Centrum voor Wiskunde en Informatica
Centre for Mathematics and Computer Science

O.J. Boxma, H. Levy, J.A. Weststrate

Efficient visit orders for polling systems

Department of Operations Research, Statistics, and System Theory Report BS-R9017 June

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.).

Efficient Visit Orders for Polling Systems

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

H. Levy
Department of Computer Science
The Raymond and Beverly Sackler Faculty of Exact Sciences
Tel-Aviv University, Tel-Aviv 69978, Israel

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

This paper deals with the issue of deriving efficient operational rules for polling systems with switchover periods. Specifically, we study the following static optimization problem: Determine the server visit order (polling table) that minimizes the mean total workload. This problem is strongly related with, and in many applications coincides with, that of minimizing the overall mean customer delay in the system. Two heuristic approaches to the polling table problem are presented, both using the exact solution of a related problem. The first approach is based on exact optimality results for a random polling system with the same traffic characteristics. The second approach uses a polling table model, in the analysis of which some of the restrictions are relaxed. Numerical experiments show that both approaches yield excellent results.

1980 Mathematics Subject Classification: 60K25, 68M20.
Key Words & Phrases: random polling, polling table, mean workload minimization.

1. Introduction

The basic polling system is a system of multiple queues, attended to by a single server in a cyclic order. Polling systems arise naturally in the modelling of many computer, communication and production networks where several users compete for access to a common resource (a central computer, a transmission channel, a carousel in an assembly line). Takagi [26,27] and Levy and Sidi [22] mention a large variety of applications.

Such applications also give rise to several variants of the basic polling system, like:
(i) probabilistic polling: the server visits the queues according to a probabilistic routing mechanism. Probabilistic polling may be used to model distributed control systems, in which the decision which station will be served next is achieved in a distributed manner, by cooperation among the stations. Cf. Kleinrock and Levy [17] who specifically mention the example of an exhaustive slotted Aloha system.
(ii) periodic polling: the server visits the queues in a fixed order specified by a polling table in which each queue occurs at least once (cf. Eisenberg [10], Baker and Rubin [1]). Some examples are provided by the token bus protocol in Local Area Networks, and by star polling at a computer with multidrop terminals (polling table [1,2,1,3,...,1,N]).

Probabilistic polling and periodic polling open useful and interesting possibilities for efficient operation and optimization, by allowing various choices of the server routing probabilities respectively the polling table. Optimization in polling systems is a subject which has so far received very little

Report BS-R9017
Centre for Mathematics and Computer Science
P.O. Box 4079, 1009 AB Amsterdam, The Netherlands

attention in queueing literature. Of the more than 450 references in Takagi's recently updated polling survey [27], almost none is concerned with optimization issues. Most polling studies do not go beyond the comparison of performance measures under different service disciplines at the queues. One of the few exceptions is the paper of Browne and Yechiali [9]. Using Markov decision processes, they determine a semi-dynamic policy in which the server, at the beginning of a cycle, chooses a visiting order of the queues for this cycle that minimizes the mean duration of the cycle. A somewhat similar problem for the case of unit buffers is considered in [8]. Another exception is the paper by Hofri and Ross [13] in which they consider a polling system of two queues and study the properties of an optimal dynamic rule for determining the service duration of each queue as a function of the queue occupancy.

The present study is devoted to optimization of polling systems. The next section contains a global discussion of several polling optimization issues. The goal there is to draw attention to this problem area and to list a number of interesting research themes. The rest of the paper deals with the problem of efficiently operating a polling system with switchover times in order to minimize the mean total workload in the system. This problem is equivalent with that of minimizing $\sum \rho_i E\mathbf{W}_i$, a weighted sum of the mean waiting times $E\mathbf{W}_i$ where the weights are the traffic loads ρ_i of the queues. In the practically relevant case that all mean service times are equal, this amounts to minimizing the overall mean customer delay $E\mathbf{W} := \sum \lambda_i E\mathbf{W}_i / \sum \lambda_i$, λ_i denoting the arrival rate at the i-th queue. $E\mathbf{W}$ is perhaps the single most important performance measure in polling systems. The strong relation between total workload and overall mean customer delay adds to the importance of the mean workload as a performance measure and as an objective function for optimization.

As will be explained in Section 2, efficient operation of a polling system can be achieved by controlling two parameters: 1) The amount of service given to each queue during its service period, and 2) The order (and frequency) by which the server visits the different queues. For each given visit order, the service strategy which minimizes mean workload is to serve in each visit as many customers as possible (cf. [23] and Section 3 below). This rule translates to using the exhaustive (and possibly the gated) service strategy at each queue. Further performance improvement can be achieved by a proper selection of the visit order (and frequency). Our focus is, therefore, on the following optimization problem: Determine the polling table that minimizes the mean total workload in a periodic polling model with switchover times and either gated or exhaustive service. To solve this problem we present two heuristic approaches which use the exact solutions of two related problems. The first determines those server visit frequencies that lead to the minimal total workload in a related *probabilistic* polling system. The second finds the optimal visit frequencies in a system in which some of the restrictions are relaxed.

The structure of this paper is as follows: Section 2 discusses the general problem of the optimization of polling systems. Section 3 prepares the ground, by reviewing recent polling results which are necessary for tackling our optimization problem. Section 4 considers the optimization of probabilistic polling systems. Section 5 considers periodic polling systems for which it proposes approximate approaches for finding efficient polling tables. Section 6 presents numerical results, and Section 7 contains a summary and some plans for the future.

Main results

We find optimal visit frequencies for the random polling system. Closed-form expressions for these frequencies are obtained for systems in which all the switchover periods are identically distributed; for the case of non-identical switchover periods these frequencies are determined numerically by solving a simple non-linear optimization problem.

We next turn to the more interesting periodic polling problem, for which we find an efficient polling order as follows:

(1) Determine the numbers of visits to the queues by using either the optimal visit frequencies derived for the random polling system with the same traffic characteristics, or by using a heuristically derived formula based on the analysis of a system with relaxed assumptions.

(2) For given numbers of visits, determine the visit order using the so-called *Golden Ratio* policy.

Note
Preliminary and partial results have appeared in the conference paper [5]. The main extensions are: a discussion of the important case of unequal switchover times is added, a wider range of numerical examples is studied, and a new approach to the periodic polling problem is proposed (which opens possibilities for minimization of a weighted sum of mean waiting times with *arbitrary* non-negative weight factors).

2. Optimization of polling systems

The ultimate goal of performance modelling and analysis is performance improvement and system optimization. Performance analysis can be applied at any stage of development, from the initial design phase to the operational phase. The range of options from which one can choose, and the optimization problems to be tackled, are mainly determined by the stage of development. For example, in designing a local area network there may first exist such channel access options as collision-detection protocols or collision-free token passing protocols. And when a token passing mechanism has been elected, the network configuration may be open for discussion: Should it be a bus, or a ring, or perhaps several interconnected rings? In the latter case, how should stations be assigned to the rings? Which static or dynamic priority rules should be implemented to give certain stations more opportunities to transmit, or longer transmission periods (thus improving some performance measure)?

Similar performance and optimization problems occur naturally in many other settings that give rise to polling models; whether it be in the design of traffic light regulation systems for signalized intersections, or in the development of a robotics system for processing several streams of parts. In the abstract setting of a single server that serves several customer classes, we now briefly discuss optimization criteria and regulation mechanisms.

Optimization criteria
In optimizing a polling system there is generally a trade-off between *efficiency* and *fairness**. From the point of view of minimizing workload in the system it may be efficient to visit heavy traffic queues frequently and for lengthy periods of time; but this may be "unfair" to the low traffic queues. Performance criteria which are often being studied in polling systems are the mean total workload, the server cycle time, and the individual mean waiting times or a weighted sum of them. Those weight factors may be chosen such as to represent costs, or generally to represent an appropriate balance between efficiency and fairness.

Regulation mechanisms
A natural regulation mechanism in many queueing systems is the customer access mechanism. A few polling studies have allowed finite buffer sizes, but to the best of our knowledge the paper of Browne and Yechiali [8], which considers the routing of the server in a system with unit buffers, is the only one of those in which optimization plays a key role. Here lies an important field of study.

A considerable part of the polling literature is devoted to detailed studies of service policies at the queues. The obtained results allow some comparison between different policies. Unfortunately, most sophisticated deterministic policies do not yield to an exact mathematical analysis. Recently some probabilistic service policies have been introduced, which may be used to prioritize the queues and affect system performance, and which may be better amenable to mathematical analysis. The latter

* Although "fairness" is considered as an important performance criterion by many designers and researchers, we do not know of any good definition of this term. The topic of "fairness" is open for future research.

statement holds in particular for the fractional service policies suggested by Levy [20,21]. In such fractional policies, queue Q_i is assigned a parameter p_i, $0<p_i\leqslant 1$, and - loosely speaking - each of the customers present when the server visits Q_i (and possibly those arriving during their service times, etc.) has a probability p_i of receiving service in this visit period. The choice of the p_i gives rise to interesting optimization problems, which will be discussed in another paper. Another probabilistic policy is the Bernoulli service policy, in which a limit to the number of customers served in a service period is set using the Bernoulli distribution. This policy seems to affect the performance more than the fractional policies, but it is less amenable to mathematical analysis (cf. Servi [25]).

Another basic regulation mechanism in polling systems is the server routing between queues. Cyclic routing is more and more becoming a naive strategy, dating from the days in which not enough computing power was available to implement something more sophisticated. A non-cyclic fixed routing scheme (*polling table*) in which it should be possible to visit some stations more frequently than others may be more attractive. Nowadays many designers try to build a good *polling table,* but there are no clear-cut rules on how to form the table. The main goal of the present paper is to provide and test such rules.

It obviously makes sense to combine consideration of service policies and server routing strategies. For example, instead of including a queue several times in the polling table and serving one customer at each visit, it may be better to visit it only once or twice and provide exhaustive service. In the present study we do not touch upon this issue, but the results that we obtain can be used for a further investigation in this direction.

One may go a step beyond fixed, static, routing schemes. In dynamic routing the server visit order is changing dynamically, being determined by the system state during its operation. For example, it may be natural to observe the contents of the queues and to serve next the most heavily loaded queue. The advantage of dynamic server routing is that it is very sensitive to the actual system state and can thus be used to improve its performance. The disadvantages are that it requires information gathering during operation and that it is generally very hard to analyze. For systems without switchover times in which the queue to be served next is chosen after each service completion on the basis of complete information about the buffer contents, and with the goal of minimizing the weighted sum with general cost coefficients $\sum c_iE\mathbf{W}_i$, a simple $c\mu$ rule holds (see [22] for some references). However, for systems with positive switchover times, results are very scarce. Hofri and Ross [13] consider a two-queue model with switchover times and their results suggest that the optimal switchover rule is of a threshold type, i.e., there exist thresholds that determine when the server switches from one queue to the other. As mentioned in Section 1, Browne and Yechiali [8,9] study semi-dynamic server routing, in which the server visits all the queues exactly once in a cycle, but chooses a new cycle order at the end of each cycle.

Above we have indicated some global optimization issues in polling systems. By now a vast body of knowledge concerning polling systems is available. We believe it is time to develop optimization techniques to improve their performance, borrowing methods and insight from such fields as nonlinear programming, Markov decision theory and control theory.

3. Model description and a brief review of workloads and waiting times in polling systems

Model description

A single server, S, serves N infinite-capacity queues (stations) $Q_1,\ldots,Q_N$, switching from queue to queue. Customers arrive at all queues according to independent Poisson processes. The arrival intensity at Q_i is λ_i, $i=1,...,N$. Customers arriving at Q_i are called class-i customers. The service times of class-i customers are independent, identically distributed stochastic variables. Their distribution $B_i(.)$ has first moment β_i and second moment $\beta_i^{(2)}$, $i=1,...,N$. The offered traffic load, ρ_i, at Q_i is defined as $\rho_i := \lambda_i\beta_i$, $i=1,...,N$, and the total offered load, ρ, as $\rho := \sum_{i=1}^N \rho_i$. The switchover times of S between the various queues are independent stochastic variables. We specify them further when the need arises.

The scheduling discipline is the procedure for deciding which customer(s) should be in service at any time. In the polling models under consideration, the scheduling discipline can be decomposed into three components: (i) the server routing between queues; (ii) the switchover times between queues; (iii) the service policy at each queue. With regard to those service policies we restrict ourselves here mainly to exhaustive service (S empties each queue that he visits) and gated service (S serves exactly those customers in the queue who were present upon his arrival at that queue).
A case can be made for not including switchover times into the scheduling discipline. We have chosen to include them, because of the crucial influence that their presence has on the concepts to be discussed in the next paragraph.

Workloads - work conservation and work decomposition
One of the most fundamental properties that single-server multi-class service systems may possess is the property of *work conservation.* The scheduling discipline (server behaviour) is work conserving if (i) S serves at constant rate, (ii) he serves if and only if at least one customer is present, and (iii) his behaviour does not affect the amount of service given to a customer, or the arrival time of any customer. In this case a sample path consideration shows that the amount of work in the system is the same, whatever server behaviour with the above-mentioned properties occurs.
In a polling system with switchover times of the server between classes, the principle of work conservation is *violated* in the sense that the service process is interrupted although work is still present. Recently it has been shown [2] that, under rather weak conditions, a simple extension of the work conservation principle holds, viz. a *work decomposition principle:*

$$\hat{\mathbf{V}} \overset{D}{=} \mathbf{V} + \mathbf{Y}. \tag{3.1}$$

In this relation $\hat{\mathbf{V}}$ is the steady-state amount of work in the system with switchover times, $\mathbf{Y}$ is the steady-state amount of work in that system at an arbitrary switchover epoch, and $\mathbf{V}$ is the steady-state amount of work in the corresponding system *without* switchover times. Here $\overset{D}{=}$ denotes equality in distribution, and 'the corresponding system without switchover times' indicates a single-server multiclass system with exactly the same arrival and service demand process as the system under consideration, but without switchover times (hence work conserving).

The work decomposition formula (3.1) is in particular valid for cyclic polling [3] and for two special polling schemes that generalize cyclic polling and that play a central role in the remainder of this study: periodic polling [4] and probabilistic (Markovian) polling [7].

In (3.1) $\mathbf{V}$ is completely independent of the scheduling discipline, but $\mathbf{Y}$, and hence also $\hat{\mathbf{V}}$, *does* depend on it. Naturally, $\mathbf{Y}$ and $\hat{\mathbf{V}}$ should decrease
(i) with decreasing switchover times;
(ii) with increasingly 'efficient' visit order;
(iii) with increasing exhaustiveness of the service at the queues.
We return to the first two properties later on in this section. The third property has been formalized in [23], where a general framework is presented for a sample-path comparison of different service policies in polling systems. A sample path comparison is made which allows the evaluation of the policies based on the total amount of work $V(t)$ in the system at any time t. This comparison concerns policies operating with the same realizations of the arrival, service and switchover processes and of the polling order. These processes and the polling order are allowed to be quite general. The only restrictions on the server behaviour outside switchover periods are that it should be work conserving and that the server does not wait idling in an empty queue. The sample path comparison leads to the following results:
(i) The workload at any time t under the exhaustive service policy is less than or equal to the workload at t under any other arbitrary policy:

$$V_{exhaustive}(t) \leqslant V_{policy}(t).$$

(ii) With a similar notation,

$$V_{k-limited}(t) \leqslant V_{m-limited}(t) \text{ for } k \geqslant m;$$

(under the c-limited policy, S serves at most c customers before leaving the queue).

Similar comparisons are made for *stochastic* policies like the Bernoulli and binomial-gated policies: an ordering is proven w.r.t. the parameter of the related (Bernoulli, binomial) probability distribution.

Waiting times - conservation laws and pseudoconservation laws

Consider a single-server multi-class system for which (3.1) holds. First we restrict ourselves to the case of zero switchover times. Introduce $\mathbf{W}_n$, the waiting time (excluding service time) of a class-n customer. Under the assumption that the scheduling discipline is non-preemptive, and that only information about the current state and the past of the queueing process is used in making scheduling decisions, it can be shown that (cf. Kleinrock [15,16]):

$$\sum_{n=1}^{N} \rho_n E\mathbf{W}_n = \rho \frac{\sum_{n=1}^{N} \lambda_n \beta_n^{(2)}}{2(1-\rho)}. \tag{3.2}$$

Kleinrock called (3.2) a *conservation law* to indicate that a change in the scheduling discipline (under the above restrictions) does not lead to a change in $\sum \rho_n E\mathbf{W}_n$.

Under the same conditions as above, one obtains [2] in the case of non-zero switchover times, using (3.1):

$$\sum_{n=1}^{N} \rho_n E\mathbf{W}_n = \rho \frac{\sum_{n=1}^{N} \lambda_n \beta_n^{(2)}}{2(1-\rho)} + E\mathbf{Y}. \tag{3.3}$$

This has been coined a *pseudoconservation law:* a change in the visit order or service policy at a queue generally *does* lead to a change in $E\mathbf{Y}$, and hence in the lefthand side of (3.3). We shall specify $E\mathbf{Y}$ for the following server visit orders: (I) cyclic polling, (II) periodic polling, and (III) Markovian polling. In the sequel, the groups of queues that are being served under the exhaustive (gated) service policy are denoted by e (g).

I. Cyclic polling

Formula (3.3) now reduces to the following pseudoconservation law [3]:

$$\sum_{n=1}^{N} \rho_n E\mathbf{W}_n = \rho \frac{\sum_{n=1}^{N} \lambda_n \beta_n^{(2)}}{2(1-\rho)} + \rho\frac{s^{(2)}}{2s} + \frac{s}{2(1-\rho)}[\rho^2 - \sum_{n=1}^{N} \rho_n^2] + \frac{s}{1-\rho}\sum_{n \in g} \rho_n^2. \tag{3.4}$$

Here s and $s^{(2)}$ denote the mean and second moment of the sum of the switchover times in one cycle. It should be noted that $E\mathbf{Y}$ appears to be roughly linearly dependent on the mean total switchover time s; $E\mathbf{Y}$ and $E\hat{\mathbf{V}}$ appear to increase roughly linearly with increasing switchover times. It is also noteworthy that the order of the queues in the cycle does not influence the mean workload of the system or the weighted sum of mean waiting times in (3.4) as long as this order does not affect s.

II. Periodic polling (Polling Table)

First some additional notation. The order in which S visits the queues is specified in a polling table $T = \{T(m), m=1,...,M\}$. The i-th entry $T(i)$ is the index of the i-th queue polled in the cycle that is created by the polling table. This queue is referred to as the i-th 'pseudostation'. For example, $T = \{1,2,1,3\}$ denotes a cycle in which Q_1, Q_2, Q_1, Q_3 are consecutively visited. The first and third pseudostation both refer to Q_1. s_m and $s_m^{(2)}$ indicate the mean and second moment of the switchover time between the m-th and $(m+1)$-st pseudostations; s denotes the mean of the total switchover time in one cycle. The mean value of the time spent by S at pseudostation m is denoted by $E\mathbf{VI}_m$. Generally there is no simple expression available for these mean visit times, but they can be obtained by solving a simple set of linear equations, cf. [1,4].

Finally we introduce the $M \times M$ $(0,1)$ matrix $Z = (z_{ij})$, where $z_{ij} = 0$ unless none of the table entries $T(i+1), \ldots, T(j)$ equals $T(i)$, in which case $z_{ij} = 1$. The following pseudoconservation law has been proven in [4] ($\tilde{g}$ denotes the group of gated pseudostations):

$$\sum_{n=1}^{N} \rho_n E\mathbf{W}_n = \rho \frac{\sum_{i=1}^{N} \lambda_i \beta_i^{(2)}}{2(1-\rho)} + \rho \sum_{m=1}^{M} \frac{s_m^{(2)}}{2s} + \tag{3.5}$$

$$\sum_{k=1}^{M} \rho_{T(k)} \sum_{m \neq k} \frac{s_m}{s} z_{km} \sum_{j=k}^{m-1} (s_j + E\mathbf{VI}_{j+1}) + \sum_{j \in \tilde{g}} \rho_{T(j)} E\mathbf{VI}_j \sum_{m=1}^{M} \frac{s_m}{s} z_{jm} + \sum_{m \in \tilde{g}} \rho_{T(m)} \frac{s_m}{s} E\mathbf{VI}_m.$$

The sum $\sum_{j=k}^{m-1}$, with $1 \leqslant k, m \leqslant M$, should be interpreted as a cyclic sum.

III. Markovian polling

Again, we first need some notation. S is assumed to move between the N queues according to an irreducible, positive recurrent discrete-time parameter Markov chain $\{\mathbf{d}_n, n=0,1,...\}$ with stationary transition probabilities $p_{ij} = Pr\{\mathbf{d}_{n+1}=j \mid \mathbf{d}_n=i\}$, $i,j=1,...,N$, $n=0,1,...$. The limiting and stationary distribution of this Markov chain is denoted by $q_i = \lim_{n \to \infty} Pr\{\mathbf{d}_n = i\}$, $i=1,...,N$. The switchover times of S between Q_i and Q_j are i.i.d. stochastic variables with mean s_{ij} and second moment $s_{ij}^{(2)}$.

An important quantity in this model is $\mathbf{T}_{ki}$, the time between a departure of S from Q_i and the last previous departure from Q_k, $k,i=1,...,N$. Generally, determination of all $E\mathbf{T}_{ki}$ requires the solution of N sets of N linear equations. In [7] the following pseudoconservation law has been proven:

$$\sum_{n=1}^{N} \rho_n E\mathbf{W}_n = \rho \frac{\sum_{i=1}^{N} \lambda_i \beta_i^{(2)}}{2(1-\rho)} + \frac{\sigma}{1-\rho} \sum_{k \in g} \frac{\rho_k^2}{q_k} + \tag{3.6}$$

$$\frac{\rho}{2\sigma} \sum_{i=1}^{N} q_i \sum_{j=1}^{N} p_{ij} s_{ij}^{(2)} + \frac{1}{\sigma} \sum_{i=1}^{N} q_i \sum_{j=1}^{N} p_{ij} s_{ij} \sum_{k \neq i} \rho_k E\mathbf{T}_{ki},$$

with

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

Kleinrock & Levy [17] restrict themselves to the case that $p_{ij} = p_j$ (this is referred to as *random polling)* and $s_{ij} = s_i$, $s_{ij}^{(2)} = s_i^{(2)}$ for all $i,j \in \{1,...,N\}$. In this case $q_k = p_k$, $k = 1,...,N$, $E\mathbf{T}_{ki} = (\sigma/(1-\rho))[(\rho_i/q_i) - (\rho_k/q_k) + (1/q_k)]$, $k,i=1,...,N$, and (3.6) reduces to:

$$\sum_{n=1}^{N} \rho_n E\mathrm{W}_n = \rho \frac{\sum_{i=1}^{N} \lambda_i \beta_i^{(2)}}{2(1-\rho)} - \frac{\sum_{i=1}^{N} p_i s_i}{1-\rho} \sum_{k \in e} \frac{\rho_k^2}{p_k} + \frac{\sum_{i=1}^{N} p_i s_i}{1-\rho} \sum_{k=1}^{N} \frac{\rho_k}{p_k} - \sum_{i=1}^{N} \rho_i s_i + \rho \frac{\sum_{i=1}^{N} p_i \, s_i^{(2)}}{2\sum_{i=1}^{N} p_i s_i}. \tag{3.7}$$

In Sections 4 and 5 we shall use some of the above conservation laws to attempt minimization of the mean workload in a polling system with either Markovian polling or periodic polling, when the server transition probabilities, respectively the table, can be freely chosen.

Remark 3.1
The principle of work conservation implies that, in the case of zero switchover times, *any* server visit order leads to the same mean workload. But for positive switchover times and given service policies at the queues, the mean total workload does depend on the visit frequency and the visit pattern. In the next two sections we shall investigate this in detail for random polling and periodic polling.

Remark 3.2
From the equations determining the mean visit times $E\mathrm{VI}_m$ in (3.5), respectively the mean interdeparture times $E\mathrm{T}_{ki}$ in (3.6) (cf. [4] respectively [7]), it can be seen that these quantities depend on λ_j and β_j only through their product ρ_j. Hence only the first term in the righthand sides of (3.5) and (3.6) depends on individual arrival rates and service time moments; and this first term does not depend on the choice of the polling table respectively the server transition probabilities. The implication is that for the optimal choice of the table or the transition probabilities, only traffic loads matter and not individual arrival rates and service time moments.

Remark 3.3
Note that higher moments of the service time distributions appear in the pseudoconservation laws (3.5) and (3.6) only in the first term in the righthand side. The optimal choice of the visit probabilities and visit pattern, respectively, is not affected by these higher moments. Similarly, this choice is insensitive to second and higher moments of the switchover time distributions, as long as these distributions are all the same.

4. Optimization of random polling systems

Consider the random polling system described at the end of Section 3. In the present section we are interested in the following problem. Suppose that for given arrival, service and switchover processes and service disciplines at the queues, the system designer still has the freedom to choose the server transition probabilities p_j, $j=1,...,N$. He wants to choose them such that the mean steady-state amount of work in the system is minimized. This can be done using a simple numerical procedure for solving a constrained minimization problem which amounts to minimizing the righthand side of (3.7).

Intuitively one expects that queues with heavy traffic should be visited more frequently than low traffic queues. But how much more frequently? In purely cyclic polling it is well known that the ratios of mean visit times of the queues are equal to the ratios of the offered traffic loads. Should visit frequencies in random polling obey the same rule in order to minimize mean workload? Does the choice of service discipline matter?

To get qualitative insight into these questions we first consider a special case of the general optimization problem, in which all switchover times are identically distributed; at least, all switchover times have the same first moment $\hat{\sigma}$ and second moment $\hat{\sigma}^{(2)}$. The resulting expressions for optimal visit frequencies appear to be extremely appealing, being simple, robust and elegant.

It follows from (3.7) and the fact that, cf. [7],

$$E\hat{V} = \sum_{n=1}^{N} \rho_n E\mathbf{W}_n + \sum_{n=1}^{N} \frac{\beta_n^{(2)}}{2\beta_n}\rho_n, \tag{4.1}$$

that minimization of $E\hat{V}$ w.r.t. $p_1, \dots, p_N$, under the conditions $p_1 + \cdots + p_N = 1$, $p_1 \geqslant 0, \cdots, p_N \geqslant 0$, amounts to the following problem.

$$Min\ [\frac{\sum_{i=1}^{N} \lambda_i \beta_i^{(2)}}{2(1-\rho)} - \frac{\hat{\sigma}}{1-\rho}\sum_{k\in e}\frac{\rho_k^2}{p_k} + \frac{\hat{\sigma}}{1-\rho}\sum_{k=1}^{N}\frac{\rho_k}{p_k} - \rho\hat{\sigma} + \frac{\rho}{2\hat{\sigma}}\hat{\sigma}^{(2)}] \tag{4.2}$$

s.t.

$$p_1 + \cdots + p_N = 1, \quad p_1 \geqslant 0, \ \cdots, p_N \geqslant 0.$$

This is a classical non-linear optimization problem with linear constraints. Introducing the Lagrange multiplier L, and omitting all terms in (4.2) that do not involve the probabilities p_i, we want to minimize the unconstrained Lagrangian function

$$F := -\frac{\hat{\sigma}}{1-\rho}\sum_{k\in e}\frac{\rho_k^2}{p_k} + \frac{\hat{\sigma}}{1-\rho}\sum_{k=1}^{N}\frac{\rho_k}{p_k} + L(\sum_{k=1}^{N} p_k - 1) \tag{4.3}$$

in the non-negative 2^N-tant. The Kuhn-Tucker points of this expression are obtained by putting $\frac{\partial F}{\partial p_k} = 0$, $k = 1,\dots,N$ and $\frac{\partial F}{\partial L} = 0$, yielding

(i) if Q_k has exhaustive service:

$$-\frac{\hat{\sigma}}{1-\rho}\frac{\rho_k - \rho_k^2}{p_k^2} + L = 0;$$

(ii) if Q_k has gated service:

$$-\frac{\hat{\sigma}}{1-\rho}\frac{\rho_k}{p_k^2} + L = 0.$$

The convexity of F in $(p_1,\dots,p_N)$ readily implies that the admissible stationary point yields the minimum of $E\hat{V}$:
If Q_k has exhaustive service:

$$p_k^* = \frac{\sqrt{\rho_k(1-\rho_k)}}{\sum_{j\in e}\sqrt{\rho_j(1-\rho_j)} + \sum_{j\in g}\sqrt{\rho_j}}; \tag{4.4}$$

and if Q_k has gated service:

$$p_k^* = \frac{\sqrt{\rho_k}}{\sum_{j\in e}\sqrt{\rho_j(1-\rho_j)} + \sum_{j\in g}\sqrt{\rho_j}}. \tag{4.5}$$

As announced, these optimal server routing probabilities are remarkably simple. Allocation is according to a square root rule, and only the offered traffic loads play a role. In the exhaustive case the influence of this load is quite small. It should be noted that the visit frequency for a queue Q_k with exhaustive service is *decreasing* in ρ_k for $\rho_k > 0.5$, the other loads being kept constant. Still, it is easily seen that among any two queues with exhaustive service the one with higher load has a higher visit frequency.

In light traffic, for exhaustive service as well as for gated service, all routing probabilities are allocated according to the square root of the traffic loads; in fact, this can easily be shown to be true for *any* service policy.

Remark 4.1

When S meets n customers in Q_k, his mean visit period equals $n\beta_k$ in the case of gated service, and $n\beta_k/(1-\rho_k)$ in the case of exhaustive service. This gives some feeling as to why a queue with exhaustive service should receive fewer visits than a queue with gated service and the same traffic load, and why relatively few visits should be made to a queue with exhaustive service in heavy traffic.

Remark 4.2

We have also performed the above minimization for the cases of binomial-gated and binomial-exhaustive service. In binomial-gated service, when S finds m customers present at Q_k he serves n out of those m with probability $\binom{m}{n}\alpha_k^n(1-\alpha_k)^{m-n}$ $(0<\alpha_k\leqslant 1)$ and then leaves the queue; in the same situation under binomial-exhaustive service, S selects n out of those m customers with probability $\binom{m}{n}\gamma_k^n(1-\gamma_k)^{m-n}$ $(0<\gamma_k\leqslant 1)$ and serves those customers, and the ones arriving during their service, etc. Denoting the binomial-exhaustive (binomial-gated) queues by *be (bg)*, and now including e in *be* $(\gamma_k=1)$ and g in *bg* $(\alpha_k=1)$, we find if Q_k has binomial-exhaustive service:

$$p_k^* = \frac{\sqrt{\rho_k(1-\rho_k)/\gamma_k}}{\sum\limits_{j\in be}\sqrt{\rho_j(1-\rho_j)/\gamma_j}+\sum\limits_{j\in bg}\sqrt{\rho_j/\alpha_j}}, \qquad (4.6)$$

and if Q_k has binomial-gated service:

$$p_k^* = \frac{\sqrt{\rho_k/\alpha_k}}{\sum\limits_{j\in be}\sqrt{\rho_j(1-\rho_j)/\gamma_j}+\sum\limits_{j\in bg}\sqrt{\rho_j/\alpha_j}}. \qquad (4.7)$$

Remark 4.3

It follows from (4.1) that, in the practically relevant case that all mean service times are equal, minimizing $E\hat{\mathbf{V}}$ amounts to minimizing $\sum\lambda_i E\mathbf{W}_i \;/\; \sum\lambda_i$, the overall mean waiting time.

We next turn to the case of random polling with $s_{ij}=s_i$, $s_{ij}^{(2)}=s_i^{(2)}$ for all $i,j\in\{1,...,N\}$. This is the case mentioned below (3.6); minimization of $E\hat{\mathbf{V}}$ w.r.t. $p_1,\dots,p_N$, under the condition $p_1+\cdots+p_N=1, p_1\geqslant 0,\cdots,p_N\geqslant 0$, amounts to

$$Min\left[\frac{\sum\limits_{i=1}^{N}\lambda_i\beta_i^{(2)}}{2(1-\rho)}-\frac{\sum\limits_{i=1}^{N}p_is_i}{1-\rho}\sum_{k\in e}\frac{\rho_k^2}{p_k}+\frac{\sum\limits_{i=1}^{N}p_is_i}{1-\rho}\sum_{k=1}^{N}\frac{\rho_k}{p_k}-\rho_is_i+\rho\frac{\sum\limits_{i=1}^{N}p_is_i^{(2)}}{2\sum\limits_{i=1}^{N}p_is_i}\right], \qquad (4.8)$$

s.t.

$$p_1+\cdots+p_N=1,\quad p_1\geqslant 0,\quad\cdots,p_N\geqslant 0.$$

This problem has to be solved numerically. Application of a modified-Newton algorithm (procedure E04LAF from the NAG-library) yields the optimal visit probabilities in a straightforward manner.

5. Optimization of polling tables

Consider the polling system with a polling table, as described in Section 3. In the present section we are interested in the following problem. Suppose that for given arrival, service and switchover processes and service disciplines at the queues, the system designer still has the freedom to choose the polling table. He wants to choose it such that the mean steady-state amount of work in the system is minimized.

As in the case of the random polling system, the mean workload $E\hat{\mathbf{V}}$ is linearly related to $\sum \rho_n E\mathbf{W}_n$ according to (4.1). Hence, for a polling table with only exhaustive and gated service, minimization of $E\hat{\mathbf{V}}$ over all possible polling tables amounts to minimization of the expression in the righthand side of the pseudoconservation law (3.5) over all such tables. If (an upper bound on) the size of the table, M, is given, then this requires the solution of an integer programming problem (S. Browne [private communication]). Below we will be concerned with the case in which there is no restriction on the table size. The number of possible tables is now unlimited, and it is a priori not clear whether a given 'good' table cannot be improved upon by taking a much larger table with a very similar structure (for example: replace a 60-entry table with 59 $\times Q_1$ followed by once Q_2, by a 6001 entry table composed of 99 subsequent such 60-entry patterns followed by 60 $\times Q_1$ and once Q_2).

In this section we present an approximate approach to the problem of choosing an optimal polling table. The approach consists of three steps:

Step 1. Determine 'good' visit frequencies of all queues in the table.
Step 2. Based on these frequencies, determine a 'good' table size M and the numbers of occurrence of each queue in the table.
Step 3. Given this M and these numbers of occurrence, determine a 'good' ordering of the queues.

Below we discuss each of these steps in some detail. Subsequently we present numerical results to illustrate the accuracy of the procedure.

Step 1. Determination of visit frequencies

This step seems to be the most important of the three. It is described in detail in Subsections 5.1, 5.2 and 5.3. We propose two different approaches for it. Each of these approaches resorts to another system, whose performance (in terms of mean workload) could be radically different from that of the system with a polling table. However, the shape of the performance curve and the optimal operation point of all three systems are very similar to each other. The two approaches can be considered as upper and lower bound approximations, in the sense that their performance curves respectively lie above and below that of the system under consideration.

Step 2. Determination of the table size

Let $f_1,...,f_N$ (with $\sum f_i = 1$) be the visit frequencies obtained in step 1. We want to choose a table size M such that $Mf_1, \ldots, Mf_N$ either are integers or are within a predetermined small positive distance ϵ from an integer (such that the sum of these integers equals M). The resulting integers $m_1, \ldots, m_N$ will be the numbers of occurrence of the N stations. ϵ determines how accurately we wish to approximate the visit frequencies. In a different context, this procedure has been proposed by Panwar et al. [24]. Our experience with that procedure suggests that it is not necessary to take ϵ very small; we have several examples where $\epsilon = 0.25$ (leading to a small table) yields a better result than a much smaller ϵ (that leads to a larger table). We might add that for practical purposes the table can be quite large; a table of several hundred entries should not pose any difficulty in most systems.

Step 3. Determination of the order within the table
The previous steps have determined the table size M and the numbers of occurrence $m_1, \ldots, m_N$ of the queues in this table, with $m_i \approx Mf_i$. We would like to find a table order in which, for each i, the numbers of visits to other queues between consecutive visits to Q_i are (nearly) equal. The following example demonstrates that exact equality cannot always be reached. Let $M=6$, $m_1=1$, $m_2=2$ and $m_3=3$. There is no order in which Q_2 is visited each third time and in which Q_3 is visited each second visit.

This example was taken from Hofri & Rosberg [12]. They consider a conflict-free distributed protocol for the access of N transmission stations to a common channel. They use a weighted Time Division Multiplexing (TDM) protocol; the weight factors refer to the frequencies with which time slots are assigned to the stations. TDM systems are very similar to polling systems. Two main differences, which make TDM better amenable to an exact analysis, are: In TDM each station is visited by the server for a fixed time slot, regardless of whether there are messages present; and TDM does not require switchover times between stations. Hofri & Rosberg investigate two weighted TDM policies for assigning the slots to the stations, for given weight factors $f_1, \ldots, f_N$. One is a 'random' control policy in which each slot is with probability f_i assigned to the i-th station (note the similarity with random polling, where a visit period is assigned instead of a time slot). The other one is a deterministic policy, the 'Golden Ratio policy', which appears to be much better than the random policy. We describe this policy in detail, because we propose to use it also for determining a 'good' polling order.

Let $\phi^{-1} := \frac{1}{2}(\sqrt{5}-1)=0.618034....$ (ϕ^{-1} is also known as the Golden Ratio; it is related to the Fibonacci numbers $F_1, F_2, ...$ via $F_k = [\phi^k-(1-\phi)^k]/\sqrt{5}$.) Put the M numbers $\phi^{-1} mod 1, 2\phi^{-1} mod 1, \ldots, M\phi^{-1} mod 1$ in increasing order (this corresponds to placing them on a circle of unit circumference). Let the j-th smallest number correspond to the j-th position in the table. Assign $\phi^{-1} mod 1, 2\phi^{-1} mod 1, \ldots, m_1\phi^{-1} mod 1$ to Q_1, $(m_1+1)\phi^{-1} mod 1, \ldots, (m_1+m_2)\phi^{-1} mod 1$ to Q_2, etc. The table is thus determined.

Hofri & Rosberg [12], and in particular also Itai & Rosberg [14], discuss a number of properties of the thus obtained assignment. Theorem 5.1 of Itai & Rosberg [14] states that the circle of unit circumference is divided into intervals of at most three different lengths (two if M is a Fibonacci number). As a corollary, they conclude that for each station i, too, there are at most three different interval lengths between successive placements (two if n_i is a Fibonacci number). Consequently, distances between consecutive occurrences of station i in the polling table are also quite evenly spaced. This provides the motivation for using the Golden Ratio (GR) policy in our periodic polling problem.

For the same reason, GR has been applied to several other problems where more or less equidistant spacings of several kinds of items have to be accomplished. See Knuth [18] for an extensive discussion of its properties, and its application to open address hashing (how to distribute keys uniformly over a hashing table); see Itai & Rosberg [14], Hofri & Rosberg [12] and Panwar et al. [24] for performance studies of multi-access protocols which use the Golden Ratio policy, and for a discussion of properties of this policy.

In the next two subsections we propose two different approximations for step 1, the determination of the visit frequencies.

5.1. Determination of visit frequencies: the random polling approximation

Consider a random polling system with the same arrival, service and switchover distributions and the same service policies as in the periodic polling system. In Section 4 it has been shown that the square root probability assignments (4.4) (for exhaustive service) and (4.5) (for gated service) minimize the mean workload in the random polling system with identical switchover time distributions; for the case of different switchover time distributions, the optimal visit probabilities are easily determined numerically by solving the non-linear optimization problem posed in (4.8). In the 'random polling' approximation we propose to choose the visit frequencies $f_1, \ldots, f_N$ of the various queues in the polling table according to exactly the same assignments as have been obtained in the equivalent random polling system. Obviously there is no guarantee that this yields an optimal ratio. On the other hand, since the two systems possess the same properties apart from the fact that in the first one the queues are chosen in random order and in the second one in periodic order, it seems natural that approximately the same visit frequencies should optimize both.

5.2. Determination of visit frequencies: a lower bound approach

Our lower bound approach aims at finding the optimal visit frequencies by focusing at each queue and attempting to place its visits within the cycle in the best (but not necessarily feasible) way for that queue, neglecting the constraints imposed by the visits that have to be paid to the other queues. The approach consists of two stages. First, under the assumption that m_1, m_2, $\cdots, m_N$ are the numbers of visits respectively paid to queues $Q_{1,} \cdots, Q_N$, we attempt to find the best placement of these visits within the cycle. Second we use the delay expressions for these optimal placements and find the optimal values of the variables m_1, m_2, $\cdots, m_N$.

We start with the problem of placing the visits of each queue over the cycle. Since this is a very difficult problem we resort to an "optimistic" approach which attempts to place the visits of queue Q_i within the cycle by ignoring the effects the placement has on the waiting times in the other queues and on the conflicts that may result from such placement. Consider Q_i which receives m_i visits within the cycle. Let $\mathbf{V}_i(j)$ and $\mathbf{I}_i(j)$ ($j = 1, \ldots, m_i$) denote the durations of the jth visit time and the jth intervisit time (succeeding the visit time) of this queue within the cycle. Let $\mathbf{U}_i(j)$ be the amount of work at Q_i at the end of the jth visit.

Using these variables we can now calculate $\rho_i E\mathbf{W}_i$ by calculating the pseudoconservation law for a system consisting of only this queue and in which the intervisits play the role of the switchover periods. Thus we get:

$$\rho_i E\mathbf{W}_i = \rho_i \frac{\lambda_i \beta_i^{(2)}}{2(1-\rho_i)} + \frac{\sum_{j=1}^{m_i} E\mathbf{I}_i(j) E\mathbf{U}_i(j)}{\sum_{j=1}^{m_i} E\mathbf{I}_i(j)} + \frac{\rho_i \sum_{j=1}^{m_i} E\mathbf{I}_i^2(j)}{2\sum_{j=1}^{m_i} E\mathbf{I}_i(j)} \tag{5.1}$$

where the first term is the M/G/1 term and the other two represent $E\mathbf{Y}$, the mean amount of work during the switchover periods; the second one represents the mean amount of work present at the beginning of switchover periods, and the third one the mean amount of work accumulating during switchover periods.

For exhaustive service at Q_i we have $E\mathbf{U}_i(j) = 0$. For gated service at Q_i we get a set of equations $E\mathbf{U}_i(j+1) = [E\mathbf{U}_i(j) + E\mathbf{I}_i(j)\rho_i]\rho_i$, leading to:

$$E\mathbf{U}_i(j) = \frac{\rho_i}{1-\rho_i^{m_i}} \left[\sum_{k=0}^{m_i-1} E\mathbf{I}_i(j+k)\rho_i^{m_i-k} \right] \tag{5.2}$$

where the indices are modulo m_i. Thus we have:

$$\rho_i E\mathbf{W}_i = \rho_i \frac{\lambda_i \beta_i^{(2)}}{2(1-\rho_i)} + \mathbb{1}(gated)\cdot \frac{\sum_{j=1}^{m_i} E\mathbf{I}_i(j) \frac{\rho_i}{1-\rho_i^{m_i}} \sum_{k=0}^{m_i-1} E\mathbf{I}_i(j+k)\rho_i^{m_i-k}}{\sum_{j=1}^{m_i} E\mathbf{I}_i(j)} + \frac{\rho_i \sum_{j=1}^{m_i} E\mathbf{I}_i^2(j)}{2\sum_{j=1}^{m_i} E\mathbf{I}_i(j)} \tag{5.3}$$

where $\mathbb{1}(gated)$ is 1 if the service policy is gated and 0 otherwise (here: exhaustive).

Note that now $\rho_i E\mathbf{W}_i$ is expressed in terms of the first two moments of the variables $\mathbf{I}_i(j)$. The optimization of this expression depends on the actual distribution of these variables, whose determination seems extremely difficult due to their dependence on the visits of the other queues. However, we may continue with our "optimistic" approach and calculate the value of $E\mathbf{I}_i^2(j)$ under the best possible situation for queue i. This situation is achieved when $E\mathbf{I}_i^2(j)$ gets its minimal value, namely when $\mathbf{I}_i(j)$ is deterministically distributed. Thus, the third term now becomes: $\rho_i \sum_{j=1}^{m_i} [E\mathbf{I}_i(j)]^2 \,/\, [2\sum_{j=1}^{m_i} E\mathbf{I}_i(j)]$.

We are now interested in minimizing $\rho_i E\mathbf{W}_i$ by a proper selection of the values of $E\mathbf{I}_i(j)$ ($j = 1, \cdots, m_i$). This minimization has to be done while preserving the invariant of the mean cycle time (which is independent of the cycle chosen!): $EC = s/(1-\rho)$, which in turn implies: $\sum_{j=1}^{m_i} E\mathbf{I}_i(j) = s(1-\rho_i)/(1-\rho)$. Thus, our minimization problem becomes:

$$\text{MIN} \quad \rho_i \frac{\lambda_i \beta_i^{(2)}}{2(1-\rho_i)} + \mathbb{1}(gated)\cdot \frac{\dfrac{\rho_i}{1-\rho_i^{m_i}} \sum\limits_{j=1}^{m_i} E\mathbf{I}_i(j) \sum\limits_{k=0}^{m_i-1} E\mathbf{I}_i(j+k)\rho_i^{m_i-k}}{\sum\limits_{j=1}^{m_i} E\mathbf{I}_i(j)} + \frac{\rho_i \sum\limits_{j=1}^{m_i} [E\mathbf{I}_i(j)]^2}{2\sum\limits_{j=1}^{m_i} E\mathbf{I}_i(j)} \qquad (5.4)$$

$$\text{s.t.} \quad \sum_{j=1}^{m_i} E\mathbf{I}_i(j) = \frac{s(1-\rho_i)}{1-\rho}.$$

Minimization of both the second and the third term of (5.4) under the above constraint is achieved when $E\mathbf{I}_i(j) = E\mathbf{I}_i(k)$ for every $1 \leqslant j,k \leqslant m_i$ and thus the optimal values of $E\mathbf{I}_i(j)$ are given by

$$E\mathbf{I}_i(j) = E\mathbf{I}_i = \frac{s}{1-\rho}\,\frac{1-\rho_i}{m_i}.$$

Substituting the value of $E\mathbf{I}_i$ and $s = \sum_{j=1}^{N} m_j s_j$ yields the optimal value of $\rho_i E\mathbf{W}_i$:

$$\rho_i E\mathbf{W}_i = \rho_i \frac{\lambda_i \beta_i^{(2)}}{2(1-\rho_i)} + \frac{\sum\limits_{j=1}^{N} m_j s_j}{1-\rho}\,\frac{1-\rho_i}{m_i} \left[\mathbb{1}(gated)\cdot \frac{\rho_i^2}{1-\rho_i} + \frac{\rho_i}{2} \right]. \qquad (5.5)$$

Having determined the mean waiting time at queue i (under the optimal placement of its visits) for the numbers of visit $m_1, m_2, \cdots, m_N$, we are now ready to optimize $\sum_{i=1}^{N} \rho_i E\mathbf{W}_i$ with respect to the selection of $m_1, m_2, \cdots, m_N$. This is achieved by solving the (unconstrained) minimization problem:

$$\underset{m_1, m_2, \cdots, m_N}{\text{MIN}} \left[\sum_{j=1}^{N} m_j s_j \right] \left[\sum_{i \in g} \frac{\rho_i(1+\rho_i)}{m_i} + \sum_{i \in e} \frac{\rho_i(1-\rho_i)}{m_i} \right]. \qquad (5.6a)$$

Now we may transform the discrete optimization variables (the m_i's) to continuous ones by considering the variables $f_i = m_i / \sum_{j=1}^{N} m_j$, where f_i is the visit frequency of queue i. The optimization of the visit frequencies now becomes:

$$\underset{f_1, f_2, \cdots, f_N}{\text{MIN}} \left[\sum_{j=1}^{N} f_j s_j \right] \left[\sum_{i \in g} \frac{\rho_i(1+\rho_i)}{f_i} + \sum_{i \in e} \frac{\rho_i(1-\rho_i)}{f_i} \right]. \qquad (5.6b)$$

Note that this is a homogeneous problem in the sense that if $f_1^*, \cdots f_N^*$ is a solution then $xf_1^*, \cdots xf_N^*$ is a solution too (for any x). This is taken care of by the fact that the f_i's are frequencies, namely by adding the constraint $\sum_{i=1}^{N} f_i = 1$. The results of this optimization problem provide the "optimal" visit frequencies, given by:

If Q_i has exhaustive service:

$$f_i^* = \frac{\sqrt{\rho_i(1-\rho_i)/s_i}}{\sum\limits_{j \in g} \sqrt{\rho_j(1+\rho_j)/s_j} + \sum\limits_{j \in e} \sqrt{\rho_j(1-\rho_j)/s_j}}. \qquad (5.7a)$$

If Q_i has gated service:

$$f_i^* = \frac{\sqrt{\rho_i(1+\rho_i)/s_i}}{\sum_{j\in g}\sqrt{\rho_j(1+\rho_j)/s_j} + \sum_{j\in e}\sqrt{\rho_j(1-\rho_j)/s_j}}. \tag{5.7b}$$

Remark 5.1
Note that from (5.3) we can get the value of EW_i and then formulate the more general optimization problem of minimizing the general weighted sum of the mean waiting times: $\sum_{i=1}^{N} C_i EW_i$ where C_i are arbitrary parameters. This problem is outside the scope of this paper and we elaborate on its formulation and solution in [6]. This more general formulation also leads to a generalization of the results derived by Kruskal [19] for a polling model in which all variables are deterministic and the service policy is either exhaustive or gated at all queues.

Remark 5.2
Above we assumed that all the intervisit times are deterministically distributed (as a lower bound for finding the best possible performance of each queue). This assumption can be generalized to other distributions in which $E\mathbf{I}_i^2(j) = \mathbf{K}[E\mathbf{I}_i(j)]^2$ for some arbitrary $\mathbf{K}$. We report on these results in [6].

5.3. A special case supporting the approximate approaches

In Subsections 5.1 and 5.2 we have proposed two approaches for deriving approximations for the optimal visit frequencies to be paid to each station. Both approaches generally yield very similar results, as will be demonstrated in several numerical examples in Section 6. In the case that all switchover time distributions are the same, comparison of (4.4), (4.5) and (5.7a), (5.7b) shows that both approaches yield the same result for exhaustive service at all queues, and slightly different results for gated service at all queues.

As an additional support to these rules we next consider a system for which the optimal visit frequencies (and pattern) can be derived analytically. The idea to study that system is due to Giannakouros and Laloux [11]. It is a system in which all stations belong to either of two types, h or l. There are N_h stations, each with parameters λ_h, β_h, ρ_h, $\beta_h^{(2)}$, s_h and $s_h^{(2)}$, and N_l stations, each with parameters λ_l, β_l, ρ_l, $\beta_l^{(2)}$, s_l and $s_l^{(2)}$. Let m_h and m_l be the number of visits given to an h-station and to an l-station respectively; obviously, all h-stations should receive the same number of visits and the same holds for the l-stations. Without loss of generality let us assume that the h-stations are to receive more attention then the l-stations and let K be the ratio of these numbers: $K = m_h/m_l$.

Giannakouros and Laloux [11] propose the following visit pattern, assuming that K divides N_l and dividing the l-stations into K equal size distinct groups. First, all the N_h h-stations are visited, then the first group of l-stations is visited, then the h-stations are visited again in the same order, then the second group of the l-stations is visited, and so on, where the cycle ends with a visit to the last group of l-stations. Thus, all l-stations occur exactly once in the table.

For this pattern we may derive the pseudoconservation law in a closed-form expression. If the service of the h-stations and the l-stations is exhaustive, then the part of the pseudoconservation law that is relevant for optimization becomes:

$$\begin{aligned} &\frac{\rho_h s_h N_h(N_h-1)}{2(1-\rho)} + \frac{\rho_h s_l N_l N_h}{2K(1-\rho)} + \frac{\rho_l s_l N_l(N_l-1)}{2(1-\rho)} + \frac{\rho_l s_h N_l N_h K}{2(1-\rho)} \\ &+ \frac{\rho}{2}\frac{KN_h s_h^{(2)} + N_l s_l^{(2)}}{KN_h s_h + N_l s_l} + \left[s_h - \frac{\rho_h s}{K(1-\rho)} + \frac{\rho_l s}{1-\rho} - s_l\right]\left[\frac{\rho_h s_l N_l N_h}{2s} - \frac{\rho_l s_h N_l N_h K}{2s}\right], \end{aligned} \tag{5.8}$$

where $s = KN_h s_h + N_l s_l$. Considering this problem as if K is a continuous variable and differentiating (5.8) with respect to K (which is equivalent to the differentiation of the pseudoconservation law itself) yields a condition for the optimal value of K, K^*:

$$\frac{s_h \rho_l(1-\rho_l)N_l N_h}{2(1-\rho)} - \frac{s_l \rho_h(1-\rho_h)N_h N_l}{2K^{*2}(1-\rho)} + \frac{\rho}{2s^2}\left([s_h^{(2)} - s_h^2]s_l - [s_l^{(2)} - s_l^2]s_h\right)N_l N_h = 0. \tag{5.9}$$

Now, if either (1) All switchover periods are identically distributed (i.e., $s_l = s_h$, $s_l^{(2)} = s_h^{(2)}$), or (2) All switchover periods are deterministic (i.e., $s_h^{(2)} = s_h^2$, $s_l^{(2)} = s_l^2$), then the optimal visit ratio is:

$$\frac{m_h^*}{m_l^*} = K^* = \frac{\sqrt{\rho_h(1-\rho_h)/s_h}}{\sqrt{\rho_l(1-\rho_l)/s_l}}. \tag{5.10}$$

This analysis can also be carried out for the cases in which 1) Service of all stations is gated, 2) Service of h-stations is gated and of l-stations is exhaustive, and 3) Service of h-stations is exhaustive and of l-stations is gated. The results of all four cases yield the final result for the optimal visit frequency, f_i^*, for each station:
If Q_i has exhaustive service:

$$f_i^* = \frac{\sqrt{\rho_i(1-\rho_i)/s_i}}{\sum\limits_{j \in e} \sqrt{\rho_j(1-\rho_j)/s_j} + \sum\limits_{j \in g} \sqrt{\rho_j(1+\rho_j)/s_j}}. \tag{5.11}$$

If Q_i has gated service:

$$f_i^* = \frac{\sqrt{\rho_i(1+\rho_i)/s_i}}{\sum\limits_{j \in e} \sqrt{\rho_j(1-\rho_j)/s_j} + \sum\limits_{j \in g} \sqrt{\rho_j(1+\rho_j)/s_j}}. \tag{5.12}$$

This result, which coincides with the rule proposed in (5.7a), (5.7b) and thus supports the approximation approach proposed in Subsection 5.2, is a generalization of the result derived by Giannakouros and Laloux [11]; they consider the case with exhaustive service and deterministic and identical switchover periods between all stations.

6. Numerical results

In this section we numerically evaluate the quality of the approximation algorithms proposed in this paper. Three steps have been recommended in Section 5: 1) Visit frequency selection, 2) Table size selection, and 3) Visit order selection. As stated before it seems that the selection of the table size does not have a significant effect on the results, especially since in most applications there is no difficulty in using tables of tens or even hundreds of entries. The crucial steps are therefore 1) and 3) to which we devote this section. Below we first examine the selection of the visit frequencies and then examine the whole combined procedure.

6.1. Selection of visit frequencies

This step, for which we have proposed two similar approximations (the random polling approximation, and the lower bound approximation), is the most crucial one. In Figure 1, placed after the list of references, we examine the properties of this step (while applying the random polling approximation) by considering a case in which steps 2 and 3 are not required. The example considered represents a typical communications system consisting of many light traffic stations (to be called l-stations) and one heavy traffic station (to be called h-station). Since all l-stations are statistically identical, an optimal visit pattern will consist of one visit to the h-station, followed by a visit to each of the first k l-stations, followed by a visit to the h-station, followed by a visit to each of the next k l-stations, and so on (cf. Subsection 5.3). The only optimization question left open is how large should k be, which is equivalent to the selection of the visit frequency ratio: f_h/f_l.

In the specific example considered we have 12 l-stations whose arrival rates are $\lambda_l = 0.02$; the switchover times and the service times are all deterministic of 1 unit (thus we have $\rho_h = \lambda_h$, $\rho_l = \lambda_l$) and the service policy is gated.

In Figure 1 we plot the exact overall mean waiting time in the system (which in this case is a scale-up of the pseudoconservation law: $EW = (\sum_{i=1}^{N} \rho_i EW_i)/\rho)$ as a function of the visit ratio f_h/f_l. The plot is given for five different values of ρ_h: 0.02, 0.08, 0.18, 0.32 and 0.72. We provide these plots for the random polling system and for the polling table system in Figure 1(a) and 1(b), respectively (to enhance presentation we needed to scale down the $\rho_h = 0.72$ curves by a factor of 9). In the figure we may make the following observations:

1) The shape of each curve in the polling table system (1(b)) resembles that of its mate in the random polling system (1(a)). This is true despite the fact that the levels of the curves are considerably different from each other. Moreover, the minimum of each curve (marked by "*") is achieved at the same location for which the minimum of its mate is achieved.

2) The curves are very shallow at the neighbourhood of their minimum and quite steep at their ends.

These observations suggest the following properties of our approximation:

a) The location of the minimum in the random polling system is an excellent predictor for the location of the minimum in the polling table system.

b) The shallowness of these curves at the neighbourhood of their minimum suggests that even if this prediction somewhat misses (and hits only in the *neighbourhood* of the real minimum) it will still predict an operation point with excellent (very close to minimum) performance.

An important point to note is that while the random polling approximation is very good for predicting the best operation point, it is very bad in predicting the actual performance of the polling table system (the performance values of the systems are 100% and more away from each other!)

Comparison of the Random Polling approximation and the Lower Bound approximation

For step 1 we have proposed two different approximation approaches: the random polling (RP) and the lower bound (LB). An important question, therefore, is how these approximations compare with each other. For the case of identical switchover time distributions, both approaches yield explicit formulas for the frequencies f_i, thus allowing a simple comparison. For exhaustive service systems the two approximations give in this case exactly the same prediction (see Equations (4.4) and (5.7a)). The question then boils down mainly to the comparison of gated service systems.

For these systems the RP approximation suggests that f_i should be proportional to $\sqrt{\rho_i}$ while the LB approximation suggests that it should be proportional to $\sqrt{\rho_i(1+\rho_i)}$. Note that in most cases this difference should be relatively small and that the largest relative difference between these predictions is bounded by $\sqrt{2}$. To see the effect of the difference consider again Figure 1. The locations of the minima of the 5 curves are predicted by the RP approximation to be the points 1, 2, 3, 4 and 6 (marked by '*'). In contrast, the LB approximation will predict these minima to be at the points 1, 2.06, 3.23, 4.55 and 7.79. The reader may realize that, due to the shallowness of these curves, the values of the curves at these points should be very similar to the values marked by '*'. This implies that the performance values of the two approximations are very close to each other.

To provide a more precise comparison of this case we examine the performance of the polling table system when operated under the frequencies suggested by the LB approximation. For all five cases we find that the relative performance difference between the table suggested by RP and the table suggested by LB (either using the GR arrangement or using manual arrangement of the visit order) is less than 1.5%; the largest difference is found for the case $\rho_h = 0.72$.

6.2. Evaluation of the complete algorithm

The rest of this section is devoted to an extensive examination of a variety of cases in which we compare the performance of the polling strategy suggested by the complete approximation algorithm (the combination of either RP or LB with the GR - Golden Ratio - policy for the visit order) with that of the optimal strategy. The results of these comparisons are presented in Tables I.A, I.B, II, III and IV, that appear at the end of this paper. In all the cases considered the service time distributions used are negative exponential; note that due to remark 3.3 the selection of the service time distribution should not affect any of the optimization results.

Discussion of Tables I.A and I.B

Tables I.A and I.B are concerned with an extensive analysis of a 2-queue system. In Table I.A the switchover periods are identical (deterministic with $s_1 = s_2 = 1$) and both queues receive gated service. Q_1 has heavy traffic, whereas Q_2 has low traffic. Step 1 with RP recommends a visit frequency of 3:2, whereas Step 1 with LB recommends a visit frequency of 5:3. In the table we have investigated all 24 ratios m_1:m_2 ranging from 1:1 to 6:4. For each of these ratios we have selected m_1 and m_2 to be the numbers of visits paid to the queues, getting $M = m_1 + m_2$, and we have found the best ordering consisting of m_1, m_2 visits by calculating $E\hat{V}$ for all those orderings. Under the heading 'opt' the lowest such $E\hat{V}$ and the corresponding table have been displayed. Under the heading 'gr', we display the result of applying GR to each (m_1,m_2) combination. The results of gr coincide with those of opt in 18 out of the 24 cases; the largest workload difference in the other 6 cases is 4.1%. This supports the usefulness of the GR policy (step 3). The ratio 5:3 suggested by LB appears to be the optimal ratio; the ratio 3:2 suggested by RP yields a mean workload that is only 0.07% higher. The behaviour of $E\hat{V}$ is here so robust that even taking a ratio ρ_1:ρ_2 = 9:4 would have led to a GR result that is acceptable (5% worse). But it is easy to find examples where such a naive allocation rule would give dramatic errors. E.g., [5] contains a 2-queue example where the ratio ρ_1:ρ_2 gives a 47% error.

Table I.B is similar to Table I.A but considers a case in which $s_1 = 1$ and $s_2 = 1/9$. Naturally the results of gr coincide with those of opt in the same 18 cases; the largest workload difference here slightly exceeds 3%. Again LB gives the 'optimal' ratio, while RP suggests a ratio yielding a 0.31% higher mean workload.

Perturbation of GR

These tables, and many more examples we have investigated, suggest that there are usually many points around the optimum (with a different visit frequency and/or a different table ordering) that yield almost as small a mean workload; but when the visit frequency is too far removed from the optimal one, then $E\hat{V}$ shoots up sharply. This property is best illustrated in Figure 1. This suggests a refinement of the GR policy: apply GR not only to the case (m_1, $\cdots$,m_N), but also to all its $3^N - 1$ neighbours (achieved by adding 1, 0, or -1 to each of the components of the vector (m_1, $\cdots$,m_N)) and then take the best among all these results. A simpler, much cheaper and also very effective alternative would be to check only the $2N$ *immediate* neighbours, obtained by changing just one m_i at a time.

Discussion of Tables II and III

In Table II we provide an extensive examination of a 2-queue system. The table contains 18 cases in which we vary the utilizations, the switchover period durations (mean value) and the service disciplines (denoted by e and g for exhaustive and gated respectively). Under the headline "optimal" we provide the 'optimal' visit pattern (found by an extensive search, like in Tables Ia, Ib) and the corresponding value of $E\hat{V}$. Under the headline "RP-GR approximation" we provide the results achieved for the RP-GR approximations: we report the value of $E\hat{V}$, the relative performance degradation (compared to the optimal $E\hat{V}$) and the predicted visit pattern. In addition, the last column under this headline reports the performance degradation of the best point found by examining the $2N$ immediate neighbours of the point predicted by RP ('%n'). Under the headline "LB-GR approximation" we provide the same data for the LB-GR approximation.

The performance degradation of the RP-GR approximation and of the LB-GR approximation (in comparison to the optimal $E\hat{V}$), in all these cases, is less than 1.5%. The neighbourhood examination of all these cases (for both approximations) almost always yields the optimal value (0% error), with a largest error of 0.4%. In case 1b (exhaustive service) it appears to be optimal to visit Q_1 twice in a row. In the case of exhaustive service it may in most applications be unnatural to have a positive switchover time between consecutive visits to the same queue. Still, if such positive switchover times exist, apparently examples can be given in which repeated visits to the same queue yield a lower $E\hat{V}$

than alternate visit patterns.

Very similar results as for Table II are obtained in Table III, which considers eighteen 3-queue cases. Here the difference between the 'optimal' results and those of RP-GR (also LB-GR) once equals 7.1%; step 1 does suggest the right ratio, but GR does not produce the most sensible order. In the rather extreme case of a mean switchover time s_2 which is 16 times the mean service times and the other mean switchover times, LB-GR and RP-GR produce an error of 6.7%. The $2N$ neighbour examination still gives a 1.7% error.

Discussion of Table IV

In Table IV we examine the quality of the approximations for a variety of switchover time distributions. In all cases examined we hold the arrival parameters the same; service times are negative-exponential with mean 1 and $\rho_1=0.63$, $\rho_2=0.18$; the service discipline at both queues is gated (this corresponds to Tables Ia, Ib). We examine 18 cases varying the first moments and the second moments of the switchover periods, taking deterministic and exponential switchover times, and switchover times with $s_i^{(2)} = 10s_i^2$. The structure of the table is similar to that of Table II. The effect of the second moment of the switchover time distributions is not very pronounced (in fact, exhaustive service would lead to a more pronounced effect). Although the LB-GR approximation does not take the second moment of the switchover time distribution into account, the results are so robust that the largest error still is below 1.5%.

6.3. Conclusions from the numerical experiments

The assignment of visit frequencies (step 1) performs excellently, for the RP approach as well as for the LB approach. Ratios in the direct neighbourhood of the obtained ratio usually yield results of comparable quality.

The procedure for determining the table size (step 2) seems to be the least crucial part of the approach. In fact, in all the examples considered we have not seen a single case in which the selection of the table size affected performance considerably.

The Golden Ratio policy for determining the exact visit order of the queues in the table generally works very well. It is extremely easy to apply, and the mean workload that it produces hardly ever exceeds the mean workload for the best order for given table entries $(m_1, \ldots, m_N)$ by more than a few percent.

The combined procedure, with the refinement of applying GR also to the neighbours of the table entry vector $(m_1, \ldots, m_N)$ found via steps 1 and 2, has been tested for a large number of queueing models with high, medium and low traffic, and with exhaustive and/or gated service disciplines. In each case, the mean workload was also calculated for ALL tables of 'reasonable' size and table entry vector. In almost all cases these approaches led to the same result; the largest relative difference was 1.7%.

7. Summary and plans for the future

This paper has been devoted to studying the efficient operation and optimization of polling systems. We have focused on the problem of deriving an efficient polling table which will minimize the mean workload in the system. We have proposed an approximation algorithm which derives an efficient polling table using two major steps: 1) Selection of visit frequencies (using either the random polling approach or the lower bound approach) and 2) Selection of visit order using the Golden Ratio procedure. The numerical examination shows that the approximation algorithm produces polling tables which are very close to the optimal ones (and in many cases are identical to them). Moreover, due to the shallowness of the objective function at the neighbourhood of its minimum, even if the approximation algorithm slightly misses the optimal operation point, the performance of the table produced by the algorithm is extremely close to that of the optimal one.

Presently we are working on extending the approximation approach to the more general objective function $\sum_{i=1}^{N} C_i EW_i$. The quality of the results obtained in this paper gives us the hope that a good algorithm can be devised for that optimization problem as well. The results of that study are to be reported in [6]. We also hope that similar rules work reasonably well for a much larger class of systems; e.g., systems with more general arrival processes. This issue is left for future investigation.

Acknowledgement
The authors are indebted to J.F.A. Staaijen for his contribution to the numerical calculations.

References

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

[2] Boxma, O.J. (1989). Workloads and waiting times in single-server systems with multiple customer classes. *Queueing Systems* **5**, 185-214.

[3] Boxma, O. J., Groenendijk, W.P. (1987). Pseudo-conservation laws in cyclic-service systems. *J. Appl. Prob.* **24**, 949-964.

[4] 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; to appear in IEEE Trans. Commun.*

[5] Boxma, O.J., Levy, H., Weststrate, J.A. (1990). Optimization of polling systems. *To appear in: P.J.B. King, I. Mitrani, R.J. Pooley (eds.), Performance '90 (North-Holland Publ. Cy., Amsterdam).*

[6] Boxma, O.J., Levy, H., Weststrate, J.A. (1990). Paper in preparation.

[7] Boxma, O.J., Weststrate, J.A. (1989). Waiting times in polling systems with Markovian server routing. *In: G. Stiege and J.S. Lie (eds.), Messung, Modellierung und Bewertung von Rechensystemen und Netzen (Springer, Berlin) pp. 89-104.*

[8] Browne, S., Yechiali, U. (1988). Dynamic scheduling in single server multi-class service systems with unit buffers. *Report Graduate School of Business, Columbia University (NY).*

[9] Browne, S., Yechiali, U. (1989). Dynamic priority rules for cyclic-type queues. *Adv. Appl. Prob.* **21**, 432-450.

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

[11] Giannakouros, N.P., Laloux, A. (1989). On the usefulness of the pseudoconservation law in the performance analysis of service systems with deterministic polling. *Report Telecommunications Laboratory, University of Louvain.*

[12] Hofri, M., Rosberg, Z. (1987). Packet delay under the Golden Ratio weighted TDM policy in a multiple-access channel *IEEE Trans. Inform. Theory,* Vol. IT-**33**, 341-349.

[13] Hofri, M., Ross, K.W. (1987). On the optimal control of two queues with server set-up times and its analysis. *SIAM J. on Computing* **16**, 399-419.

[14] Itai, A., Rosberg, Z. (1984). A Golden Ratio control policy for a multiple-access channel. *IEEE Trans. Autom. Control,* Vol. AC-**29**, 712-718.

[15] Kleinrock, L. (1964). *Communication Nets - Stochastic Message Flow and Delay.* Dover, New York.

[16] Kleinrock, L. (1965). A conservation law for a wide class of queueing disciplines. *Naval Res. Logist. Quart.* **12**, 181-192.

[17] Kleinrock, L., Levy, H. (1988). The analysis of random polling systems. *Oper. Res.* **36**, 716-732.

[18] Knuth, D.E. (1973). *The Art of Computer Programming, Vol. 3.* Addison-Wesley, Reading (MA).

[19] Kruskal, J.B. (1969). Work-scheduling algorithms: a nonprobabilistic queuing study (with possible application to No. 1 ESS). *Bell System Techn. J.* **48**, 2963-2974.

[20] Levy, H. (1988). Optimization of polling systems: The fractional exhaustive service method.

Report Department of Computer Science, Tel-Aviv University.

[21] Levy, H. (1989). Analysis of cyclic-polling systems with binomial-gated service. *In: T. Hasegawa, H. Takagi and Y. Takahashi (eds.), Performance of Distributed and Parallel Systems (North-Holland, Amsterdam) pp. 127-139.*

[22] Levy, H., Sidi, M. (1990). Polling systems: applications, modeling and optimization. To appear in *IEEE Trans. Commun.*

[23] Levy, H., Sidi, M., Boxma, O.J. (1990). Dominance relations in polling systems. *Queueing Systems* **6**, 155-171.

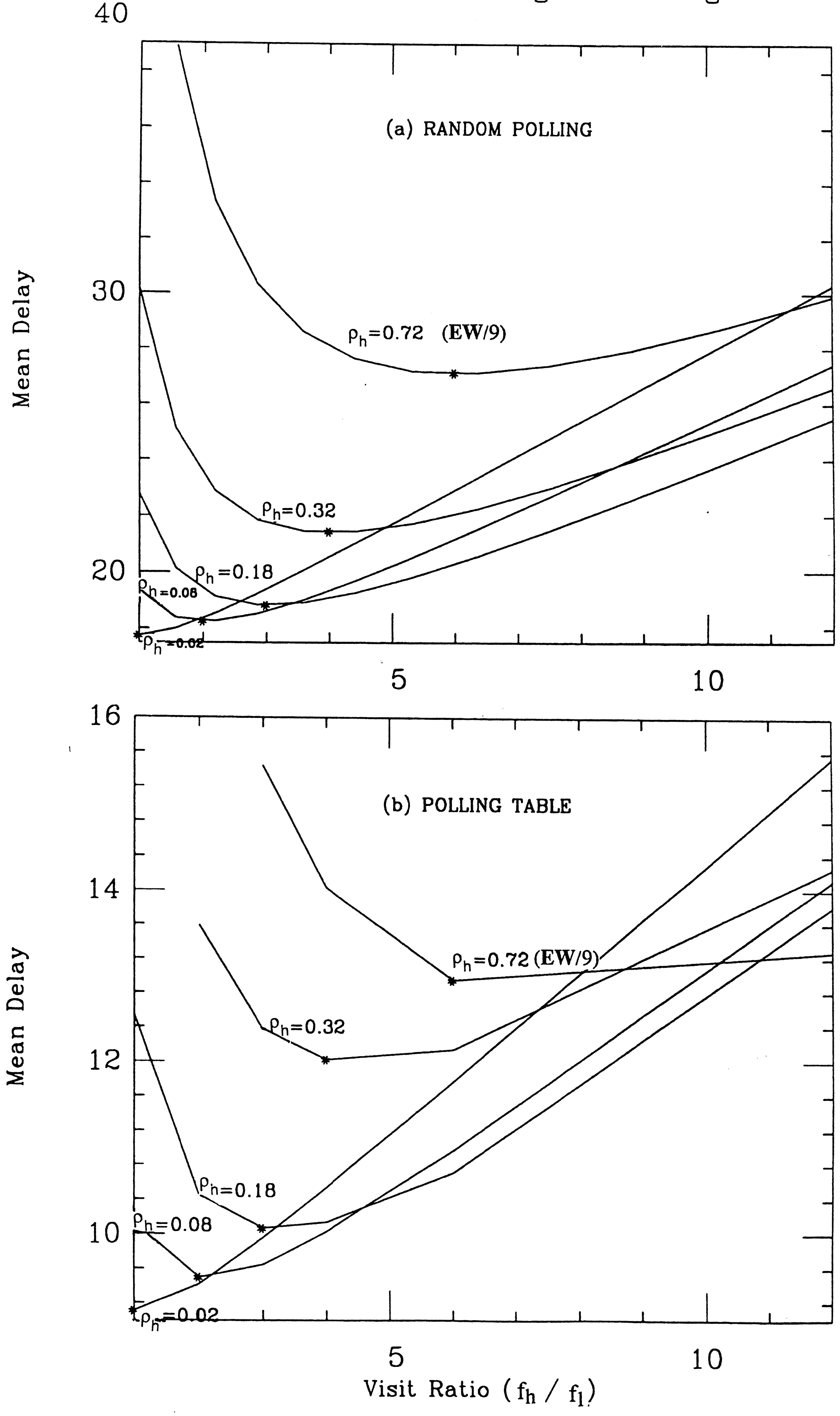
[24] Panwar, S.S., Philips, T.K., Chen, M.-S. (1988). Golden Ratio scheduling for low delay flow control in computer networks. *Report RC 13642, IBM Thomas J. Watson Research Center, Yorktown Heights (NY).*

[25] Servi, L.D. (1986). Average delay approximation of M/G/1 cyclic service queues with Bernoulli schedules. *IEEE J. Sel. Areas Commun.,* Vol. SAC-4, 813-822.

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

[27] Takagi, H. (1990). Queueing analysis of polling models. In: *H. Takagi (ed.), Stochastic Analysis of Computer and Communication Systems (North-Holland Publ. Cy., Amsterdam).*

Figure 1: Random Polling vs. Polling Table



$\rho_1 = 0.63,\ \rho_2 = 0.28.$

$s_1 = 1,\ s_2 = 1;\ s_1^{(2)} = 1,\ s_2^{(2)} = 1;$

Service strategy at each queue is gated.

Table Ia

A two-queue case:

optimal and golden ratio ordening for 24 $m_1 : m_2$ ratios;

comparison with the Random Polling and Lower Bound approximations.

	m_2	1		2		3		4	
m_1		table	$\hat{EV}$	table	$\hat{EV}$	table	$\hat{EV}$	table	$\hat{EV}$
1	opt	12	25.503	122	31.007	1222	37.250	12222	43.701
	gr	21	25.503	212	31.007	2212	37.250	22212	43.701
2	opt	211	24.951	2121	25.503	12122	28.264	122122	31.007
	gr	112	24.951	1212	25.503	21212	28.264	212122	31.487
3	opt	2111	26.369	21211	24.959	121212	25.503	2212121	27.363
	gr	1211	26.369	21211	24.959^{r}	212121	25.503	2122121	27.363
4	opt	21111	28.343	211211	24.951	1121212	25.084	21212121	25.503
	gr	21111	28.343	211121	25.561	2121121	25.084	21211212	26.551
5	opt	211111	30.580	2112111	25.664	11211212	24.942^{*}	112121212	25.171
	gr	111121	30.580	1121121	25.664	11211212	24.942^{l}	112121212	25.171
6	opt	2111111	32.972	21112111	26.369	112112112	24.951	1121211212	24.959
	gr	1121111	32.972	11211112	26.719	112121112	25.401	1212121112	25.475

* indicates the 'optimal' table and its value for $\hat{EV}$;

r indicates the table and its value for $\hat{EV}$ as found by the Random Polling approximation;

l indicates the table and its value for $\hat{EV}$ as found by the Lower Bound approximation.

$\rho_1 = 0.63$, $\rho_2 = 0.28$.
$s_1 = 1$, $s_2 = 1/9$; $s_1^{(2)} = 1$, $s_2^{(2)} = 1/81$.
Service strategy at each queue is gated.

Table Ib

A two-queue case:

optimal and golden ratio ordening for 24 $m_1 : m_2$ ratios;

comparison with the Random Polling and Lower Bound approximations.

	m_2	1		2		3		4	
m_1		table	$\hat{EV}$	table	$\hat{EV}$	table	$\hat{EV}$	table	$\hat{EV}$
1	opt	12	18.661	122	18.666	1222	19.234	12222	19.917
	gr	21	18.661	212	18.666[r]	2212	19.234	22212	19.917
2	opt	211	20.520	2121	18.661	12122	18.608[*]	122122	18.666
	gr	112	20.520	1212	18.661	21212	18.608[l]	212122	18.874
3	opt	2111	22.690	21211	19.661	121212	18.661	2212121	18.618
	gr	1211	22.690	21211	19.661	212121	18.661	2122121	18.618
4	opt	21111	25.006	211211	20.520	1121212	19.367	21212121	18.661
	gr	21111	25.006	211121	20.940	2121121	19.367	21211212	19.253
5	opt	211111	27.423	2112111	21.661	11211212	19.974	112121212	19.212
	gr	111121	27.423	1121121	21.661	11211212	19.974	112121212	19.212
6	opt	2111111	29.919	21112111	22.690	112112112	20.520	1121211212	19.661
	gr	1121111	29.919	11211112	22.954	112121112	20.830	1212121112	19.989

[*] indicates the 'optimal' table and its value for $\hat{EV}$;

[r] indicates the table and its value for $\hat{EV}$ as found by the Random Polling approximation;

[l] indicates the table and its value for $\hat{EV}$ as found by the Lower Bound approximation.

Table II

Some two-queue cases: comparison of the 'optimal' polling table with the RP-GR and LB-GR approximations.

	Q_1			Q_2			optimal		RP-GR approximation				LB-GR approximation			
case	dis	ρ_1	s_1	dis	ρ_2	s_2	$\hat{EV}$	table	$\hat{EV}$	%	table	%n	$\hat{EV}$	%	table	%n
1a	e	0.66	1	e	0.06	1	3.574	12	3.597	-0.6	112	0.0	3.597	-0.6	112	0.0
1b	"	"	"	"	"	2.26	4.175	211	4.199	-0.6	1211	0.0	4.199	-0.6	1211	0.0
1c	"	"	"	"	"	0.028	3.087	12	3.124	-1.2	22212	-0.4	3.112	-0.8	2212	-0.3
2a	e	0.12	1	e	0.05	1	0.389	12	0.395	-1.5	21211	0.0	0.395	-1.5	21211	0.0
2b	"	"	"	"	"	1.799	0.463	12	0.465	-0.4	112	0.0	0.465	-0.4	112	0.0
2c	"	"	"	"	"	0.450	0.339	12	0.339	0.0	21	0.0	0.339	0.0	21	0.0
3a	g	0.80	1	g	0.05	1	12.555	211111	12.600	-0.4	21111	0.0	12.555	0.0	111121	0.0
3b	"	"	"	"	"	0.016	10.721	12	10.784	-0.6	212	0.0	10.752	-0.3	21212	0.0
3c	"	"	"	"	"	0.328	11.592	2111	11.670	-0.7	112	0.0	11.592	0.0	1211	0.0
4a	g	0.62	1	g	0.356	1	81.200	1121212	81.200	0.0	2121121	0.0	81.209	-0.0	21211	-0.0
4b	"	"	"	"	"	1.882	97.042	211	97.042	0.0	112	0.0	97.042	0.0	112	0.0
4c	"	"	"	"	"	0.118	58.999	122	58.999	0.0	212	0.0	58.999	0.0	212	0.0
5a	g	0.563	1	g	0.25	1	10.484	11211212	10.487	-0.0	21211	-0.0	10.484	0.0	11211212	0.0
5b	"	"	"	"	"	0.1998	8.173	12	8.272	-1.2	21212	0.0	8.238	-0.8	2122121	0.0
5c	"	"	"	"	"	3.196	15.004	2111	15.004	0.0	1211	0.0	15.004	0.0	1211	0.0
6a	e	0.724	1	g	0.05	1	4.541	12	4.573	-0.7	112	0.0	4.573	-0.7	112	0.0
6b	"	"	"	"	"	2.36	5.271	211	5.301	-0.6	1211	0.0	5.301	-0.6	1211	0.0
6c	"	"	"	"	"	0.263	4.130	12	4.130	0.0	21	0.0	4.130	0.0	21	0.0

Table III

Some three-queue cases: comparison of the 'optimal' polling table with the RP-GR and LB-GR approximations.

	Q_1			Q_2			Q_3			optimal		RP-GR approximation				LB-GR approximation			
case	dis	ρ_1	s_1	dis	ρ_2	s_2	dis	ρ_3	s_3	$\hat{EV}$	table	EV	%	table	%n	EV	%	table	%n
1a	e	0.617	1	e	0.158	1	e	0.015	1	6.087	2121213	6.224	-2.3	21211213	0.0	6.224	-2.3	21211213	0.0
1b	"	"	"	"	"	2.25	"	"	0.25	6.173	213	6.304	-2.1	1312	0.0	6.304	-2.1	1312	0.0
1c	"	"	"	"	"	2.25	"	"	1	7.238	121213	7.478	-3.3	2131121	0.0	7.478	-3.3	2131121	0.0
2a	e	0.325	1	e	0.11	1	e	0.025	1	1.697	121213	1.697	0.0	212131	0.0	1.697	0.0	212131	0.0
2b	"	"	"	"	"	0.198	"	"	0.44	1.370	123	1.441	-5.2	212132	-0.5	1.441	-5.2	212132	-0.5
2c	"	"	"	"	"	1.23	"	"	2.78	2.037	12121213	2.052	-0.7	112131212	-0.3	2.052	-0.7	112131212	-0.3
3a	e	0.385	1	e	0.385	1	e	0.015	1	6.480	212121213	6.943	-7.1	212131212	-0.2	6.943	-7.1	212131212	-0.2
3b	"	"	"	"	"	1	"	"	0.25	6.141	12123	6.141	0.0	31212	0.0	6.141	0.0	31212	0.0
3c	"	"	"	"	"	16	"	"	1	23.936	121213	25.536	-6.7	211131	-1.7	25.536	-6.7	211131	-1.7
4a	g	0.625	1	g	0.225	1	g	0.025	1	18.647	1211211213	18.781	-0.7	112131212	-0.7	18.962	-1.7	1312121112	-0.7
4b	"	"	"	"	"	0.611	"	"	0.227	16.092	121213	16.264	-1.1	31212	0.0	16.092	0.0	212131	0.0
4c	"	"	"	"	"	0.017	"	"	0.025	12.193	2123	12.309	-1.0	222132	-0.1	12.309	-1.0	222132	-0.1
5a	g	0.80	1	g	0.05	1	g	0.05	1	22.857	11112113	23.067	-0.9	211131	-0.0	22.857	0.0	1113112	0.0
5b	"	"	"	"	"	0.328	"	"	0.328	19.719	12113	19.956	-1.2	1312	0.0	19.719	0.0	31211	0.0
5c	"	"	"	"	"	0.228	"	"	0.829	20.786	12112113	20.969	0.9	2131121	0.0	20.786	0.0	11211213	0.0
6a	e	0.66	1	g	0.10	1	g	0.025	1	5.875	121213	5.875	0.0	212131	0.0	5.875	0.0	212131	0.0
6b	"	"	"	"	"	0.122	"	"	0.114	4.686	123	4.721	-0.7	2312	0.0	4.721	-0.7	2312	0.0
6c	"	"	"	"	"	4.401	"	"	1.03	8.433	1213	8.710	-3.3	31211	0.0	8.710	-3.3	31211	0.0

ρ_1=0.63, ρ_2=0.28.
The service strategy at both queues is gated.

Table IV

Influence of the second moment of the switchover time distribution.

	switch characteristics				optimal		RP-GR approximation			LB-GR approximation		
case	s_1	$s_1^{(2)}$	s_2	$s_2^{(2)}$	$\hat{EV}$	$f_1:f_2$	$\hat{EV}$	%	$f_1:f_2$	$\hat{EV}$	%	$f_1:f_2$
1a	0.25	0.0625	1.75	3.0625	20.316	4:1	20.316	0.0	4:1	20.572	-1.3	5:1
1b	"	0.125	"	6.125	20.864	4:1	20.864	0.0	4:1	21.084	-1.1	5:1
1c	"	0.625	"	30.625	25.179	5:1	25.179	0.0	5:1	25.179	0.0	5:1
2a	0.5	0.25	1.5	2.25	22.384	3:1	22.384	0.0	3:1	22.384	0.0	3:1
2b	"	0.5	"	4.5	22.839	3:1	22.839	0.0	3:1	22.839	0.0	3:1
2c	"	2.5	"	22.5	25.569	3:1	25.569	0.0	3:1	25.569	0.0	3:1
3a	0.9	0.81	1.1	1.21	24.465	2:1	24.667	-0.8	3:2	24.465	0.0	2:1
3b	"	1.62	"	2.42	24.909	2:1	25.117	-0.8	3:2	24.909	0.0	2:1
3c	"	8.1	"	12.1	28.461	2:1	28.461	0.0	2:1	28.461	0.0	2:1
4a	1	1	1	1	24.951	5:3	24.959	-0.1	3:2	24.942	0.0	5:3
4b	"	2	"	2	25.397	5:3	25.414	-0.1	3:2	25.397	0.0	5:3
4c	"	10	"	10	29.037	5:3	29.054	-0.1	3:2	29.037	0.0	5:3
5a	1.5	2.25	0.5	0.25	25.503	1:1	25.503	0.0	1:1	25.503	0.0	1:1
5b	"	4.5	"	0.5	26.072	1:1	26.072	0.0	1:1	26.072	0.0	1:1
5c	"	22.5	"	2.5	30.662	1:1	31.034	-1.3	3:4	30.662	0.0	1:1
6a	1.75	3.0625	0.25	0.0625	25.503	1:1	25.857	-1.4	1:2	25.587	-0.3	2:3
6b	"	6.125	"	0.125	26.214	1:1	26.502	-1.1	1:2	26.262	-0.2	2:3
6c	"	30.625	"	0.625	31.659	1:2	31.659	0.0	1:2	31.669	-0.0	2:3