient and steady state distributions in polling models by numerical transform inversion. *Perf. Eval.* **25**, 267-292.

[8] 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.

[9] 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.*

[10] Eisenberg, M. (1972). Queues with periodic service and changeover times. *Oper. Res.* **20**, 440-451.

[11] Eisenberg, M. (1994). The polling system with a stopping server. *Queueing Systems* **18**, 387-431.

[12] Fricker, C. and Jaïbi, M.R. (1994). Monotonicity and stability of periodic polling models. *Queueing Systems* **15**, 211-238.

[13] Konheim, A.G., Levy, H. and Srinivasan, M.M. (1994). Descendant set: an efficient approach for the analysis of polling systems. *IEEE Trans. Commun.* **42**, 1245-1253.

[14] Kroese, D.P. (1997). Heavy traffic analysis for continuous polling models. *J. Appl. Prob.* **34**, 720-732.

[15] Leung, K.K. (1991). Cyclic service systems with probabilistically-limited service. *IEEE J. Sel. Areas Commun.* **9**, 185-193.

[16] Levy, H. and Sidi, M. (1991). Polling models: applications, modeling and optimization. *IEEE Trans. Commun.* **38**, 1750–1760.

[17] Reiman, M.I. and Wein, L.M. (1994). Dynamic scheduling of a two-class queue with setups. To appear in *Oper. Res.*

[18] Resing, J.A.C. (1993). Polling systems and multitype branching processes. *Queueing Systems* **13**, 409-426.

[19] Srinivasan, M.M., Niu, S.-C. and Cooper, R.B. (1995). Relating polling models with zero and nonzero switchover times. *Queueing Systems* **19**, 149-168.

[20] Takagi, H. (1986). *Analysis of Polling Systems* (The MIT Press, Cambridge, MA).

[21] Takagi, H. (1991). Applications of polling models to computer networks. *Comp. Netw. ISDN Syst.* **22**, 193-211.

[22] Van der Mei, R.D. and Borst, S.C. (1997). Analysis of multiple-server polling systems by means of the power-series algorithm. *Stoch. Mod.* **13**, 339-369.

[23] Van der Mei, R.D. and Levy, H. (1996). Expected delay analysis of polling systems in heavy traffic. To appear in *Adv. Appl. Prob.*

[24] Van der Mei, R.D. (1997). Periodic polling systems in heavy traffic. In: *Ten Years LNMB*, eds. W.K. Klein Haneveld, O.J. Vrieze and L.C.M. Kallenberg (CWI Tract 122), 179-189.

[25] Van der Mei, R.D. and Levy, H. (1997). Polling systems in heavy traffic: exhaustiveness of the service disciplines. *Queueing Systems* **27**, 227-250.

[26] 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. Ramaswami and P.E. Wirth (Elsevier, Amsterdam), 275-284.

APPENDIX

Proof of Part 2 of Theorem 1

It is convenient to write equations (10) in matrix notation. To this end, let $\underline{\alpha}^{(2)}_{(\cdot,c),k}$ be the vector whose i-th component is $\alpha^{(2)}_{(i,c),k}$. Define for $c = 1, 2, \ldots,$ and $k = 1, \ldots, M$,

$$\underline{y}_{0,k} = \mathbf{P}_1 \cdots \mathbf{P}_{k-1}\underline{e}_k + \sum_{j=1}^{k-1} \frac{b^{(2)}_{T(j)}}{b^2_{T(j)}(1 - \rho_{T(j)} I_{\{T(j)\in E\}})} \alpha^2_{(j,0),k} \mathbf{P}_1 \cdots \mathbf{P}_{j-1}\mathbf{e}_j, \quad (34)$$

$$\underline{y}_{c,k} = \sum_{j=1}^{M} \frac{b^{(2)}_{T(j)}}{b^2_{T(j)}(1 - \rho_{T(j)} I_{\{T(j)\in E\}})} \alpha^2_{(j,c),k} \mathbf{P}_1 \cdots \mathbf{P}_{j-1}\underline{e}_j. \quad (35)$$

Using these definitions, equation (10) can be expressed in matrix notation as follows: for $k = 1, \ldots, M$, $c = 1, \ldots,$

$$\underline{\alpha}^{(2)}_{(\cdot,0),k} = \underline{y}_{0,k}, \quad \underline{\alpha}^{(2)}_{(\cdot,c),k} = \mathbf{M}\underline{\alpha}^{(2)}_{(\cdot,c-1),k} + \underline{y}_{c,k} = \sum_{j=0}^{c} \mathbf{M}^j \underline{y}_{c-j,k}. \quad (36)$$

Denoting $\hat{\underline{y}}_k = \sum_{c=0}^{\infty} \underline{y}_{c,k}$, equation (36) leads to the following expression: For $k = 1, \ldots, M$,

$$\sum_{c=0}^{\infty} \underline{\alpha}^{(2)}_{(\cdot,c),k} = \sum_{j=0}^{\infty} \mathbf{M}^j \sum_{c=0}^{\infty} \underline{y}_{c,k} = \sum_{j=0}^{\infty} \mathbf{M}^j \hat{\underline{y}}_k = \frac{1}{1-\gamma} \underline{u}\, \underline{w}^\top \hat{\underline{y}}_k + \underline{\epsilon}, \quad (37)$$

where $\underline{\epsilon}$ consists of lower-order terms in the sense that they become negligible when $\rho \uparrow 1$. The first equality follows from equation (36), the second equality is trivial, and the last equality follows from Lemma 4.1.

We first prove that in the limiting case $\rho \uparrow 1$ the series $\sum_{c=0}^{\infty} \alpha^{(2)}_{(i,c),k}$ is proportional to $b_{T(i)}$. This observation follows from the following sequence of equalities: for $i, j, k = 1, \ldots, M$,

$$\lim_{\rho\uparrow 1} \frac{\sum_{c=0}^{\infty} \alpha^{(2)}_{(i,c),k}}{\sum_{c=0}^{\infty} \alpha^{(2)}_{(j,c),k}} = \lim_{\rho\uparrow 1} \frac{u_i \underline{w}^\top \hat{\underline{y}}_k}{u_j \underline{w}^\top \hat{\underline{y}}_k} = \lim_{\rho\uparrow 1} \frac{u_i}{u_j} = \frac{b_{T(i)}}{b_{T(j)}}. \quad (38)$$

The first equality follows from (37), the second equality is trivial and the third equality follows from Lemma 4.2 and the continuity of the eigenvalues and eigenvectors with respect to ρ.

It remains to prove that in the limiting case $\rho \uparrow 1$ the series $\sum_{c=0}^{\infty} \alpha^{(2)}_{(i,c),k}$ is proportional to x_k^2. Denoting $\hat{\underline{y}}_k = (\hat{y}_{1k}, \ldots, \hat{y}_{Mk})$, the result follows from

the following equalities: for $i, k, l = 1, \ldots, M$,

$$\lim_{\rho \uparrow 1} \frac{\sum_{c=0}^{\infty} \alpha^{(2)}_{(i,c),k}}{\sum_{c=0}^{\infty} \alpha^{(2)}_{(i,c),l}} = \lim_{\rho \uparrow 1} \frac{\hat{y}_{ik}}{\hat{y}_{il}} = \lim_{\rho \uparrow 1} \frac{\sum_{c=0}^{\infty} \alpha^{2}_{(i,c),k}}{\sum_{c=0}^{\infty} \alpha^{2}_{(i,c),l}} = \lim_{\rho \uparrow 1} \frac{\hat{u}^2_k}{\hat{u}^2_l} = \frac{x^2_k}{x^2_l}. \tag{39}$$

The first equality follows from (37). The second equality follows from the fact that $\underline{y}_{c,k}$, and hence also $\hat{\underline{y}}_k$, depends on k only through $\alpha_{(i,c),k}$ (see (34)-(35)). The third equality follows from Lemmas 4.1 and the last equality follows from Lemma 4.2 and the continuity of the eigenvectors and eigenvalues with respect to ρ.

Received: 6/9/1997
Revised: 5/27/1998
Accepted: 8/15/1998