Centrum voor Wiskunde en Informatica
Centre for Mathematics and Computer Science

O.J. Boxma, J.A. Weststrate

Waiting times in polling systems with Markovian server routing

Department of Operations Research, Statistics, and System Theory Report BS-R8905 March

The Centre for Mathematics and Computer Science is a research institute of the Stichting Mathematisch Centrum, which was founded on February 11, 1946, as a nonprofit institution aiming at the promotion of mathematics, computer science, and their applications. It is sponsored by the Dutch Government through the Netherlands Organization for the Advancement of Research (N.W.O.).

Copyright © Stichting Mathematisch Centrum, Amsterdam

Waiting Times in Polling Systems with Markovian Server Routing

O.J. Boxma
Centre for Mathematics and Computer Science
P.O. Box 4079, 1009 AB Amsterdam, The Netherlands
Faculty of Economics, Tilburg University
P.O. Box 90153, 5000 LE Tilburg, The Netherlands

J.A. Weststrate
Faculty of Economics, Tilburg University
P.O. Box 90153, 5000 LE Tilburg, The Netherlands

This study is devoted to a queueing analysis of polling systems with a probabilistic server routing mechanism. A single server serves a number of queues, switching between the queues according to a discrete time parameter Markov chain. The switchover times between queues are nonnegligible. It is observed that the total amount of work in this Markovian polling system can be decomposed into two independent parts, viz., (i) the total amount of work in the corresponding system without switchover times and (ii) the amount of work in the system at some epoch covered by a switching interval. This work decomposition leads to a pseudoconservation law for mean waiting times, i.e., an exact expression for a weighted sum of the mean waiting times at all queues. The results generalize known results for polling systems with strictly cyclic service.

1980 Mathematics Subject Classification: 60K25, 68M20.
Key Words & Phrases: Markovian polling, work decomposition, pseudoconservation law, mean waiting times

1. Introduction

A system in which one server visits a set of queues, in some order, is commonly referred to as a polling system. A large number of queueing theoretic studies about polling systems has been published. The vast majority of these studies considers polling systems in which the server serves the queues in a strictly cyclic order. Several service strategies at the queues have been investigated and implemented in actual computer-communication networks; these strategies range from exhaustive (a queue is served until it is empty) to 1-limited (when the queue is non-empty, the server serves exactly one customer).

The main performance measures of cyclic polling systems are the mean waiting times at the various queues. When all queues have an exhaustive service strategy, the exact mean waiting times at the queues can be determined by solving a system of linear equations. For most other service strategies exact mean waiting times are only known in special cases; see the surveys of Takagi [1986,1988] for detailed results and further references.

Recently some generalizations have been considered, which encompass a much larger class of cyclic polling systems. One generalization of purely cyclic polling is a polling system with a service order table, i.e., a list of stations which the server must successively visit. Stations can be given higher priority by listing them more often in the table. See Boxma et al. [1988]. The polling table [1,2,...,N] gives the purely cyclic case, and [1,2,1,3,...,1,N] represents the important star polling scheme.

Another generalization concerns polling systems with a random polling scheme. In a random polling scheme, the polling order is not fixed but determined by some random mechanism. In a recent study, Kleinrock and Levy [1988] analyzed the behaviour of a random polling system in which the next station polled will be the jth station with probability p_j, independent of the present station. A large p_j corresponds to a high priority for the jth station. Kleinrock and Levy [1988] consider three

different systems. In each one, all stations have the same service strategy: exhaustive, gated or 1-limited. For the exhaustive and gated strategies, they give the individual mean response times (waiting times plus service times) as the solution of a system of linear equations. For the 1-limited strategy they determine the mean response time for the special case of a completely symmetric system. They state that their results can be used to predict the expected delay in an exhaustive slotted ALOHA system; the random polling mechanism represents the random scheme according to which it is decided which station will transmit during the next slot.

In this study we also consider a polling system with probabilistic server routing. In our case the next station polled will be determined by a discrete time parameter Markov chain. We shall sometimes speak of *Markovian polling*. This includes cyclic polling and the purely random polling scheme of Kleinrock and Levy as special cases. The service strategies at the various queues may be different (exhaustive at one queue, 1-limited at the next one, etc.).

The switchover times of the server between queues are assumed to be nonnegligible. Hence there is no *work conservation*. However, the work conservation principle can be extended to a *work decomposition* principle (Boxma and Groenendijk [1987], Boxma [1989]). This work decomposition principle states that the amount of work in a polling system with switchover times can be decomposed into two independent parts, viz., (i) the amount of work in the same polling system but without switchover times and (ii) the amount of work in the system at some epoch covered by a switching interval.

Boxma and Groenendijk [1987] have used the work decomposition principle for cyclic polling systems to derive a *pseudoconservation law* for such systems, viz., an exact expression for a weighted sum of the mean waiting times at the queues. These results yield new insight into the behaviour of polling systems and can be used to obtain approximations for the individual mean waiting times. In the present study a pseudoconservation law will be derived for Markovian polling.

This study has been motivated by various considerations. Firstly, Markovian polling provides a theoretically interesting generalization of cyclic polling, and therefore we have considered it worthwhile to try and generalize the work decomposition principle and pseudoconservation law of Boxma and Groenendijk [1987] to the case of Markovian polling. Secondly, Markovian polling appears to have some interesting practical applications. While cyclic polling has been successfully used to model systems where a central controller polls many stations, Markovian polling may be used to model *distributed systems* (Kleinrock and Levy [1988]). As an example, Levy [1984] uses a random polling model to predict the mean delay in a slotted Aloha system. We believe that a second example may be found in the Orwell ring protocol (Mitrani, Adams and Falconer [1986]). In this protocol, c slots of equal length rotate around a ring. Each slot can accommodate one packet. A packet in a slot filled by a station Q_i is addressed to station Q_j with a certain probability. Q_j empties the slot and passes it on empty to the next downstream station. This is a major departure from other slotted ring protocols, where a slot can be released only by the station which filled it. For the case of $c = 1$ slot, it seems an interesting possibility to use a Markovian polling model to approximate the performance of the Orwell ring protocol. Such a model captures the stochastic character of the order of service of the stations, although it ignores the fact that the server transition probabilities in reality depend on whether or not a packet is waiting for transmission (whether or not a queue is non-empty).

A final reason for studying Markovian polling (and polling tables, for that matter) is that they open possibilities for optimization by considering various choices of the transition probabilities (and of the polling table). As a first step towards obtaining insight into this matter, we have compared the performance of polling systems with either cyclic or star polling and the performance of Markovian polling systems with the same server visit frequencies.

The paper is organized as follows. In section 2 a model description is presented. Section 3 contains some preliminaries concerning the mean visit times of the server at the various queues, and a brief discussion of ergodicity conditions. Section 4 gives the work decomposition for Markovian polling systems. Section 5 is devoted to the main result of this study, the derivation of the pseudoconservation law. The determination of the mean interdeparture times of the server between queues will be central in that analysis. It will be shown that a system of N^2 equations in the N^2 unknown mean

interdeparture times of the server, can be related to a system of N^2 equations in the N^2 unknown mean entrance times of the underlying reversed Markov chain (in fact, the system can be decomposed into N sets of N linear equations). For some special cases, explicit expressions for the mean interdeparture times are derived.

2. Model description

In this paper we consider a (continuous time) queueing system with N stations (queues), $Q_1,...,Q_N$, where each station has an infinite buffer capacity to store waiting messages (customers).

Message arrival process.

Customers arrive at all queues according to independent Poisson processes. The arrival intensity at Q_i is λ_i, $i=1,...,N$. The total arrival rate is given by:

$$\Lambda := \sum_{i=1}^{N} \lambda_i.$$

Customers who arrive at Q_i are called type-i customers.

Service process.

The service times of type-i customers are independent, identically distributed stochastic variables. Their distribution $B_i(.)$ has first moment β_i and second moment $\beta_i^{(2)}$. The offered traffic, ρ_i, at Q_i is defined as:

$$\rho_i := \lambda_i \beta_i, \quad i=1,...,N,$$

and the total offered traffic, ρ, as:

$$\rho := \sum_{i=1}^{N} \rho_i.$$

Polling strategy.

The N stations are served by a single server S who visits the stations according to a Markovian polling scheme. The next station to be polled is determined according to an irreducible positive recurrent discrete time parameter Markov chain $M = \{\mathbf{d}_n, n=0,1,...\}$ with state space $I = \{1,...,N\}$. With $\{\mathbf{d}_n = i\}$ we denote the event that the nth station polled after $t=0$ is station Q_i, $i \in I$.

We assume that the Markov chain M has stationary one-step transition probabilities, i.e., the conditional probabilities $Pr\{\mathbf{d}_{n+1}=j|\mathbf{d}_n=i\}$, $i,j \in I$, are independent of n.

Define:

$$p_{ij} := Pr\{\mathbf{d}_{n+1}=j|\mathbf{d}_n=i\}, \quad i,j \in I, \quad n=0,1,... , \tag{2.1}$$

$$q_i := \lim_{n\to\infty} Pr\{\mathbf{d}_n=i\}, \quad i \in I, \quad n=0,1,... . \tag{2.2}$$

For the waiting time analysis in Section 5 it will turn out to be essential to consider the *time reversed process* of the Markov chain M, $\tilde{M} = \{\tilde{\mathbf{d}}_n, n=0,1,...\}$, obtained from M by reversing the time parameter. The following theorem is proved in Kelly [1979], pp. 28,29:

Theorem 2.1

If M is a stationary discrete time parameter Markov chain with state space I, one-step transition probabilities p_{ij}, $i,j \in I$, and with equilibrium distribution $\{q_j, j \in I\}$, then the reversed process $\tilde{M}$ is a stationary discrete time parameter Markov chain with state space I, one-step transition probabilities

$$\tilde{p}_{ij} := Pr\{\tilde{\mathbb{d}}_{n+1}=j|\tilde{\mathbb{d}}_n=i\} = \frac{q_j}{q_i}p_{ji}, \quad i,j\in I, \tag{2.3}$$

and with the same equilibrium distribution $\{q_j,\ j\in I\}$.

In the sequel $\tilde{M}$ will be called the reversed Markov chain.

Service strategy.
For the service strategies at the stations there are various possibilities, which differ in the number of customers who may be served in a queue during a visit of S to that queue. Assume that S visits Q_i. If Q_i is not empty S acts as follows, depending on the service strategy at Q_i:

I Exhaustive service (E): S serves type-i customers until Q_i is empty;
II Gated service (G): S serves exactly those type-i customers present upon his arrival at Q_i (a gate closes upon his arrival);
III 1-Limited service (1-L): S serves exactly one type-i customer.

In the sequel we will allow mixed service strategies (e.g., exhaustive at Q_1, 1-limited at Q_2 and Q_4, gated at Q_3, etc.).
After the visit period at Q_i (which has length zero when Q_i is empty) S switches with probability p_{ij} to Q_j, $i,j\in I$.

REMARK 2.1
We have restricted ourselves to the above-mentioned three main disciplines in polling systems. We could have included other strategies; in fact we show in Section 5 how 1-limited service in Markovian polling leads to the Bernoulli service discipline (cf. Keilson and Servi [1986]) in cyclic polling.

Switching process.
A switchover time is needed to switch from Q_i to Q_j, $i,j\in I$. The switchover times of the server between Q_i and Q_j are independent, identically distributed stochastic variables with mean s_{ij} and second moment $s_{ij}^{(2)}$.

The message arrival processes, the service demand processes and the switching processes are assumed to be mutually independent.

3. PRELIMINARY RESULTS

First some definitions. We define the visit time of the server at Q_i, V_i, as:

$$V_i := \text{the time between the arrival of the server at } Q_i \text{ and its subsequent departure from } Q_i,\ i\in I. \tag{3.1}$$

NOTE
If $p_{ii}>0$ and if, moreover, the switchover time from Q_i to Q_i is zero with positive probability, then a visit to Q_i may immediately be followed by another such visit.

We also define

$$\mathbb{u}_n := \text{the time between the departure of the server from the (n-1)th station polled and its departure from the nth station polled after } t=0\ ;$$

$\mathbf{u}_n$ is the sum of the switchover time from the $(n-1)$th station polled after $t=0$ to the nth station, and the visit time to the latter station. The process $SM = \{(\mathbf{d}_n, \mathbf{u}_n),\ n=0,1,...\}$, with $M = \{\mathbf{d}_n,\ n=0,1,...\}$ the discrete time parameter Markov chain described in Section 2, is a semi-Markov process (cf. Cinlar [1975], Chapter 10, Section 5). Note that when the $(n-1)$th station polled after $t=0$ is Q_i and the nth station polled after $t=0$ is Q_j:

$$E\mathbf{u}_n = s_{ij} + E\mathbf{V}_j.$$

We also have to introduce the server's interdeparture time between Q_i and Q_j, $\mathbf{T}_{ij}$, $i,j \in I$:

$\mathbf{T}_{ij} :=$ the time between a departure of S from Q_j and its last previous departure from Q_i. (3.2)

So during the time span $\mathbf{T}_{ij}$, for $j \neq i$, S has not returned to Q_i but there may have been several visits to Q_j.
From a work balancing argument it follows that

$$E\mathbf{V}_i = \rho_i E\mathbf{T}_{ii},\quad i \in I. \tag{3.3}$$

Another balancing argument yields:

$$\frac{q_i E\mathbf{V}_i}{q_j E\mathbf{V}_j} = \frac{\rho_i}{\rho_j},\quad i,j \in I, \tag{3.4}$$

(cf. Cinlar [1975], p. 341), i.e. the ratio of the average amount of time S spends at Q_i and the average amount of time S spends at Q_j equals the ratio of the average traffic loads at Q_i and Q_j.
Combining (3.3) and (3.4) gives:

$$q_i E\mathbf{T}_{ii} = q_j E\mathbf{T}_{jj},\ i,j \in I, \tag{3.5}$$

hence $q_i E\mathbf{T}_{ii}$, $i \in I$, is a constant.

Ergodicity conditions
A necessary condition for ergodicity of the system is $\rho<1$. When the service strategy at each queue is either exhaustive or gated this condition can be shown to be also sufficient. Without proof we observe that a necessary condition for ergodicity for a 1-limited station Q_i is:

$$\lambda_i E\mathbf{T}_{ii} < 1. \tag{3.6}$$

Indeed, $\lambda_i E\mathbf{T}_{ii}$ equals the mean number of arrivals to Q_i between two successive visits (and potential services) of the server at Q_i. For the mixed service strategies that we allow, condition (3.6) should be added to the stability condition $\rho<1$ for those queues at which we have a 1-limited service strategy. In the sequel it will be assumed that the ergodicity conditions are fulfilled, and that the system is in equilibrium.

4. Work decomposition

In this section we state that the amount of work in the Markovian polling system with switchover times can be decomposed into two independent terms, one of which is the amount of work in the same system but *without* switchover times. Before we give the work decomposition theorem we introduce the notion of 'corresponding M/G/1-system'. The corresponding M/G/1-system indicates a single-server system with exactly the same arrival processes, service demand processes and scheduling

disciplines (i.e. a procedure for deciding which customer, if any, should be in service at any time) as the Markovian polling system under consideration, but without switchover times. The principle of work conservation implies that the amount of work in the latter system is independent of the service discipline: Markovian polling and FCFS service for all customers, irrespective of the queue they join, give rise to identical amounts of work at all times.
Define:

$\mathbf{V}_{MP}$:= steady-state amount of work in the Markovian polling system,
$\mathbf{V}$:= steady-state amount of work in the corresponding M/G/1-system,
$\mathbf{Y}$:= steady-state amount of work in the Markovian polling system at some epoch covered by a switching interval.

We relate these quantities in the next theorem.

THEOREM 4.1
Consider a single-server multi-queue Markovian polling system as described in Section 2. Suppose the system is ergodic and stationary. Then the steady-state amount of work in this system, $\mathbf{V}_{MP}$, is distributed as the sum of the steady-state amount of work in the corresponding M/G/1-system, $\mathbf{V}$, and the steady-state amount of work at some epoch covered by a switching interval, $\mathbf{Y}$:

$$\mathbf{V}_{MP} \overset{D}{=} \mathbf{V} + \mathbf{Y}, \tag{4.1}$$

where $\overset{D}{=}$ stands for equality in distribution. Furthermore, $\mathbf{V}$ and $\mathbf{Y}$ are independent.

PROOF
See Boxma [1989].

5. THE PSEUDOCONSERVATION LAW

In this section we use Theorem 4.1 to derive an expression for a weighted sum of the mean waiting times. As a consequence of Theorem 4.1:

$$E\mathbf{V}_{MP} = E\mathbf{V} + E\mathbf{Y}, \tag{5.1}$$

and hence from M/G/1 theory, cf. Cohen [1982]:

$$E\mathbf{V}_{MP} = \frac{\sum_{i=1}^{N} \lambda_i \beta_i^{(2)}}{2(1-\rho)} + E\mathbf{Y}. \tag{5.2}$$

On the other hand, when $\mathbf{X}_i$ denotes the number of waiting type-i customers and $\mathbf{W}_i$ the waiting time of a type-i customer in the Markovian polling system with switchover times:

$$E\mathbf{V}_{MP} = \sum_{i=1}^{N} \beta_i E\mathbf{X}_i + \sum_{i=1}^{N} \rho_i \frac{\beta_i^{(2)}}{2\beta_i} = \sum_{i=1}^{N} \rho_i E\mathbf{W}_i + \frac{1}{2}\sum_{i=1}^{N} \lambda_i \beta_i^{(2)}, \tag{5.3}$$

the first equality following from the fact that service is non-preemptive, and the second equality following from Little's formula. Combination of (5.2) and (5.3) yields:

$$\sum_{i=1}^{N} \rho_i E\mathbf{W}_i = \rho \frac{\sum_{i=1}^{N} \lambda_i \beta_i^{(2)}}{2(1-\rho)} + E\mathbf{Y}. \tag{5.4}$$

To obtain an expression for this weighted sum of mean waiting times it remains to determine $E\mathbf{Y}$, the mean amount of work at some epoch covered by a switching interval. Denote by $\mathbf{Y}_{ij}$ the amount of work in the Markovian polling system at some epoch covered by a switchover from Q_i to Q_j, $i,j \in I$. $E\mathbf{Y}$ can be expressed as a weighted sum of all $E\mathbf{Y}_{ij}$, averaging over the switchover durations and the frequencies with which transitions between queues occur:

$$E\mathbf{Y} = (1/\sigma)\sum_{i=1}^{N} q_i \sum_{j=1}^{N} p_{ij}s_{ij}E\mathbf{Y}_{ij}, \tag{5.5}$$

with

$$\sigma := \sum_{i=1}^{N} q_i \sum_{j=1}^{N} p_{ij}s_{ij}, \tag{5.6}$$

the average mean switchover time. Note that in the purely cyclic case ($p_{i,i+1}=1$, $i \in I$): $\sigma=(1/N)\sum_{i=1}^{N} s_{i,i+1}$, with $\sum_{i=1}^{N} s_{i,i+1}$ the mean total switchover time in one cycle.
It remains to determine $E\mathbf{Y}_{ij}$, $i,j \in I$. $E\mathbf{Y}_{ij}$ is composed of three terms, only one of which depends on j:

$E\mathbf{M}_i^{(1)}$:= the mean amount of work in Q_i at a departure epoch of S from Q_i,

$E\mathbf{M}_i^{(2)}$:= the mean amount of work in $Q_1,\ldots,Q_{i-1},Q_{i+1},\ldots,Q_N$, at a departure epoch of S from Q_i,

$\rho\frac{s_{ij}^{(2)}}{2s_{ij}}$:= the mean amount of work that arrived in the system during the past part of the switching interval (from Q_i to Q_j) under consideration.

So we can write:

$$E\mathbf{Y}_{ij} = E\mathbf{M}_i^{(1)} + E\mathbf{M}_i^{(2)} + \rho\frac{s_{ij}^{(2)}}{2s_{ij}}. \tag{5.7}$$

It will turn out that $E\mathbf{M}_i^{(1)}$ is the only term in the righthand side of (5.7) that depends on the service strategy at Q_i. It can only be specified when the service strategy at Q_i is known.

We shall first consider $E\mathbf{M}_i^{(2)}$, the mean amount of work in $Q_1,\ldots,Q_{i-1},Q_{i+1},\ldots,Q_N$ at a departure epoch of the server from Q_i. Q_k ($k\neq i$) can make two contributions to $E\mathbf{M}_i^{(2)}$:

- the mean amount of work left behind in Q_k by S at his last departure from Q_k,
- the mean amount of work that has arrived in Q_k during $\mathbf{T}_{ki}$, the server's interdeparture time between Q_k and Q_i (cf. (3.2)).

We obtain the following relation:

$$E\mathbf{M}_i^{(2)} = \sum_{k\neq i} E\mathbf{M}_k^{(1)} + \sum_{k\neq i} \rho_k E\mathbf{T}_{ki}, \quad i \in I. \tag{5.8}$$

Substitution of (5.8) in (5.7) gives:

$$E\mathbf{Y}_{ij} = \sum_{k=1}^{N} E\mathbf{M}_k^{(1)} + \sum_{k\neq i} \rho_k E\mathbf{T}_{ki} + \rho\frac{s_{ij}^{(2)}}{2s_{ij}}, \quad i,j \in I. \tag{5.9}$$

$E\mathbf{M}_k^{(1)}$ and $E\mathbf{T}_{ki}$ still have to be determined. The $E\mathbf{M}_k^{(1)}$ are derived below for an exhaustive, gated or

1-limited service strategy at Q_k:

(i) Q_k has an exhaustive service strategy: Q_k is left behind empty by S, so

$$E\mathbf{M}_k^{(1)} = 0. \tag{5.10}$$

(ii) Q_k has a gated service strategy: $E\mathbf{M}_k^{(1)}$ equals ρ_k times the mean visit time of S at Q_k, hence (cf. (3.3)),

$$E\mathbf{M}_k^{(1)} = \rho_k E\mathbf{V}_k = \rho_k^2 E\mathbf{T}_{kk}. \tag{5.11}$$

(iii) Q_k has a 1-limited service strategy: a similar derivation as in Boxma and Groenendijk [1987] leads to,

$$E\mathbf{M}_k^{(1)} = \lambda_k E\mathbf{T}_{kk}[\rho_k(E\mathbf{W}_k+\beta_k)] + (1-\lambda_k E\mathbf{T}_{kk})0 = \rho_k\lambda_k E\mathbf{T}_{kk}E\mathbf{W}_k + \rho_k^2 E\mathbf{T}_{kk}, \tag{5.12}$$

($\lambda_k E\mathbf{T}_{kk}$ is the fraction of visits of S to Q_k that result in a service, and $\rho_k(E\mathbf{W}_k+\beta_k)$ equals the mean amount of work that has arrived during the sojourn time of a departing customer).
Substituting (5.10),...,(5.12) in (5.9) gives:

$$E\mathbf{Y}_{ij} = \sum_{k\in g,1-l} \rho_k^2 E\mathbf{T}_{kk} + \sum_{k\in 1-l} \rho_k\lambda_k E\mathbf{T}_{kk}E\mathbf{W}_k + \sum_{k\neq i} \rho_k E\mathbf{T}_{ki} + \rho\frac{s_{ij}^{(2)}}{2s_{ij}}, \quad i,j\in I, \tag{5.13}$$

with g and $1-l$ denoting the group of queues with gated and 1-limited service strategies respectively.

It now remains to determine $E\mathbf{T}_{ki}$, $k,i\in I$. We first introduce the event
B_{ji} := 'the last visit before a visit of S to Q_i was to Q_j'.
For all $k,i\in I$:

$$E\mathbf{T}_{ki} = \sum_{j=1}^{N} E\{\mathbf{T}_{ki}|B_{ji}\}Pr\{B_{ji}\}. \tag{5.14}$$

Determination of $E\mathbf{T}_{ki}$ requires looking backwards in time (cf. Theorem 2.1). We can write for all $i,j\in I$:

$$Pr\{B_{ji}\} = Pr\{\mathbf{d}_{n-1}=j|\mathbf{d}_n=i\} = Pr\{\tilde{\mathbf{d}}_{n+1}=j|\tilde{\mathbf{d}}_n=i\} = \tilde{p}_{ij}, \tag{5.15}$$

the one-step transition probabilities of the reversed Markov chain $\tilde{M}$. It easily follows that:

$$\begin{aligned} E\{\mathbf{T}_{ki}|B_{ji}\} &= E\{\mathbf{T}_{kj}\} + s_{ji} + E\mathbf{V}_i \quad \text{if } j\neq k, \\ E\{\mathbf{T}_{ki}|B_{ji}\} &= s_{ki} + E\mathbf{V}_i \quad \text{if } j=k. \end{aligned} \tag{5.16}$$

Substitution of (5.15) and (5.16) into (5.14) gives, for $k,i\in I$:

$$E\mathbf{T}_{ki} = \sum_{j\neq k} [E\mathbf{T}_{kj} + s_{ji} + E\mathbf{V}_i]\tilde{p}_{ij} + [s_{ki} + E\mathbf{V}_i]\tilde{p}_{ik}. \tag{5.17}$$

If we define

$$f(i) := E\mathbf{V}_i + \sum_{j=1}^{N} s_{ji}\tilde{p}_{ij}, \quad i\in I, \tag{5.18}$$

then we can rewrite (5.17) as:

$$E\mathbf{T}_{ki} = f(i) + \sum_{j \neq k} E\mathbf{T}_{kj}\, \tilde{p}_{ij}, \quad i,k \in I. \tag{5.19}$$

Clearly, the set of N^2 linear equations (5.19) can be decomposed into N sets of N linear equations. In the next lemma it will be shown that the N^2 unknown mean interdeparture times $E\mathbf{T}_{ki}$, $k,i \in I$ can be expressed in the N^2 mean entrance times between queues in the underlying reversed Markov chain $\tilde{M}$. The mean entrance time between Q_i and Q_j, $\tilde{v}_{ij}$, in the reversed Markov chain $\tilde{M}$ is defined as:

$$\tilde{v}_{ij} := E\{\#\text{ steps required for the first entrance into } Q_i \text{ starting from } Q_j\}, \quad i,j \in I, \tag{5.20}$$

(cf. Cohen [1982], p. 33).
Note that from the theory of Markov chains and from Theorem 2.1 we have,

$$\tilde{v}_{ii} = \frac{1}{q_i}, \quad i \in I. \tag{5.21}$$

We now formulate:

LEMMA 5.1
For all $i,k \in I$:

$$E\mathbf{T}_{ki} = f(i) + \sum_{l \neq k} f(l) \frac{\tilde{v}_{ik} + \tilde{v}_{kl} - \tilde{v}_{il}}{\tilde{v}_{ll}}. \tag{5.22}$$

PROOF
Denote by $\tilde{S}$ the server for the reversed Markov chain $\tilde{M}$. We start with two observations:
(i) $E\mathbf{T}_{ki}$ is in the reversed Markov chain $\tilde{M}$ the average time between an arrival of $\tilde{S}$ at Q_i and his first subsequent arrival at Q_k.
(ii) $f(i)$ is in the reversed Markov chain $\tilde{M}$ the average time between an arrival of $\tilde{S}$ at Q_i and his arrival at the next station to be visited after Q_i (possibly again Q_i).

Using these observations we can write for $i,k \in I$:

$$E\mathbf{T}_{ki} = f(i) + \sum_{l \neq k} f(l)\, E\{\#\text{ times } \tilde{S} \text{ visits } Q_l \text{ before it visits } Q_k \text{ starting from } Q_i\}. \tag{5.23}$$

Further on we can write for $i,k,l \in I$ and $k \neq l$ (see Chung [1967], p.46):

$$E\{\#\text{ times } \tilde{S} \text{ visits } Q_l \text{ before it visits } Q_k \text{ starting from } Q_i\} = \sum_{n=1}^{\infty} {}_k\tilde{p}_{il}^{(n)},$$

with

$${}_k\tilde{p}_{il}^{(n)} = Pr\{\tilde{\mathbf{d}}_n = l,\ \tilde{\mathbf{d}}_m \neq k,\ m = 1,\ldots,n-1 | \tilde{\mathbf{d}}_0 = i\},$$

i.e., the probability of going from Q_i to Q_l in n steps without visiting Q_k. Using Corollary 2 on page 65 of Chung [1967] we find:

$$\sum_{n=1}^{\infty} {}_k\tilde{p}_{il}^{(n)} = [\tilde{v}_{ik} + \tilde{v}_{kl} - \tilde{v}_{il}]/\tilde{v}_{ll}.$$

So we obtain:

$$E\{\#\text{ times } \tilde{S} \text{ visits } Q_l \text{ before it visits } Q_k \text{ starting from } Q_i\} = \tag{5.24}$$
$$[\tilde{v}_{ik}+\tilde{v}_{kl}-\tilde{v}_{il}]/\tilde{v}_{ll},\quad i,k,l\in I,\ k\neq l.$$

Combining (5.23) and (5.24) gives relation (5.22).

REMARK 5.1
An alternative way to prove Equation (5.22) is by means of matrix manipulations. The following steps are required:
(1) Denote by $\bar{T} = (E\mathbf{T}_{11},\ldots,E\mathbf{T}_{1N},E\mathbf{T}_{21},\ldots,E\mathbf{T}_{NN})^t$ the N^2-dimensional column vector of the unknown mean interdeparture times. Then we can write (5.19) in the following form:

$$\bar{T} = A\bar{T}+\bar{b},$$

with obvious definitions of the N^2 by N^2 matrix A and the N^2-dimensional vector $\bar{b}$.
(2) Because the eigenvalues of A are all less than one (Seneta [1981]), $\bar{T}$ can be written as:

$$\bar{T} = [I-A]^{-1}\bar{b} = [\sum_{n=0}^{\infty} A^{(n)}]\bar{b} = [I + \sum_{n=1}^{\infty} A^{(n)}]\bar{b}.$$

(3) It appears that $\sum_{n=1}^{\infty} A^{(n)}$ is a blockdiagonal matrix with N blocks of size $N\times N$. Denoting the (i,l)th element of the kth block by $C_k(i,l)$, it can be shown that for $i,l\in\{1,\ldots,N\}$, $k\in\{1,\ldots,N\}$,

$$C_k(i,l) = 0,\quad l=k, \qquad C_k(i,l) = \sum_{n=1}^{\infty} {}_k\tilde{p}_{il}^{(n)},\quad l\neq k.$$

Combination of steps 2 and 3 now yields (5.22).

So, to determine $E\mathbf{T}_{ki}$, $k,i\in I$ we can determine the mean entrance times, $\tilde{v}_{ik}$, $i,k\in I$, of the underlying reversed Markov chain $\tilde{M}$. It is known that these are the solution of

$$\tilde{v}_{ik} = 1 + \sum_{j\neq k} \tilde{p}_{ij}\tilde{v}_{jk},\quad i,k\in I; \tag{5.25}$$

cf. Cohen [1982].

From a theoretic point of view it is interesting to make the link between interdeparture times of S, and entrance times of the underlying reversed Markov chain (the more so because the semi-Markov process and its underlying (reversed) Markov chain arise so naturally in the present queueing model). From a numerical point of view it constitutes no real advantage to solve the set of equations (5.25) instead of the set of equations (5.19).

For $k=i$ Lemma 5.1 yields, with $\tilde{v}_{ll} = 1/q_l$, $l\in I$,

$$q_i E\mathbf{T}_{ii} = \sum_{l=1}^{N} q_l f(l),\ i\in I. \tag{5.26}$$

This demonstrates the fact that $q_i E\mathbf{T}_{ii}$ does not depend on i (cf. (3.5)). We now determine $C := q_i E\mathbf{T}_{ii}$, successively using (5.26), (5.18), (2.3), (5.6) and (3.3):

$$C = \sum_{l=1}^{N} q_l f(l) = \sum_{l=1}^{N} q_l \sum_{m=1}^{N} \tilde{p}_{lm}s_{ml} + \sum_{l=1}^{N} q_l E\mathbf{V}_l =$$
$$\sum_{m=1}^{N} q_m \sum_{l=1}^{N} p_{ml}s_{ml} + \sum_{l=1}^{N} q_l E\mathbf{V}_l = \sigma + \sum_{l=1}^{N} \rho_l q_l E\mathbf{T}_{ll} = \sigma + \rho C.$$

Hence

$$C = \frac{\sigma}{1-\rho}, \tag{5.27}$$

so

$$E\mathbf{T}_{ii} = \frac{1}{q_i}\frac{\sigma}{1-\rho}, \tag{5.28}$$

$$E\mathbf{V}_i = \frac{1}{q_i}\frac{\rho_i\sigma}{1-\rho}. \tag{5.29}$$

Substituting (5.28) in (5.13) gives:

$$E\mathbf{Y}_{ij} = \frac{\sigma}{1-\rho_{k\in g,\,1-l}}\sum\frac{\rho_k^2}{q_k} + \frac{\sigma}{1-\rho_{k\in 1-l}}\sum\frac{\rho_k}{q_k}\lambda_k E\mathbf{W}_k + \sum_{k\neq i}\rho_k E\mathbf{T}_{ki} + \rho\frac{s_{ij}^{(2)}}{2s_{ij}},\quad i,j\in I, \tag{5.30}$$

with $E\mathbf{T}_{ki}$ as in Lemma 5.1.

Combining (5.4), (5.5) and (5.30) gives our main result which is formulated in Theorem 5.1 below. As before, denote by g and $1-l$ the group of queues with gated and 1-limited service strategies, and further denote by e the group of queues with exhaustive service strategies.

THEOREM 5.1

Consider an ergodic and stationary single-server multi-queue Markovian polling system with mixed service strategies as described in Section 2. Then:

$$\sum_{k\in e,g}\rho_k E\mathbf{W}_k + \sum_{k\in 1-l}\rho_k[1-\frac{\lambda_k}{q_k}\frac{\sigma}{1-\rho}]E\mathbf{W}_k = \rho\frac{\sum_{i=1}^{N}\lambda_i\beta_i^{(2)}}{2(1-\rho)} + \frac{\sigma}{1-\rho_{k\in g,\,1-l}}\sum\frac{\rho_k^2}{q_k} + \frac{\rho}{2\sigma}\sum_{i=1}^{N}q_i\cdot\sum_{j=1}^{N}p_{ij}s_{ij}^{(2)} + \frac{1}{\sigma}\sum_{i=1}^{N}q_i\sum_{j=1}^{N}p_{ij}s_{ij}\sum_{k\neq i}\rho_k E\mathbf{T}_{ki}, \tag{5.31}$$

with $E\mathbf{T}_{ki}$ as in Lemma 5.1.

REMARK 5.2

In the purely cyclic case, $p_{i,i+1}=1$, $i\in I$, (5.31) reduces to (3.22) in Boxma and Groenendijk [1987]. Note that $(\sigma/q_k) = n\sigma$ in (5.31) corresponds to the total mean switchover time, s, in (3.22) of that publication, cf. below (5.6).

REMARK 5.3

Theorem 5.1 can be generalized to the case of a batch arrival process with correlated sizes of the batches simultaneously arriving at the various queues (cf. the cyclic polling model of Levy and Sidi [1988]).

REMARK 5.4

Kleinrock and Levy [1988] restrict themselves to the special case that $p_{ij} = p_j$ (random polling) and $s_{ij} = s_i$, $s_{ij}^{(2)} = s_i^{(2)}$ for all $i,j\in I$. In this case $q_k = p_k$, $k\in I$, and (5.17) reduces to:

$$ET_{ki} = \frac{\sigma}{1-\rho}\left[\frac{\rho_i}{q_i} - \frac{\rho_k}{q_k} + \frac{1}{q_k}\right], \quad k,i \in I. \tag{5.32}$$

To interpret this formula, note that

$$ET_{ki} - EV_i + EV_k = \frac{1}{q_k}\frac{\sigma}{1-\rho} = ET_{kk}, \quad k,i \in I; \tag{5.33}$$

and observe that in this case M is reversible, so that $ET_{ki} - EV_i + EV_k$ also equals the mean time between a departure from Q_i (or Q_k, as $p_{ij}=p_j$ for all i) and the first subsequent departure from Q_k. Formula (5.31) reduces to:

$$\sum_{k\in e,g} \rho_k EW_k + \sum_{k\in 1-l} \rho_k\left[1-\frac{\lambda_k}{p_k}\frac{\sigma}{1-\rho}\right]EW_k = \tag{5.34}$$

$$\rho\frac{\sum_{i=1}^{N}\lambda_i\beta_i^{(2)}}{2(1-\rho)} + \frac{\sigma}{1-\rho}\sum_{k\in g,\,1-l}\frac{\rho_k^2}{p_k} - \frac{\sigma}{1-\rho}\sum_{k=1}^{N}\frac{\rho_k^2}{p_k} + \frac{\sigma}{1-\rho}\sum_{k=1}^{N}\frac{\rho_k}{p_k} -$$

$$\sum_{i=1}^{N}\rho_i s_i + \frac{\rho}{2\sigma}\sum_{i=1}^{N} p_i\, s_i^{(2)}.$$

Kleinrock and Levy, for the cases of exhaustive service and of gated service at all queues, give the individual mean waiting times as the solution of a set of $O(N^2)$ linear equations. For the completely symmetric case, they determine the mean waiting times (which now are all the same) explicitly for the exhaustive, gated and 1-limited strategies. It can easily be shown that (5.34) leads to the expression found by Kleinrock and Levy for this completely symmetric case.

REMARK 5.5
At this stage we'd like to point at the relative simplicity of the pseudoconservation law formulated in Theorem 5.1. In the general case the righthand side of (5.31) can be evaluated after N sets of N linear equations have been solved; in special cases such as purely random polling, the ET_{ki} can be determined explicitly in a straightforward manner. This should be contrasted with the fact that, apart from two-queue models and completely symmetric models, the *individual* mean waiting times are only known for purely random polling with exhaustive or gated service at all queues. For the more general Markovian polling, the individual mean waiting times might again be determined for these two service disciplines, following the approach of Kleinrock and Levy [1988]; but this seems to require the solution of a set of $O(N^3)$ linear equations.

REMARK 5.6
It is interesting to compare cyclic polling and random polling with equal visit probabilities ($p_{ij}\equiv 1/N$) for all queues. We restrict ourselves to the case that, for both models, all queues have exactly the same traffic characteristics, while all switchover times are independent, identically distributed s.v. with mean r. With an obvious notation, the difference between the mean workloads in both models is (cf. (5.1)):

$$EV_{MP} - EV_{cycl} = EY_{MP} - EY_{cycl}. \tag{5.35}$$

Comparison of the pseudoconservation laws for cyclic polling (Boxma and Groenendijk [1987]) and random polling (formula (5.34)) yields:

$$EV_{MP} - EV_{cycl} = \frac{N-1}{2}\frac{r\rho}{1-\rho}. \tag{5.36}$$

Not surprisingly, the random character of the server visits in random polling leads to a higher mean workload than for cyclic polling.

When the service strategy at all queues is the same, then (5.36) leads to the following results of Kleinrock and Levy [1988]: for exhaustive and gated service, with an obvious notation,

$$EW_{MP} - EW_{cycl} = \frac{N-1}{2}\frac{r}{1-\rho}; \tag{5.37}$$

for 1-limited service,

$$EW_{MP} - EW_{cycl} = \frac{N-1}{2}\frac{r}{1-\rho-N\lambda_1 r}. \tag{5.38}$$

REMARK 5.7

We have also compared V_{MP} with V_{star}, the amount of work in a single-server N-queue system ($N \geqslant 2$) with star polling, i.e. server visits according to the polling table [1,2,1,3,...,1,N]. We have assumed that

- Q_1 receives exhaustive service, whereas $Q_2, \ldots, Q_N$ receive 1-limited service (in both models);
- both models have the same traffic characteristics;
- all switchover times are equal to the constant r;
- $p_{ij} = p_j$, with $p_1 = \frac{1}{2}$, $p_2 = \cdots = p_N = 1/(2(N-1))$.

Comparison of EV_{MP} and EV_{star} amounts to comparison of EY in both models. Using the evaluations of EY in the present paper and in Boxma et al. [1988], it can be shown that

$$EV_{MP} \geqslant EV_{star}; \tag{5.39}$$

in fact, if $\rho_2 = \cdots = \rho_N$, then

$$EV_{MP} - EV_{star} = \tfrac{1}{2}r\rho + r\frac{\rho-\rho_1}{1-\rho}[N-2+2\rho_1]. \tag{5.40}$$

It should be noted that this difference is

- roughly linearly increasing in N;
- tending to zero for $r \to 0$;
- dependent on λ_n and β_n only via their product ρ_n;
- equal to $\frac{1}{2}r\rho$ for $\rho = \rho_1$ ($Q_2, \ldots, Q_N$ receive no traffic), a result which is easily explained by noting that the comparison in this case amounts to a comparison of two M/G/1 queues with vacations.

Another special case of the Markovian polling model introduced in this paper is the cyclic service model with a Bernoulli schedule, as introduced by Keilson and Servi [1986]. This schedule operates as follows. If there are still customers present in Q_i after a service completion in this queue, the server decides with probability $1-p_i$ to serve the next customer at Q_i, and with probability p_i he switches to Q_{i+1}. He also takes the latter action when there are no more customers present in Q_i. Tedijanto [1988] has derived a pseudoconservation law for the cyclic service system with a Bernoulli schedule. Below we show how this pseudoconservation law follows as a special case of the pseudoconservation law in Theorem 5.1. Assume that all stations have a 1-limited service strategy, and take, for all $i \in I$,

$$p_{ij} = 1-p_i \quad \text{if } j = i,$$

$$p_{ij} = p_i \qquad \text{if } j = i+1,$$
$$p_{ij} = 0 \qquad \text{else ;}$$
$$s_{ii} = 0,$$
$$s_{i,i+1} = s_i,$$
$$s^{(2)}_{i,i+1} = s^{(2)}_i.$$

It should be noted that the server pays a geometrically distributed number of consecutive visits, with mean number $1/p_i$, to Q_i; even when Q_i has become empty, the server may still return a number of times, but this does not take time because $s_{ii}=0$. It follows that the server spends on the average $E\mathbf{V}_m/p_m$ at Q_m before he switches to Q_{m+1}. A work balancing argument now implies that

$$\frac{E\mathbf{V}_m}{p_m} = \rho_m \frac{D}{1-\rho}, \quad m \in I, \tag{5.41}$$

with

$$D := \sum_{i=1}^{N} s_i.$$

For this special case we also have:

$$f(i) = p_i[s_{i-1} + \rho_i \frac{D}{1-\rho}], \quad i \in I, \tag{5.42}$$

$$q_i = \frac{1}{p_i} \Big/ \Big(\sum_{m=1}^{N} \frac{1}{p_m}\Big), \quad i \in I, \tag{5.43}$$

$$\sigma = D \Big/ \Big(\sum_{m=1}^{N} \frac{1}{p_m}\Big), \tag{5.44}$$

and (cf. (5.25))

$$\tilde{v}_{ij} = \sum_{k=j}^{i-1} \frac{1}{p_k}, \quad i,j \in I, \tag{5.45}$$

(note that this is a cyclic sum; for $j > i-1$, $\tilde{v}_{ij}$ is the sum of $1/p_j, \ldots, 1/p_N, 1/p_1, \ldots, 1/p_{i-1}$). Using the above formulas, (5.22) reduces to:

$$E\mathbf{T}_{ki} = \sum_{m=k+1}^{i} \frac{E\mathbf{V}_m}{p_m} + \sum_{m=k}^{i-1} s_m, \quad k,i \in I. \tag{5.46}$$

Indeed, in the cyclic service model with a Bernoulli schedule, $E\mathbf{T}_{ki}$ equals the mean amount of time the server spends at $Q_{k+1},\ldots,Q_i$, plus the sum of the mean switchover times between Q_k and Q_i.

Using (5.41),...,(5.46), we obtain the pseudoconservation law for the cyclic service model with a Bernoulli schedule at all queues:

$$\sum_{k=1}^{N} \rho_k[1-\lambda_k p_k \frac{D}{1-\rho}]E\mathbf{W}_k = \tag{5.47}$$
$$\rho \frac{\sum_{i=1}^{N} \lambda_i \beta_i^{(2)}}{2(1-\rho)} + \frac{D}{1-\rho} \sum_{k=1}^{N} \rho_k^2 p_k + \rho \frac{D^{(2)}}{2D} + \frac{D}{2(1-\rho)}[\rho^2 - \sum_{i=1}^{N} \rho_i^2],$$

with $D^{(2)}$ the second moment of the sum of the N switchover times between Q_1 and Q_2, ..., Q_N and Q_1.

Expression (5.47) is the same as (3.6.6) in Tedijanto [1988]. Note that (5.47) reduces to the pseudoconservation law for cyclic polling with exhaustive (respectively 1-limited) service at all queues if all $p_k = 0$ (respectively all $p_k = 1$).

ACKNOWLEDGMENT

The authors are indebted to J.W. Cohen, F.A. van der Duyn Schouten, W.P. Groenendijk and H. Levy for useful comments and stimulating discussions.

REFERENCES

1. BOXMA, O.J. (1989). *Workloads and waiting times in single-server systems with multiple customer classes.* To appear in *Queueing Systems.*
2. BOXMA, O. J., GROENENDIJK, W.P. (1987). *Pseudo-conservation laws in cyclic-service systems.* J. Appl. Prob. **24**, 949-964.
3. BOXMA, O.J., GROENENDIJK, W.P., WESTSTRATE, J.A. (1988). *A pseudoconservation law for service systems with a polling table.* Report Centre for Mathematics and Computer Science, Amsterdam.
4. CINLAR, E. (1975). *Introduction to Stochastic Processes* (Prentice Hall, Englewood Cliffs, NJ).
5. COHEN, J.W. (1982). *The Single Server Queue* (North-Holland, Amsterdam; 2nd ed.).
6. KEILSON, J., SERVI, L.D. (1986). *Oscillating random walk models for GI/G/1 vacation systems with Bernoulli schedules.* J. Appl. Prob. **23**, 790-802.
7. KELLY, F.P. (1979). *Reversibility and Stochastic Networks* (Wiley, New York).
8. KLEINROCK, L., LEVY, H. (1988). *The analysis of random polling systems.* Oper. Res. **36**, 716-732.
9. LEVY, H. (1984). *Non-Uniform Structures and Synchronization Patterns in Shared-Channel Communication Networks.* CSD-840049, Computer Science Department, University of California, Los Angeles, Ph.D. Dissertation.
10. LEVY, H., SIDI, M. (1988). *Correlated arrivals in polling systems.* Report Department of Computer Science, Tel Aviv University.
11. MITRANI, I., ADAMS, J.L., FALCONER, R.M. (1986). *A modelling study of the Orwell ring protocol.* In: Teletraffic Analysis and Computer Performance Evaluation, eds. O.J. Boxma, J.W. Cohen and H.C. Tijms (North-Holland, Amsterdam), pp. 429-438.
12. E. SENETA (1981). *Non-negative Matrices and Markov Chains* (Springer, New York; 2nd ed.).
13. TAKAGI, H. (1986). *Analysis of Polling Systems* (The MIT Press, Cambridge, MA).
14. TAKAGI, H. (1988). *Queuing analysis of polling models.* ACM Comput. Surveys **20**, 5-28.
15. TEDIJANTO (1988). *Exact results for the cyclic-service queue with a Bernoulli schedule.* Report Electrical Engineering Department and Systems Research Center, University of Maryland.