1991

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

Efficient visit frequencies for polling tables: minimization of waiting cost

Department of Operations Research, Statistics, and System Theory Report BS-R9111 April

CWI, nationaal instituut voor onderzoek op het gebied van wiskunde en informatica

CWI is the research institute of the Stichting Mathematisch Centrum, which was founded on February 11, 1946, as a non-profit institution aiming at the promotion of mathematics, computer science, and their applications. It is sponsored by the Dutch Government through the Netherlands organization for scientific research (NWO).

Efficient Visit Frequencies for Polling Tables:

Minimization of Waiting Cost

O.J. Boxma
CWI
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

Polling systems have been used as a central model for the modeling and analysis of many communication systems. Examples include the Token Ring network and a communications switch. The common property of these systems is the need to efficiently share a single resource (server) among N entities (stations). In spite of the massive research effort in this area, very little work has been devoted to the issue of how to *efficiently operate* these systems.

In the present paper we deal with this problem, namely with how to efficiently allocate the server's attention among the N stations. We consider a framework in which a predetermined fixed visit order (*polling table*) is used to establish the order by which the server visits the stations, and we address the problem of how to construct an efficient (optimal) polling table. In selecting a polling table the objective is to minimize the mean waiting cost of the system, a weighted sum of the mean delays with arbitrary cost parameters. Since the optimization problem involved is very hard, we use an approximate approach. Using two independent analyses, based on a lower bound and on mean delay approximations, we derive very simple rules for the determination of efficient polling tables. The two rules are very similar and even coincide in most cases. Extensive numerical examination shows that the rules perform well and that in most cases the system operates very close to its optimal operation point.

1980 Mathematics Classification: 60K25, 68M20.
Keywords: polling table, mean waiting times, minimization of waiting cost.
This is an invited paper for the *Special Issue on Communication Systems* of *Queueing Systems.*

Report BS-R9111
CWI
P.O. Box 4079, 1009 AB Amsterdam, The Netherlands

1. Introduction

Polling systems have been used to model a large variety of applications in which a single resource is shared among customers accumulating in N distinct queues. These applications include: 1) The Token Ring network in which a single communication channel is shared among N stations each equipped with a buffer to store its messages, 2) A switching node in a communication network which processes messages coming from N different sources, 3) Many non-generic applications in which a single program processes requests (messages) coming from N different entities. Many other applications in computer communications and in other fields are commonly represented by this model; for details the reader may refer to Levy and Sidi [1990] and Grillo [1990].

The use of this model in many scientific and engineering applications has caused the emergence of several variations of the model and of various strategies for operating these systems. This emergence, in turn, has triggered a massive research effort spanned over the last two decades. Nonetheless, this research has focused almost solely on analysis issues, namely on devising mathematical procedures for deriving the performance measures of these models. Almost no work has been done on the optimization of these models, namely on the problem of how to operate polling systems efficiently in order to improve their performance. As a result, designers of these systems have not been equipped with effective tools and guidelines for their efficient operation.

This paper is motivated by the need to derive rules for the efficient operation of polling systems. In Boxma, Levy and Weststrate [1990a,b] we addressed the problem of finding efficient polling tables in order to minimize the mean amount of work or, equivalently, the weighted sum $\sum_{i=1}^{N} \rho_i E\mathbf{W}_i$ (where ρ_i and $\mathbf{W}_i$ are the utilization of and the waiting time at queue i) in systems with exhaustive and gated service.

In this work, we are interested in studying the efficient operation of polling systems in a significantly wider framework. First, the objective function considered here is the sum $\sum_{i=1}^{N} \lambda_i c_i E\mathbf{W}_i$ in which λ_i is the arrival rate at queue i and c_i is an arbitrary parameter reflecting the cost of waiting one unit of time at queue *i*. This sum, therefore, reflects the expected waiting cost per time unit under arbitrarily selected (linear) cost parameters. This is perhaps the single most important performance measure of the system. Second, in terms of modeling aspects, we are concerned here with a framework wider than the one used in Boxma, Levy and Weststrate [1990a,b]: we consider systems in which the service policy is either exhaustive, gated or limited-1 or a mixture of those (while previously only exhaustive and gated service were considered).

The model under consideration consists of a single server and N queues, $Q_1, \cdots, Q_N$. The server visits the queues in a fixed order specified by a polling table (periodic polling) in which each queue occurs at least once (cf. Eisenberg [1972], Baker and Rubin [1987]). Common examples are the cyclic order (the table $1, 2, \cdots, N$) and the star topology order (the table $1, 2, 1, 3, 1, ..., 1, N$). Prioritization of the different queues and effect on the system performance can be achieved by controlling the polling table. Our objective is to find a polling table which minimizes the mean waiting cost, $\sum_{i=1}^{N} \lambda_i c_i E\mathbf{W}_i$. Recognizing that the system under consideration is too complicated to yield itself to optimization (in particular, $E\mathbf{W}_i$ is generally unknown), we do not attempt to derive absolute optimal rules of operation. Rather, following the approach taken in Boxma, Levy and Weststrate [1990a,b], we resort to approximate methods.

The main results of this study are formulas and procedures for determining the number of visits to be given to each queue during a cycle in order to optimize the system performance. These formulas are the key component in a procedure for determining efficient visit patterns. We derive these using two different independent approximations. Support for the quality of the rules derived is provided by the fact that the formulas obtained by the different schemes are very similar to each other. Additional support for the rules derived is provided by an extensive numerical examination of a variety of cases in which the polling tables produced by the approximations perform very close to the optimal tables.

The structure of this paper is as follows: In Section 2, we present the model and formulate the problem. In Section 3 a brief review of the literature and preliminary results are presented. In Section 4, we describe the general methodology for table determination. In Sections 5 and 6, we respectively derive efficient visit frequencies based on a lower bound approach and on a mean delay approximation. Section 7 contains numerical results for evaluating the quality of the approaches. Discussion of the operation rules and of the numerical results is given in Section 8.

2. Model and Problem Description

A single server 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, $i=1,\ldots,N$, and the total arrival rate is $\lambda = \sum_{i=1}^{N} \lambda_i$. Customers arriving at Q_i are called class-i customers; their service times are independent random variables B_i with mean β_i and second moment $\beta_i^{(2)}, i=1,\ldots,N$. The offered traffic load, ρ_i, at Q_i is defined as $\rho_i = \lambda_i \beta_i$, $i=1,\ldots,N$, and the total offered load is $\rho = \sum_{i=1}^{N} \rho_i$. When swapping out of Q_i the server incurs a switchover period of type i; switchover durations (of type i) are independent random variables S_i with mean s_i and second moment $s_i^{(2)}$. The interarrival, service and switchover processes are independent stochastic processes.

We consider three service policies in this paper: exhaustive, gated and limited-1. In the exhaustive policy when visiting Q_i, the server will serve this queue until it is completely empty. In the gated policy the server will serve all customers found at the queue at the beginning of the service period. In the limited-1 policy the server will serve exactly one customer at every visit to Q_i, if any customer is present at the queue. The order of service *within a queue* can be chosen arbitrarily as long as the server does not elect customers for service in a way that is based on their actual service time. The order by which the server visits the queues (visit order) is a fixed predetermined arbitrary order. This order is described in a table (*polling table*) in which the number of visits given to Q_i, m_i, is at least 1. The size of a table, which is the total number of visits in a cycle (period) is $M = \sum_{j=1}^{N} m_j$. Reference to all indices in this paper is done modulo M. The *visit frequency* of a station, f_i is given by $f_i = m_i / (\sum_{j=1}^{N} m_j)$. The total switchover time in a cycle is $S := \sum_{i=1}^{N} m_i S_i$, whose mean value is $s = \sum_{i=1}^{N} m_i s_i$.

The waiting time at Q_i is $\mathbf{W}_i$ and the cost imposed on the system of having a customer waiting one unit of time at Q_i is c_i. The expected cost of operating the system per unit of time is thus $\sum_{i=1}^{N} \lambda_i c_i E\mathbf{W}_i$. The problem of interest is that of finding a polling table which minimizes the expected operating cost $\sum_{i=1}^{N} \lambda_i c_i E\mathbf{W}_i$.

The order in which the server 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. The mean value of the time spent by the server 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. Boxma, Groenendijk and Weststrate [1990]. Finally we introduce the $M \times M$ (0,1) matrix $Z = (z_{ij})$, where $z_{ij} = 1$ if none of the table entries $T(i+1), \ldots, T(j)$ equals $T(i)$, and $z_{ij} = 0$ otherwise.

3. A Brief Review of the Literature and Preliminary Results

The mean waiting times $E\mathbf{W}_i$, i=1, $\cdots$,N, in a polling system with an arbitrary polling table and either exhaustive or gated service can be derived, using relatively efficient numerical procedures (Eisenberg [1972], Baker and Rubin [1987], Choudhury [1989]). This derivation requires the solution of a linear set of equations. The derivation of these measures for a system with limited-1 service is much more difficult; recently two numerical procedures have been proposed by Leung [1990] and Blanc [1990a,b]. Nonetheless, since the complexity of these numerical procedures is exponential in the system size they are applicable only to relatively small systems. Alternative analytical methods which are commonly used for large systems are approximations. Such approximations, based on pseudoconservation laws, have been provided for cyclic polling systems with limited-1 service by Boxma and Meister [1986], Srinivasan [1988], Fuhrmann and Wang [1988] and Groenendijk [1989]. These approximations usually perform well under moderate parameters; nonetheless, due to the difficulty of the problem, they may be quite inaccurate for fairly asymmetric systems under heavy load.

The mean duration of the cycle time, EC (the mean time to make one round of the table), under ergodicity conditions is expressed by a very simple formula:

$$EC = \frac{\sum_{i=1}^{N} m_i s_i}{1-\rho} \tag{3.1}$$

The expression can be derived using balance arguments. Note that the expression holds for all three service policies.

The ergodicity condition for systems with either exhaustive or gated service is $\rho < 1$. For systems in which limited-1 service is given to some (or all) of the queues an additional ergodicity condition holds for every queue Q_i which is served according to the limited-1 fashion: $\frac{\lambda_i \sum_{j=1}^{N} m_j s_j}{1-\rho} < m_i$.

A Pseudoconservation Law for Polling Tables

Boxma, Groenendijk and Weststrate [1990] have derived a pseudoconservation law, an exact expression for the sum $\sum_{i=1}^{N} \rho_i E\mathbf{W}_i$, in polling tables with either exhaustive or gated service. The expression of this sum is given by:

$$\sum_{i=1}^{N} \rho_i E\mathbf{W}_i = \rho \frac{\sum_{i=1}^{N} \lambda_i \beta_i^{(2)}}{2(1-\rho)} + \rho \sum_{m=1}^{M} \frac{\sigma_m^{(2)}}{2\sigma} + \tag{3.2}$$
$$\sum_{k=1}^{M} \rho_{T(k)} \sum_{m \neq k} \frac{\sigma_m}{\sigma} z_{km} \sum_{j=k}^{m-1} (\sigma_j + E\mathbf{VI}_{j+1}) + \sum_{j \in \bar{g}} \rho_{T(j)} E\mathbf{VI}_j \sum_{m=1}^{M} \frac{\sigma_m}{\sigma} z_{jm} + \sum_{m \in \bar{g}} \rho_{T(m)} \frac{\sigma_m}{\sigma} E\mathbf{VI}_m$$

where $\bar{g}$ represents the set of gated pseudostations, σ_i, $\sigma_i^{(2)}$ are the first and second moments of the switchover taken after the visit of pseudostation i, and $\sigma = \sum_{i=1}^{M} \sigma_i$. Note that this notation allows a model which is more general than the one presented above; to match this law to our model we have to set $\sigma_i := s_{T(i)}$, $\sigma_i^{(2)} := s_{T(i)}^{(2)}$ and $\sigma := s$.

A special case of (3.2) which will be required later, is a polling table consisting of a single queue, say Q_1, which is visited M times; in this case all pseudostations are of the same index and thus $z_{ij} = 0$ for all i and j. Equation (3.2) reduces in this case to:

$$\rho_1 E\mathbf{W}_1 = \rho_1 \frac{\lambda_1 \beta_1^{(2)}}{2(1-\rho_1)} + \rho_1 \sum_{m=1}^{M} \frac{\sigma_m^{(2)}}{2\sigma} + \mathbb{1}(gated) \sum_{m=1}^{M} \rho_1 \frac{\sigma_m}{\sigma} E\mathbf{VI}_m \qquad (3.3)$$

where $\mathbb{1}(gated)$ is 1 if the service is gated and 0 if it is exhaustive.

4. General Methodology

In Boxma, Levy and Weststrate [1990a,b] we suggested a methodology for determining efficient polling tables, consisting of three steps:

1) Determine relative visit frequencies, $f_i, i = 1, \ldots, N$ for the visits of the different queues (note that $\sum_{i=1}^{N} f_i = 1$).

2) Determine the size of the polling table M, and the number of visits m_i to be given in a cycle to Q_i, $i = 1,\ldots,N$.

3) Determine the specific order by which to arrange the visits assigned in step 2 above.

As observed in Boxma, Levy and Weststrate [1990a,b] the most critical step in deriving efficient polling tables is step 1, on which this work focuses. Step 2 does not seem to be very critical and a table consisting of up to several tens of queues will usually be a good choice. See the above-mentioned papers for a description of the rounding-off procedure. As for step 3, in Boxma, Levy and Weststrate [1990a,b] we recommended the use of the *Golden Ratio* policy (GR) to "evenly" spread the visits of the different queues over the cycle. The detailed discussions of step 3 can be found there. As stated above, the focus of this paper is on the determination of efficient visit frequencies (step 1)).

5. Approximate Operation Rules Based on A Lower Bound

In this approximation we consider a system whose waiting cost forms a lower bound on the waiting cost of the polling system under consideration. We derive the optimal visit frequencies of the lower bound system and suggest to use them as the visit frequencies of the original system (based on the fact that the two systems possess similar characteristics). We start with a derivation of the lower bound.

A Lower Bound for the Mean Delay at Q_i: Exhaustive and Gated Service

To derive the lower bound we focus on Q_i under the assumption that it receives m_i visits in a cycle, and observe that in order to derive its waiting time one can view it as a polling table system consisting of a single queue. In this view Q_i receives m_i visit periods in a cycle and the periods between the visits are called *intervisits*. Let $\mathbf{I}_i(j)$, $\mathbf{VI}_i(j)$ be random variables respectively denoting the durations of jth intervisit and visit periods of Q_i. Note that $\mathbf{I}_i(j)$ consists of the visits given to the other queues (and the switchover periods in between) in the period between the $j-1$st and the jth visits of Q_i.

Using these variables it is now easy to derive an approximate "closed form" expression for the mean delay at Q_i. This is simply done by viewing the system as a single queue polling system (with m_i visits) and computing the pseudoconservation law from (3.3), in which the duration of the jth switchover period is $\mathbf{I}_i(j)$,and in which the dependence of the intervisit times and visit times is ignored. This yields (note that we can admit any *service order within Q_i* which does not elect customers for service in a way that is based on their actual service time, and which hence does not influence $E\mathbf{W}_i$):

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

where $\mathbb{1}(gated)$ is 1 if Q_i is a gated service station and 0 if Q_i is an exhaustive service station. The values $E\mathbf{VI}_i(j)$ $(j=1,\cdots,m_i)$ for a system with gated service can be determined from the simple set of linear equations:

$$E\mathbf{VI}_i(j+1) = \rho_i[E\mathbf{VI}_i(j) + E\mathbf{I}_i(j)] \tag{5.2}$$

which is the mean amount of work arriving at the system during the periods $\mathbf{VI}_i(j)$ and $\mathbf{I}_i(j)$. The solution of this set is:

$$E\mathbf{VI}_i(j) = \frac{\sum_{k=0}^{m_i-1} E\mathbf{I}_i(j+k)\rho_i^{m_i-k}}{1-\rho_i^{m_i}} \tag{5.3}$$

From (5.1) and (5.3) we get the mean waiting time at Q_i:

$$EW_i^{gated} = \frac{\lambda_i\beta_i^{(2)}}{2(1-\rho_i)} + \frac{\sum_{j=1}^{m_i} E\mathbf{I}_i^2(j)}{2\sum_{j=1}^{m_i} E\mathbf{I}_i(j)} + \frac{\sum_{j=1}^{m_i} E\mathbf{I}_i(j) \sum_{k=0}^{m_i-1} E\mathbf{I}_i(j+k)\rho_i^{m_i-k}}{(1-\rho_i^{m_i})\sum_{j=1}^{m_i} E\mathbf{I}_i(j)} \tag{5.4a}$$

$$EW_i^{exhaustive} = \frac{\lambda_i\beta_i^{(2)}}{2(1-\rho_i)} + \frac{\sum_{j=1}^{m_i} E\mathbf{I}_i^2(j)}{2\sum_{j=1}^{m_i} E\mathbf{I}_i(j)} \tag{5.4b}$$

The mean waiting time, therefore, depends on the first and second moments of the intervisit times of Q_i. While it is hard to compute the first two moments of $\mathbf{I}_i(j)$, it is easy to compute the following sum:

$$\sum_{j=1}^{m_i} E\mathbf{I}_i(j) = \frac{s(1-\rho_i)}{1-\rho}. \tag{5.5}$$

This expression results from the well known expression for the mean cycle time: $EC = s/(1-\rho)$, from the simple relation $\sum_{j=1}^{m_i}[E\mathbf{I}_i(j)+E\mathbf{VI}_i(j)] = EC$ and from the relation (being implied by load balance arguments) $\sum_{j=1}^{m_i} E\mathbf{VI}_i(j) = \rho_i EC$.

To derive the lower bound we now consider all possible sets of non-negative random variables $\{\mathbf{I}_i(j)\}$ which obey (5.5) and seek the set that minimizes Equation (5.4a-b). This minimization is achieved via the next proposition and the subsequent lemma.

Proposition 5.1: Let $\{\mathbf{I}_i^*(j)\}$ be a set of random variables which minimizes (5.4a-b). Then $\mathbf{I}_i^*(j)$, $j=1, \cdots, m_i$, is deterministically distributed.

The proposition results from the fact that for any arbitrary random variable $\mathbf{I}_i(j)$ one can select a deterministic random variable $\mathbf{I}_i^*(j)$ whose mean is $E\mathbf{I}_i^*(j) = E\mathbf{I}_i(j)$ and whose second moment obeys $E[(\mathbf{I}_i^*(j))^2] \leqslant E[(\mathbf{I}_i(j))^2]$; the equality holds only if $\mathbf{I}_i(j)$ is deterministically distributed. This selection will obviously reduce the value of (5.4a-b).

Following this proposition the search for the set $\{\mathbf{I}_i^*(j)\}$ can now concentrate on deterministic variables. Equation (5.4a-b), therefore, turns into:

$$EW_i^{gated} = \frac{\lambda_i\beta_i^{(2)}}{2(1-\rho_i)} + \frac{\sum_{j=1}^{m_i}[E\mathbf{I}_i(j)]^2}{2\sum_{j=1}^{m_i}E\mathbf{I}_i(j)} + \frac{\sum_{j=1}^{m_i}E\mathbf{I}_i(j)\sum_{k=0}^{m_i-1}E\mathbf{I}_i(j+k)\rho_i^{m_i-k}}{(1-\rho_i^{m_i})\sum_{j=1}^{m_i}E\mathbf{I}_i(j)} \tag{5.6a}$$

$$EW_i^{exhaustive} = \frac{\lambda_i\beta_i^{(2)}}{2(1-\rho_i)} + \frac{\sum_{j=1}^{m_i}[E\mathbf{I}_i(j)]^2}{2\sum_{j=1}^{m_i}E\mathbf{I}_i(j)} \tag{5.6b}$$

Lemma 5.2: The set of deterministically distributed random variables $\{\mathbf{I}_i^*(j)\}$, $j=1, \cdots, m_i$, which minimizes (5.6a-b), under the constraint (5.5), obeys:

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

Proof: The proof of the claim for (5.6b) is immediate. To prove the claim for (5.6a) we consider the equivalent problem (with simplified notation) of minimizing the function:

$$\underset{\{X_j\}}{MIN} \quad \frac{\sum_{j=1}^{m}X_j^2}{2} + \frac{\sum_{j=1}^{m}X_j\sum_{k=0}^{m-1}X_{j+k}\rho^{m-k}}{1-\rho^m} \tag{5.7}$$

$$s.t. \quad \sum_{j=1}^{m}X_j = constant$$

where reference to indices is done modulo m, X_j, $j=1,...,m$ are non-negative, and $0 \leqslant \rho < 1$.

To conduct this minimization we hold all variables fixed except for X_i and X_j and minimize (5.7) under the constraint that the sum of X_i and X_j is fixed, namely $X_i+X_j = Z$. We thus need to minimize:

$$\frac{X_i^2 + X_j^2}{2} + \frac{(X_i^2 + X_j^2)\rho^m + X_iX_j(\rho^d + \rho^{m-d})}{1-\rho^m} \tag{5.8}$$

in which $d = j-i$. By factoring out $(X_i+X_j)^2 \cdot \frac{1+\rho^m}{2(1-\rho^m)}$, it is easily seen that $X_i = X_j = Z/2$ minimizes the expression. Thus we conclude that equality of X_i and X_j minimizes (5.8) and therefore it follows that equality of all X_i's minimizes (5.7). □

From Proposition 5.1 and Lemma 5.2 we may now conclude that $\mathbf{I}_i^*(j)$ which minimizes (5.4a-b) is deterministically distributed with mean $E[\mathbf{I}_i^*(j)] = \frac{s(1-\rho_i)}{m_i(1-\rho)}$, and that a lower bound for $E\mathbf{W}_i$ is given by:

$$E\mathbf{W}_i^* \;=\; \frac{\lambda_i \beta_i^{(2)}}{2(1-\rho_i)} + \frac{\sum_{j=1}^{N} m_j s_j}{1-\rho} \frac{1-\rho_i}{m_i} \left[\mathbb{1}(gated) \cdot \frac{\rho_i}{1-\rho_i} + \frac{1}{2} \right] \tag{5.9}$$

A Conjectured Lower Bound for the Delay in a Limited-1 System

Similar to the above analysis one may attempt to derive a lower bound for the delay in a single queue polling table system with limited-1 service. This task is however more difficult than the one pursued above, since a pseudo-conservation law for a system with limited-1 service is available only when the number of visits given to each 1-limited queue is one. For this reason it may be difficult to obtain an expression like (5.1).

However, based on the above results we may conjecture that the intervisit time random variables which minimize the mean delay in a single queue polling table system with limited-1 service possess the same properties proved for the exhaustive and gated systems. This idea is expressed in the following conjecture:

Conjecture 5.3: The set of random variables $\{\mathbf{I}_i^*(j)\}, j=1, \cdots, m_i$, which minimizes the mean delay in a single queue polling table with limited-1 service, under the constraint (5.5), obeys:

(a) $\mathbf{I}_i^*(j)$ is deterministically distributed $(j=1, \cdots, m_i)$.

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

Once we adopt this conjecture, we may observe that a system which obeys these conditions behaves like a single queue cyclic polling system since all the intervisit times are identically distributed. We therefore may drop the index j from the variable $\mathbf{I}_i$, and use the pseudo-conservation law for the cyclic polling system with limited-1 service (cf. Boxma and Groenendijk [1987]) to derive the mean delay in this system:

$$E\mathbf{W}_i^* \;=\; \frac{\lambda_i \beta_i^{(2)} \;+\; E\mathbf{I}_i(1+\rho_i)}{2(1-\rho_i-\lambda_i E\mathbf{I}_i)}$$

and substituting $E\mathbf{I}_i \;=\; s(1-\rho_i)/(m_i(1-\rho))$, we get the lower bound for the mean delay:

$$E\mathbf{W}_i^* \;=\; \frac{\dfrac{m_i\lambda_i\beta_i^{(2)}(1-\rho)}{1-\rho_i} \;+\; s(1+\rho_i)}{2(m_i(1-\rho)-\lambda_i s)} \tag{5.10}$$

Optimization Via the Lower Bound

The lower bound approximation finds the set $\{m_i\}$, $i=1,...,N$ which solves for:

$$\mathbf{W}^* \;=\; \underset{m_1,\dots,\,m_N}{MIN} \sum_{i=1}^{N} c_i\lambda_i E\mathbf{W}_i^* \tag{5.11}$$

Thus we need to solve the following problems:

$$\textit{exhaustive}: \qquad \underset{m_1,\,\cdots,m_N}{MIN} \left[\sum_{j=1}^{N} m_j s_j\right]\left[\sum_{i=1}^{N}\frac{c_i\lambda_i(1-\rho_i)}{m_i}\right] \tag{5.12a}$$

$$\textit{gated}: \qquad \underset{m_1,\,\cdots,m_N}{MIN} \left[\sum_{j=1}^{N} m_j s_j\right]\left[\sum_{i=1}^{N}\frac{c_i\lambda_i(1+\rho_i)}{m_i}\right] \tag{5.12b}$$

$$\textit{limited}-1: \qquad \underset{m_1,\,\cdots,m_N}{MIN} \sum_{i=1}^{N} c_i\lambda_i \cdot \frac{\dfrac{m_i\lambda_i\beta_i^{(2)}(1-\rho)}{1-\rho_i} \;+\; s(1+\rho_i)}{2(m_i(1-\rho)-\lambda_i s)} \tag{5.12c}$$

These are homogeneous unconstrained minimization problems: if $(m_1^*, \cdots, m_N^*)$ yields a minimum then the same holds for $K(m_1^*, \cdots, m_N^*)$ for any $K \neq 0$.

These problems can be solved by simple minimization techniques. The solutions of these problems yield the optimal visit numbers (up to a multiplicative factor) for the lower bound:

$$\textit{exhaustive}: \qquad m_i \sim \sqrt{c_i\lambda_i(1-\rho_i)/s_i} \tag{5.13a}$$

$$\textit{gated}: \qquad m_i \sim \sqrt{c_i\lambda_i(1+\rho_i)/s_i} \tag{5.13b}$$

$$\textit{limited}-1: \qquad m_i \sim \lambda_i + (1-\rho-\sum_{j=1}^{N}\lambda_j s_j)\cdot\frac{\left[c_i\lambda_i\left[\frac{\lambda_i^2\beta_i^{(2)}}{1-\rho_i}+1+\rho_i\right]/s_i\right]^{1/2}}{\sum_{j=1}^{N}s_j\left[c_j\lambda_j\left[\frac{\lambda_j^2\beta_j^{(2)}}{1-\rho_j}+1+\rho_j\right]/s_j\right]^{1/2}} \tag{5.13c}$$

Remark 5.1: From (5.9), (5.10) and (5.13a-c) one can obtain the mean waiting cost $\mathbf{W}^*$. This forms a lower bound for the mean waiting cost in the related polling table system. For example, in the exhaustive system we have:

$$\mathbf{W}^* = \sum_{i=1}^{N}\frac{c_i\lambda_i^2\beta_i^{(2)}}{2(1-\rho_i)} + \frac{\left[\sum_{i=1}^{N}\sqrt{c_i\lambda_i(1-\rho_i)s_i}\right]^2}{2(1-\rho)}$$

Remark 5.2: In a similar way to (5.13a-c) one can derive the lower bound and "optimal" visit numbers for a system with mixed service policies, namely, a system in which different stations may be served by different service policies (exhaustive, gated or limited-1).

Remark 5.3: For the special case $c_i \equiv \beta_i$, the visit frequencies determined by (5.13a-b) have been obtained in Boxma, Levy and Weststrate [1990b]. There the lower bound approach (and, as an alternative, exact optimization results for visit frequencies in *random polling systems* with the same traffic characteristics) has been used for optimizing the specific objective function $\sum_{i=1}^{N}\rho_i E\mathbf{W}_i$.

Finally we state the lower bound approximation:

LOWER BOUND APPROXIMATION: Use a set of visit numbers which are proportional to the numbers derived in (5.13a-c) to obtain efficient polling tables.

6. Approximate Operation Rules Based on Mean Delay Approximation

The operation rules derived in this section are based on approximate expressions for the mean delays in the polling table system. Our approach is to derive simple expressions for the mean delay $E\mathbf{W}_i$ and to derive efficient visit frequencies by optimizing the objective function $\sum_{i=1}^{N}c_i\lambda_i E\mathbf{W}_i$ where the approximations of the mean delay are substituted for EW_i. The visit frequencies which optimize that objective function are then recommended for use in the polling table system.

To derive an approximation for the mean delay in a polling table system, we first recall previously derived approximations for the mean delay in the *purely cyclic system*. These expressions have been derived by Everitt [1986] for exhaustive and gated service, by Boxma and Meister [1986] for 1-limited service, and by Groenendijk [1989] for a mixture of those. For all three policies, the mean waiting time EW_i is related to the mean residual time ERC_i of a cycle of the server, starting and ending at Q_i:

$$\textit{exhaustive}: \quad EW_i = (1-\rho_i)ERC_i, \tag{6.1a}$$

$$\textit{gated}: \quad EW_i = (1+\rho_i)ERC_i, \tag{6.1b}$$

$$1-\textit{limited}: \quad EW_i \approx \frac{1-\rho+\rho_i}{1-\rho-\lambda_i \sum_{j=1}^{N} s_j} ERC_i. \tag{6.1c}$$

Formula (6.1a) is exact when the cycle starts with a departure of the server from Q_i, and (6.1b) is exact when the cycle starts at an arrival of the server to Q_i. Note that (6.1c) is only an approximation. Equations (6.1a), (6.1b) and (6.1c) have been used (in the references mentioned above) for deriving mean delay approximations in the purely cyclic system. This has been done by arguing that ERC_i is approximately the same for all i, and by using the pseudo-conservation law. Below we use that approach to derive the mean delay in the polling table system.

Assume that Q_i occurs m_i times in a table, $i=1,...,N$. We assume that these visits to Q_i are spread as evenly as possible. Denoting by $ERSC_i$ the mean residual time of a subcycle for Q_i, a subcycle being the interval between two successive server visits to Q_i, we propose the following approximation for EW_i:

$$\textit{exhaustive}: \quad EW_i \approx (1-\rho_i)ERSC_i, \tag{6.2a}$$

$$\textit{gated}: \quad EW_i \approx (1+\rho_i)ERSC_i, \tag{6.2b}$$

$$1-\textit{limited}: \quad EW_i \approx \frac{1-\rho+\rho_i}{1-\rho-(\lambda_i/m_i)\sum_{j=1}^{N} m_j s_j} ERSC_i. \tag{6.2c}$$

The arguments leading to (6.2a) and (6.2b) are simple extensions of those leading to (6.1a) and (6.1b), (cf. p. 209 of Boxma [1989] for a discussion of those arguments). The 1-limited approximation (6.2c) is derived as follows. A customer arriving at Q_i has to wait a residual subcycle; when he meets $\mathbf{X}_i$ customers in Q_i upon his arrival, he also has to wait $\mathbf{X}_i$ subcycles $\mathbf{SC}_i^+$ - which differ from ordinary subcycles in the sense that each of these subcycles contains a service at Q_i. Thus:

$$E\mathbf{W}_i \approx E\mathbf{RSC}_i + E\mathbf{X}_i E\mathbf{SC}_i^+. \tag{6.3}$$

Using the PASTA property (which implies that $E\mathbf{X}_i$ equals the mean number of customers waiting in Q_i at an arbitrary epoch) and Little's formula, we can write $E\mathbf{X}_i = \lambda_i E\mathbf{W}_i$. Similar to the 1-limited cyclic polling approximation in Boxma and Meister [1986], we use a traffic balance argument to write

$$E\mathbf{SC}_i^+ \approx \beta_i + \sum_{j=1}^{N} m_j s_j / m_i + (\rho - \rho_i) E\mathbf{SC}_i^+. \tag{6.4}$$

This equality is based on the argument that the subcycle $\mathbf{SC}_i^+$ consists of the service of one customer at Q_i, of the mean switchover duration in that cycle and of the amount of service being given to the other queues during an $\mathbf{SC}_i^+$ cycle. This yields

$$E\mathbf{SC}_i^+ \approx \frac{\beta_i + \sum_{j=1}^{N} m_j s_j \ / \ m_i}{1-\rho+\rho_i}. \tag{6.5}$$

Substitution of (6.5) in (6.3) leads to (6.2c). Approximations (6.2a), (6.2b) and (6.2c) can also be used for a polling table with a mixture of exhaustive, gated and 1-limited service policies.

Having derived (6.2a), (6.2b) and (6.2c), we finally need a bold assumption to get rid of the unknown $E\mathbf{RSC}_i$. Since Q_i has m_i subcycles in a cycle, the mean duration of each sub-cycle is $1/m_i$ of that of a complete cycle; this property will hold when all type-i subcycles are of the same length, which is the case in the neighbourhood of the optimal operation point. We now assume that the mean residual of a sub-cycle is a constant A times its mean value. This property will hold for a variety of cases in which the different subcycles are of similar distribution (in particular when all sub-cycles have the same Erlang or Gamma distribution). Hence

$$E\mathrm{RSC}_i = \frac{A}{m_i}E\mathrm{C}, \quad i=1,\dots,N, \tag{6.6}$$

with A some unknown constant and $E\mathrm{C}$ the mean time to complete one cycle of the table; the value of $E\mathrm{C}$ is well known and is given in (3.1).

We expect the accuracy of approximation (6.6) to increase with decreasing randomness of the cycle. In the case of low traffic, rather symmetric queues and 1-limited service (which has a relatively small coefficient of variation of the visit period of a queue), (6.6) can be expected to lead to good results.

Substitution of (6.6) into (6.1a), (6.1b) and (6.1c), leads to the following mean delay approximation:

$$\textit{exhaustive}: \; E\mathrm{W}_i \approx \frac{A}{1-\rho}\cdot(1-\rho_i)\cdot\frac{\sum_{j=1}^{N} m_j s_j}{m_i}, \tag{6.7a}$$

$$\textit{gated}: \qquad E\mathrm{W}_i \approx \frac{A}{1-\rho}\cdot(1+\rho_i)\cdot\frac{\sum_{j=1}^{N} m_j s_j}{m_i}, \tag{6.7b}$$

$$1-\textit{limited}: \; E\mathrm{W}_i \approx \frac{A}{1-\rho}\cdot\frac{1-\rho+\rho_i}{1-\rho-\lambda_i\sum_{j=1}^{N} s_j}\cdot\frac{\sum_{j=1}^{N} m_j s_j}{m_i}. \tag{6.7c}$$

Finally, optimization of $\sum_{i=1}^{N} c_i\lambda_i E\mathrm{W}_i$, using the expressions (6.7a), (6.7b) and (6.7c) is similar to the optimization problem posed in equations (5.12a-c), yielding the following results (note that there is no need to determine A):

$$\textit{exhaustive}: \quad m_i \sim \sqrt{c_i\lambda_i(1-\rho_i)/s_i} \tag{6.8a}$$

$$\textit{gated}: \quad m_i \sim \sqrt{c_i\lambda_i(1+\rho_i)/s_i} \tag{6.8b}$$

$$\textit{limited}-1 \quad m_i \sim \lambda_i + (1-\rho-\sum_{k=1}^{N}\lambda_k s_k)\frac{\sqrt{c_i\lambda_i(1-\rho+\rho_i)/s_i}}{\sum_{j=1}^{N} s_j\sqrt{c_j\lambda_j(1-\rho+\rho_j)/s_j}}. \tag{6.8c}$$

Note that in (6.8c) $\rho+\sum\lambda_k s_k = \sum\lambda_k(\beta_k+s_k)<1$, as otherwise the ergodicity condition of the system would be violated.

Remark 6.1: If we add the constraint $EC = \sum m_j s_j / (1-\rho) = C^*$ to the unconstrained problem (5.12c) for 1-limited service, thus prescribing the mean cycle time of the polling table, then the following minimization problem results:

$$Min \sum_{i=1}^{N} \frac{c_i \lambda_i (1-\rho+\rho_i)}{m_i - \lambda_i C^*} \tag{6.9}$$

$$Sub \frac{\sum_{j=1}^{N} m_j s_j}{1-\rho} = C^*. \tag{6.10}$$

The optimal solution of the homogeneous unconstrained problem can be scaled to satisfy (6.10). This implies that the visit numbers given by (6.8c), after a scaling with the factor C^*, also solve the constrained minimization problem (6.9)-(6.10). Advantages of this approach are the simple structure of (6.9-10) and the close similarity with Kleinrock's linear Capacity Assignment problem (Section 5.7 of Kleinrock [1976]), which enables one to translate Kleinrock's [1976] solution (5.26) immediately into (6.8c). Similar to Kleinrock's solution interpretation, one can interpret (6.8c): Station Q_i should be visited at least $\lambda_i C^*$ times during a cycle, as this is the mean number of arrivals during one cycle; the remaining 'capacity' is allocated according to a square root assignment. The latter assignment is very similar to those obtained for exhaustive and gated service in (6.8a) and (6.8b).

The derived rule is therefore expressed as:

MEAN DELAY APPROXIMATION: Use a set of visit numbers which are proportional to the numbers derived in (6.8a-c) to obtain efficient polling tables.

7. Numerical Results

In this section we numerically examine the quality of the approximated rules derived in Sections 5 and 6. The examination is conducted by a systematic exploration of a wide variety of cases and by comparing the mean waiting cost in a system operated by the rules suggested in this paper to that of an "optimally" operated system. The performance of an "optimally" operated system is found by an organized search of a wide variety of cases and selecting the one that yields the minimum value. In the case of gated service, we have used a numerical procedure of G. Choudhury for polling tables; in the other cases, the power series algorithm of J.P.C. Blanc has been applied.

A whole plethora of cases and effects can be studied: different service disciplines, small or large number of queues, choice of service time distributions and switchover distributions and their parameters. Our previous studies (Boxma, Levy and Weststrate [1990a,b]) have suggested that the optimal visit rules are not very sensitive to the choice of those distributions, and therefore we do not vary them much here.

In the comparison we put emphasis on examining the quality of the visit frequency determination step (step 1 in the procedure described in Section 4). To achieve this comparison we examine cases in which the determination of visit patterns for each set of visit frequencies can be easily generated manually; this manual pattern determination is done by an even spacing of the queue visits. In addition we examine the quality of the complete procedure suggested in Section 4. In this comparison the determination of the visit pattern is done using the Golden Ratio procedure.

The main two parts of this section deal with examining the approximation for the gated service system and the limited-1 system. Mixtures of exhaustive and limited-1 service are also considered. We have not included cases with exclusively exhaustive service, for the following two reasons: (i) space limitations, and (ii) the numerical experiments in Boxma et al. [1990a,b] for the case $c_i \equiv \beta_i$, and some cases we have run with $c_i \neq \beta_i$, show that our optimization approaches for exhaustive and for gated service systems are not only very similar, but also lead to approximations of very comparable (excellent) quality.

Gated Service System

In this examination we consider a system with several identical stations and an additional station that differs from them. The cases are constructed in a way that each time several parameters are controlled (by holding them fixed) while the others are examined (by being perturbed). We consider six cases in which the effects of the following parameters on the approximation quality are examined: 1)

Effect of arrival rates, 2) Effect of switch-over parameters, 3) Effect of cost parameters, 4) Effect of service time parameters, 5) Effect of the system size, and 6) Effect of mixed parameters. A detailed description of the cases is next given.

Case Gated-1: Effect of Arrival Rates We consider a system consisting of 12 identical stations and an additional station that differs from them. All parameters for all stations are identical except for the arrival rates. The identical parameters are: B_i is deterministic with mean 1, S_i is exponential with mean 1 and c_i is constant for all 13 queues: $c_i = 1.0$. The arrival rates of 12 stations are fixed to $\lambda_i = 0.02$ while that of the additional station (Q_1) is varied and gets the values 0.02, 0.075, 0.16, 0.26 and 0.49. The visit frequencies examined in order to find the optimum are: $f_1 {:} f_i$ = 4:5, 1:1, 4:3, 3:2, 2:1, 12:5, 3:1, 7:2, 4:1, 5:1, 6:1, 8:1 and 12:1. For each of these frequencies the visit pattern is selected manually by evenly spacing the visits. In Table Gated-1 we present the results of this case: The first column contains the arrival rates, the next 2 columns contain the results of the optimal pattern (the cost and the number of visits given to the queues), the next 4 columns contain the results of the pattern predicted by the approximation: the cost, the percent error (with respect to the optimum), the visit frequency ratio predicted by the approximation ($f_1 {:} f_i$) and the actual number of visits considered ($m_1 {:} m_i$, which usually results from rounding $f_1 {:} f_i$). The last two columns contain the results of the lower bound: the cost and the relative difference between the bound and the optimum found.

A similar case is examined in Table Gated-1-Det. Here everything is identical to Table Gated-1, except that the switch-over periods here are deterministic.

Case Gated-2: Effect of Switch-Over Periods This case is similar to Case Gated-1, except that here the parameters perturbed are the mean values of the switch-over periods. The identical parameters are: B_i is deterministic with mean 1, λ_i is equal to 0.06, and c_i is constant: $c_i = 1$. The switch-over periods of the 12 stations are taken to be exponential with mean 1. The switchover period of Station 1 is exponential with varying intensity: 1/36, 1/16, 1/9, 1/4, 1, 4, 9 and 16. The visit frequency ratios examined (in order to find the optimum) are: 1:5, 1:4, 1:3, 2:5, 1:2, 2:3, 1:1, 4:3, 3:2, 2:1, 12:5, 3:1, 7:2, 4:1, 5:1, 6:1, 8:1 and 12:1. Table Gated-2 is similar to Table Gated-1 except for the first

Table Gated-1: Effect of Arrival Rates (exp. switch-over) [i= 2,...,13]

	optimal pattern		mean delay approx.				lower bound	
λ_1,λ_i	cost	m_1:m_i	cost	%err	f_1:f_i	m_1:m_i	cost	diff.
0.02, 0.02	0.0501	1:1	0.0501	0	1.00:1	1:1	0.0466	7.4%
0.075, 0.02	0.0628	2:1	0.0628	0	1.99:1	2:1	0.0584	7.6%
0.16, 0.02	0.0833	3:1	0.0833	0	3.02:1	3:1	0.0770	8.2%
0.26, 0.02	0.1145	4:1	0.1145	0	4.01:1	4:1	0.1055	8.5%
0.49, 0.02	0.2713	6:1	0.2713	0	5.98:1	6:1	0.2590	8.9%

Table Gated-1-Det: Effect of Arrival Rates (det. switch-over) [i= 2,...,13]

	optimal pattern		mean delay approx.				lower bound	
λ_1,λ_i	cost	m_1:m_i	cost	%err	f_1:f_i	m_1:m_i	cost	diff.
0.02, 0.02	0.0475	1:1	0.0475	0	1.00:1	1:1	0.0466	1.8%
0.075, 0.02	0.0597	2:1	0.0597	0	1.99:1	2:1	0.0584	2.2%
0.16, 0.02	0.0793	3:1	0.0793	0	3.02:1	3:1	0.0277	3.0%
0.26, 0.02	0.1095	4:1	0.1095	0	4.01:1	4:1	0.1055	3.8%
0.49, 0.02	0.2640	6:1	0.2640	0	5.98:1	6:1	0.2490	6.0%

Table Gated-2: Effect of Mean Switch-Over (exp. distribution) [i= 2,...,13]

	optimal pattern		mean delay approx.				lower bound	
s_1,s_i	cost	m_1:m_i	cost	%err	f_1:f_i	m_1:m_i	cost	diff.
1/36, 1	23.16	6:1	23.16	0	6:1	6:1	21.42	8.1%
1/16, 1	23.45	4:1	23.45	0	4:1	4:1	21.71	8.0%
1/9, 1	23.75	3:1	23.75	0	3:1	3:1	22.01	7.9%
1/4, 1	24.34	2:1	24.34	0	2:1	2:1	22.61	7.7%
1, 1	26.20	1:1	26.20	0	1:1	1:1	24.45	7.1%
4, 1	30.42	1:2	30.42	0	1:2	1:2	28.35	7.3%
9, 1	35.51	1:3	35.51	0	1:3	1:3	32.54	9.1%
16, 1	41.69	1:4	41.69	0	1:4	1:4	37.02	12.6%

Table Gated-3: Effect of Cost Parameters [i= 2,...,13]

	optimal pattern		mean delay approx.				lower bound	
c_1,c_i	cost	m_1:m_i	cost	%err	f_1:f_i	m_1:m_i	cost	diff.
1/16, 1	23.44	1:3	23.53	0.4	1:4	1:4	21.71	8.0%
1/9, 1	23.70	1:3	23.71	0.04	1:3	1:3	22.01	7.7%
1/4, 1	24.27	1:2	24.27	0	1:2	1:2	22.61	7.3%
1, 1	26.20	1:1	26.20	0	1:1	1:1	24.45	7.1%
4, 1	30.35	2:1	30.35	0	2:1	2:1	28.36	7.0%
9, 1	34.83	3:1	34.83	0	3:1	3:1	32.56	7.0%
16, 1	39.64	4:1	39.64	0	4:1	4:1	37.05	7.0%

column which contains the (varying) mean of the switch-over periods.

Case Gated-3: Effect of Cost Parameters This case is similar to Case Gated-1 and Gated-2, but here we perturb the cost parameters. The identical parameters are: B_i is deterministic with mean 1, λ_i is equal to 0.06, and S_i is exponential with mean $s_i = 1$. The cost parameter of the 12 identical stations is set to $c_i = 1$ and that of station 1 varies: $c_1 = 1/16$, 1/9, 1/4, 1, 4, 9, 16. The visit frequency ratios examined (in order to find the optimum) are: 1:4, 2:7, 1:3, 2:5, 1:2, 2:3, 1:1, 4:3, 3:2, 2:1, 12:5, 3:1, 7:2, 4:1, 5:1, 6:1 and 12:1. Table Gated-3 depicts this case and its structure is similar to that of the previous tables.

Case Gated-4: Effect of Service Time In this case we consider a system similar to that of the previous cases and perturb the mean value of the service times. The identical parameters are: λ_i is equal to 0.028, S_i is exponential with mean $s_i = 1$, and $c_i = 1$. The service times of the 12 identical stations are deterministic with mean 1 and the mean service time at station 1 varies: 10, 20 (deterministic). To explore the effect of the Golden Ratio procedure we use it, in this case, for the determination of the visit pattern. The visit frequency ratios examined (in order to find the optimum) are: 1:3, 1:2, 2:3, 3:4, 4:5, 5:6, 10:11, 1:1, 11:10, 6:5, 5:4, 4:3, 3:2 and 2:1. The results are provided in Table Gated-4.

Case Gated-5: Effect of System Size In this case we examine whether the system size (i.e., the number of queues) significantly affects the quality of the approximation. We therefore examine a case similar to Case Gated-1 but in which the number of identical stations is 3 (rather than 12 there). The identical parameters are: B_i is deterministic with mean 1, S_i is exponential with mean 1 and c_i is constant: $c_i = 0.02$. The arrival rates of both the 3 identical stations and the single station vary as depicted in the table, in a way that the total utilization ρ remains relatively high (between 0.7 and 0.91). The construction of the polling pattern is done manually in this case. The visit frequency ratios searched are: 1:5, 1:4, 2:7, 1:3, 2:5, 1:2, 4:6, 3:4, 1:1, 3:2, 9:5, 2:1, 5:2, 3:1, 7:2 and 4:1. The case is presented in Table Gated-5.

Table Gated-4: Effect of Service Times [$i = 2,\ldots,13$]

	optimal pattern		mean delay approx.				lower bound	
β_1,β_i	cost	$m_1{:}m_i$	cost	%err	$f_1{:}f_i$	$m_1{:}m_i$	cost	diff.
10, 1	7.920	1:1	8.742	10.3	1.12:1	11:10	6.507	21.7%
20, 1	39.95	1:1	46.11	15.4	1.22:1	5:4	24.59	62.5%

Table Gated-5: Effect of Arrival Rates (Small System) [$i = 2,\ldots,4$]

	optimal pattern		mean delay approx.				lower bound	
λ_1,λ_i	cost	$m_1{:}m_i$	cost	%err	$f_1{:}f_i$	$m_1{:}m_i$	cost	diff.
0.02, 0.255	0.1945	1:4	0.1945	0	1:3.96	1:4	0.1600	21.5%
0.03, 0.221	0.1210	1:3	0.1210	0	1:2.96	1:3	0.0998	21.2%
0.07, 0.240	0.2122	1:2	0.2122	0	1:1.99	1:2	0.1761	22.2%
0.22, 0.220	0.4311	1:1	0.4311	0	1:1.00	1:1	0.3603	19.6%
0.42, 0.130	0.2381	2:1	0.2381	0	2.01:1	2:1	0.1980	20.2%
0.61, 0.100	0.5393	3:1	0.5393	0	2.99:1	3:1	0.4481	20.3%

Table Gated-6: Effect of Various Parameters [$i = 2,\ldots,4$]

	optimal pattern		mean delay approx.				lower bound	
parameters	cost	$m_1{:}m_i$	cost	%err	$f_1{:}f_i$	$m_1{:}m_i$	cost	diff.
$\lambda_1=0.40,\ \beta_1=1,\ s_1=3,\ c_1=0.02$ $\lambda_i=0.16,\ \beta_i=1,\ s_i=1,\ c_i=0.02$	0.6405	1:1	0.6405	0	1.003:1	1:1	0.5619	14.3%
$\lambda_1=0.34,\ \beta_1=1,\ s_1=0.5,\ c_1=0.02$ $\lambda_i=0.19,\ \beta_i=1,\ s_i=1,\ c_i=0.02$	0.5052	2:1	0.5052	0	2.016:1	2:1	0.4075	19.3%
$\lambda_1=0.10,\ \beta_1=1,\ s_1=3,\ c_1=0.02$ $\lambda_i=0.26,\ \beta_i=1,\ s_i=1,\ c_i=0.02$	0.5238	1:3	0.5238	0	1:2.99	1:3	0.4404	15.9%
$\lambda_1=0.45,\ \beta_1=1,\ s_1=1,\ c_1=0.0013$ $\lambda_i=0.15,\ \beta_i=1,\ s_i=1,\ c_i=0.02$	0.2653	3:4	0.2731	2.9	1:2.02	1:2	0.2118	20.3%

Case Gated-6: Effect of Varying Several Parameters While in the previous cases we studied the effect of perturbing each parameter individually, here we examine the effect of varying several parameters concurrently. We again consider a system with three identical stations and one different station. The parameters of the four subcases are given in Table Gated-6. The visit patterns here are determined manually. The visit frequency ratios examined are: 1:5, 1:4, 2:7, 1:3, 2:5, 1:2, 4:6, 3:4, 1:1, 3:2, 9:5, 2:1, 5:2, 3:1 and 4:1. The service times in all cases are deterministic and the switch-over periods are exponential.

Limited-1 Service System

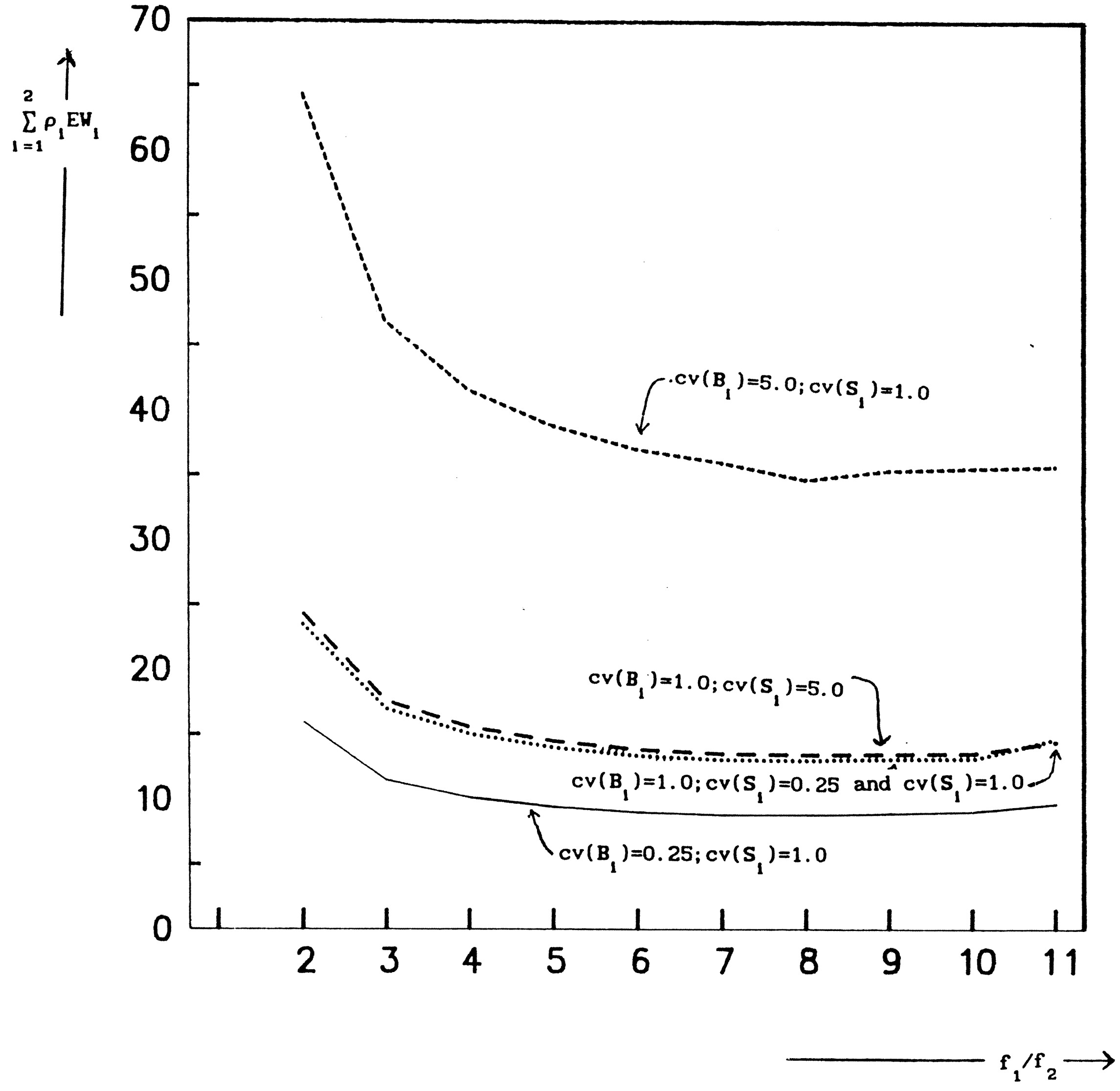
An exact analysis of limited-1 systems is only known for the cyclic fully symmetric case and the cyclic two-queue case, but Blanc's [1990a,b] power series algorithm allows the numerical evaluation of mean waiting times in limited-1 systems with a polling table. Unfortunately, the computational complexity of the algorithm in its present form is such that it is only possible to calculate mean waiting times with high accuracy when the polling table is rather small (or the traffic low). Therefore we have mainly restricted ourselves to two-queue and three-queue models. Because consideration of limited-1 service and of arbitrary cost factors c_i are both new elements in our study, we have considered a large number of examples, and in particular we have varied the cost factors extensively.

The only service time and switchover time distributions under consideration are negative exponential ones. Our experience with gated and exhaustive service (cf. Boxma et al. [1990a,b] and the cases considered earlier in this section) suggests that the optimal visit ratios for the polling tables are quite insensitive to the choice of those distributions. This is supported by some experiments for limited-1 service systems with different coefficients of variation for the service time and switchover time distributions. The results of those experiments are displayed in Figure 1.

In all examples we present the results obtained by both the mean delay approximation and the lower bound approach. The 'optimal' pattern has been determined by examining all visit number vectors $(m_1, \ldots, m_N)$ in a wide range, and all possible tables with given visit numbers. For the approxi-

Figure 1
A two-queue case: influence of the coefficient of variation of service times and switchover times.

$\lambda_1=0.765, \lambda_2=0.085; \beta_1=\beta_2=1.0; s_1=s_2=0.1.$



mations, the Golden Ratio procedure is used to determine the polling table that is most likely to be the best among all tables with given visit numbers. A detailed description of the cases is next given.

Case 1, Limited-1: This two-queue case consists of three parts, with $\beta_1 = \beta_2 = 1.0$, 0.5 and 0.1 respectively. The first part concerns a very high traffic load - certainly in view of the additional ergodicity condition that is required for stations with limited-1 service.

Case 2, Limited-1: This two-queue case also consists of three parts. The only difference between the first two parts is that $\beta_1 = \beta_2$ in part 1 and $\beta_1 = 9\beta_2$ in part 2; parts 2 and 3 only differ in the values of the arrival rates.

Case 3, Limited-1: In this two-queue case with heavy traffic and a slight asymmetry in the arrival rates and the switchover times, c_2 is varied while c_1 is being kept fixed. c_2 is chosen such that 'nice' visit ratios result, so that rounding-off errors in the visit numbers do not occur. The last two subcases are slightly different. In the 10th, c_i equals $1/(\lambda_1 + \lambda_2)$ so that the objective function equals the mean waiting time of an arbitrary customer. In the 11th, the objective function equals the sum of the two mean waiting times.

Cases 4, 6, 7, Limited-1: These are similar to case 1, but allow 3, 5 and 10 stations respectively.

Case 5, Limited-1: This case is similar to case 2, but considers 3 stations.

Figure 1: As remarked above, in the Limited-1 tables only exponential service time and switchover time distributions are considered. In Figure 1, that relates to the very first subcase of Table 1 Limited-1, we test the influence of smaller and larger coefficients of variation of the service times and switchover times on the optimal visit ratios.

A Mixture of Exhaustive and Limited-1 Service

We consider one two-queue model with exhaustive service at Q_1 and limited-1 service at Q_2. We have only developed the mean delay approximation for this mixture. It is tested in Table 1 Exhaustive/Limited-1, that consists of three parts. Parts 1 and 2 only differ in the values of the mean service times, whereas parts 1 and 3 only differ in the values of the mean switchover times.

Table 1 Limited-1
A two-queue case: asymmetric arrival rates, various traffic loads.

$\lambda_1=0.765, \lambda_2=0.085; \beta_1=\beta_2; s_1=s_2=0.1.$

	$c_1;c_2$	$\beta_1=\beta_2=1.0$									
		optimal pattern		mean delay approximation				lower bound approximation			
		cost	$m_1:m_2$	cost	%err	f_1/f_2	$m_1:m_2$	cost	%err	f_1/f_2	$m_1:m_2$
1	1.0;1.0	13.0000	8:1	13.0871	0.7	7.383	7:1	13.0000	0.0	8.251	8:1
2	1.0;0.1	8.9641	13:1	9.0418	0.1	11.716	12:1	9.0418	0.1	12.447	12:1
3	1.0;0.2	9.6881	12:1	10.2452	5.8	10.480	10:1	10.1405	4.7	11.303	11:1
4	1.0;0.5	11.3294	10:1	11.7283	3.5	8.718	9:1	11.3294	0.0	9.600	10:1
5	1.0;2.0	14.2351	6:1	14.2351	0.0	6.131	6:1	14.3242	0.6	6.937	7:1
6	1.0;5.0	16.2880	9:2	16.2880	0.0	4.699	9:2	16.5205	1.4	5.372	11:2
7	1.0;10.0	18.8084	11:3	18.9345	0.7	3.818	4:1	18.8399	0.2	4.374	9:2

	$c_1;c_2$	$\beta_1=\beta_2=0.5$									
		optimal pattern		mean delay approximation				lower bound approximation			
		cost	$m_1:m_2$	cost	%err	f_1/f_2	$m_1:m_2$	cost	%err	f_1/f_2	$m_1:m_2$
1	1.0;1.0	0.3034	9:2	0.3043	0.3	4.135	4:1	0.3034	0.0	4.395	9:2
2	1.0;0.1	0.4960	12:1	0.4976	0.3	11.302	11:1	0.4960	0.0	11.941	12:1
3	1.0;0.2	0.5144	10:1	0.5170	0.5	8.444	8:1	0.5157	0.3	8.945	9:1
4	1.0;0.5	0.5560	7:1	0.5573	0.2	5.648	11:2	0.5565	0.1	5.998	6:1
5	1.0;2.0	0.6846	3:1	0.6846	0.0	3.019	3:1	0.6846	0.0	3.210	3:1
6	1.0;5.0	0.8590	2:1	0.8590	0.0	1.995	2:1	0.8590	0.0	2.120	2:1
7	1.0;10.0	1.0879	1:1	1.1023	1.3	1.466	3:2	1.1023	1.3	1.556	3:2

	$c_1;c_2$	$\beta_1=\beta_2=0.1$									
		optimal pattern		mean delay approximation				lower bound approximation			
		cost	$m_1:m_2$	cost	%err	f_1/f_2	$m_1:m_2$	cost	%err	f_1/f_2	$m_1:m_2$
1	1.0;1.0	0.0149	10:3	0.0149	0.0	3.346	10:3	0.0149	0.0	3.356	10:3
2	1.0;0.1	0.1140	12:1	0.1142	0.2	9.747	10:1	0.1142	0.2	9.774	10:1
3	1.0;0.2	0.1203	8:1	0.1205	0.2	7.105	7:1	0.1205	0.2	7.124	7:1
4	1.0;0.5	0.1332	5:1	0.1334	0.2	4.637	9:2	0.1334	0.2	4.650	9:2
5	1.0;2.0	0.1733	9:4	0.1735	0.2	2.414	5:2	0.1735	0.2	2.420	5:2
6	1.0;5.0	0.2321	3:2	0.2321	0.0	1.572	3:2	0.2321	0.0	1.576	3:2
7	1.0;10.0	0.3077	1:1	0.3077	0.0	1.143	1:1	0.3077	0.0	1.146	1:1

Table 2 Limited-1

A two-queue case: effect of service times and traffic loads.

$\lambda_1=\lambda_2; s_1=s_2=0.1.$

	$c_1;c_2$	$\beta_1=\beta_2=0.1;\lambda_1=\lambda_2=0.75$									
		optimal pattern		mean delay approximation				lower bound approximation			
		cost	$m_1:m_2$	cost	%err	f_1/f_2	$m_1:m_2$	cost	%err	f_1/f_2	$m_1:m_2$
1	$\beta_1;\beta_2$	0.0354	1:1	0.0354	0.0	1.000	1:1	0.0354	0.0	1.000	1:1
2	1.0;0.1	0.1701	5:2	0.1701	0.0	2.495	5:2	0.1701	0.0	2.495	5:2
3	1.0;0.2	0.1974	2:1	0.1974	0.0	1.918	2:1	0.1974	0.0	1.918	2:1
4	1.0;0.5	0.2652	1:1	0.2658	0.2	1.329	4:3	0.2658	0.2	1.329	4:3
5	1.0;2.0	0.5304	1:1	0.5315	0.2	0.752	3:4	0.5315	0.2	0.752	3:4
6	1.0;5.0	0.9870	1:2	0.9870	0.0	0.521	1:2	0.9870	0.0	0.521	1:2
7	1.0;10.0	1.7017	2:5	1.7017	0.0	0.446	2:5	1.7017	0.0	0.446	2:5

	$c_1;c_2$	$\beta_1=0.9,\beta_2=0.1;\lambda_1=\lambda_2=0.75$									
		optimal pattern		mean delay approximation				lower bound approximation			
		cost	$m_1:m_2$	cost	%err	f_1/f_2	$m_1:m_2$	cost	%err	f_1/f_2	$m_1:m_2$
1	$\beta_1;\beta_2$	5.4318	5:4	6.3555	17.0	1.732	7:4	6.3253	16.5	1.806	9:5
2	1.0;0.1	5.9010	4:3	6.7197	13.9	1.754	7:4	6.6832	13.3	1.826	9:5
3	1.0;0.2	6.9663	1:1	7.8749	13.0	1.605	8:5	9.0072	29.3	1.687	5:3
4	1.0;0.5	8.2871	1:1	12.4104	49.8	1.392	7:5	13.4759	62.6	1.479	3:2
5	1.0;2.0	12.3174	4:5	14.8910	20.9	1.073	1:1	17.1761	39.4	1.154	8:7
6	1.0;5.0	18.2379	3:4	21.5138	18.0	0.894	8:9	28.0987	54.1	0.962	1:1
7	1.0;10.0	26.0193	2:3	27.7730	6.7	0.783	3:4	31.9763	22.9	0.840	5:6

	$c_1;c_2$	$\beta_1=0.9,\beta_2=0.1;\lambda_1=\lambda_2=0.5$									
		optimal pattern		mean delay approximation				lower bound approximation			
		cost	$m_1:m_2$	cost	%err	f_1/f_2	$m_1:m_2$	cost	%err	f_1/f_2	$m_1:m_2$
1	$\beta_1;\beta_2$	0.6660	2:1	0.6749	1.3	2.819	3:1	0.6749	1.3	2.993	3:1
2	1.0;0.1	0.7273	5:2	0.7335	0.9	2.919	3:1	0.7335	0.9	3.097	3:1
3	1.0;0.2	0.8270	3:2	0.8452	2.2	2.299	7:3	0.8528	3.1	2.453	5:2
4	1.0;0.5	1.0209	1:1	1.0963	7.4	1.633	5:3	1.0571	3.6	1.751	7:4
5	1.0;2.0	1.7232	1:2	1.8728	8.7	0.943	1:1	1.8728	8.7	1.014	1:1
6	1.0;5.0	2.8608	1:2	3.0525	6.7	0.656	2:3	3.0525	6.7	0.704	2:3
7	1.0;10.0	4.6420	1:3	4.7568	2.5	0.504	1:2	4.7568	2.5	0.539	1:2

Table 3 Limited-1

A two-queue case: effect of cost parameters.

$\beta_1 = \beta_2 = 1.0; \lambda_1=0.5, \lambda_2=0.25; s_1=0.1, s_2=0.2.$

	$c_1;c_2$	optimal pattern		mean delay approximation				lower bound approximation			
		cost	$m_1:m_2$	cost	%err	f_1/f_2	$m_1:m_2$	cost	%err	f_1/f_2	$m_1:m_2$
1	2.0;36.78	38.9073	1:1	38.9073	0.0	1.000	1:1	38.9073	0.0	1.039	1:1
2	2.0;15.18	23.2422	1:1	23.6934	1.9	1.250	5:4	24.2823	4.5	1.307	4:3
3	2.0;12.0	20.5422	6:5	21.0524	2.4	1.333	4:3	21.4379	4.4	1.400	7:5
4	2.0;7.92	16.8977	5:4	17.3376	2.6	1.500	3:2	17.3376	2.6	1.573	3:2
5	2.0;5.52	14.4463	7:5	14.8147	2.6	1.667	5:3	15.0740	4.4	1.750	7:4
6	2.0;3.0	11.2695	7:4	11.4088	1.2	2.000	2:1	11.4088	1.2	2.100	2:1
7	2.0;1.41	8.1689	11:4	8.1829	0.2	2.500	5:2	8.1818	0.2	2.626	8:3
8	2.0;0.75	6.2307	10:3	6.2325	0.3	3.000	3:1	6.2325	0.3	3.137	3:1
9	2.0;0.24	4.2279	4:1	4.2279	0.0	4.000	4:1	4.2279	0.0	4.171	4:1
10	1.33;1.33	6.3519	9:4	6.3850	0.0	2.259	9:4	6.3652	0.2	2.370	7:3
11	2.0;4.0	12.6553	8:5	12.8597	2.2	1.830	9:5	13.2744	4.9	1.926	2:1

Table 4 Limited-1

A three-queue case: asymmetric arrival rates.

$\beta_1=\beta_2=\beta_3=0.9;\lambda_1=0.531,\lambda_2=0.212,\lambda_3=0.106;s_1=s_2=s_3=0.1.$

	$c_1;c_2;c_3$	optimal pattern		mean delay approximation					lower bound approximation				
		cost	$m_1:m_2:m_3$	cost	%err	f_1/f_3	f_2/f_3	$m_1:m_2:m_3$	cost	%err	f_1/f_3	f_2/f_3	$m_1:m_2:m_3$
1	1.0;0.1;1.0	2.7881	5:1:1	2.8294	1.5	3.719	0.883	4:1:1	2.8294	1.5	3.689	0.858	4:1:1
2	1.0;5.0;1.0	7.9187	4:3:1	7.9187	0.0	3.899	3.023	4:3:1	7.9187	0.0	3.864	2.904	4:3:1
3	1.0;1.0;5.0	6.6573	2:1:1	6.6573	0.0	2.116	0.952	2:1:1	6.6573	0.0	2.094	0.916	2:1:1

Table 5 Limited-1

A three-queue case: asymmetric service times.

$\beta_1=0.8,\beta_2=\beta_3=0.1;\lambda_1=\lambda_2=\lambda_3=0.7;s_1=s_2=s_3=0.05.$

	$c_1;c_i$ $i=2,3$	optimal pattern		mean delay approximation				lower bound approximation			
		cost	$m_1:m_2$	cost	%err	f_1/f_i	$m_1:m_i$	cost	%err	f_1/f_i	$m_1:m_i$
1	3.55;1.0	11.0309	1:1	13.1019	18.8	2.0	2:1	13.1019	18.8	2.110	2:1
2	14.45;1.0	28.3342	2:1	29.5317	4.2	3.0	3:1	29.5317	4.2	3.133	3:1
3	51.84;1.0	74.2390	3:1	75.7466	2.0	4.0	4:1	75.7466	2.0	4.127	4:1

Table 6 Limited-1

A five-queue case: asymmetric arrival rates and switchover times.

$\beta_1=..=\beta_5=1.0;\lambda_1=0.35,\lambda_2=..=\lambda_5=0.1;s_1=0.1,\lambda_2=..=\lambda_5=0.05.$

	$c_1;c_i$, i=2,..,5	optimal pattern		mean delay approximation				lower bound approximation			
		cost	$m_1:m_i$	cost	%err	f_1/f_i	$m_1:m_i$	cost	%err	f_1/f_i	$m_1:m_i$
1	1.0;0.1	1.1446	7:1	1.2295	7.4	4.929	5:1	1.2295	7.4	4.734	5:1
2	1.0;0.5	2.4607	4:1	2.5496	3.6	2.645	5:2	2.5496	3.6	2.535	5:2
3	1.0;1.0	3.5176	5:2	3.6766	4.5	2.016	2:1	3.6766	4.5	1.933	2:1
4	1.0;2.0	5.1348	3:2	5.1348	0.0	1.548	3:2	5.1348	0.0	1.487	3:2

Table 7 Limited-1

A ten-queue case: asymmetric arrival rates.

$\beta_1=..=\beta_{10}=1.0;\lambda_1=0.27,\ \lambda_2=..=\lambda_{10}=0.03;s_1=..=s_{10}=0.05.$

	$c_1;c_i$, i=2,..,10	optimal pattern		mean delay approximation				lower bound approximation			
		cost	$m_1:m_i$	cost	%err	f_1/f_i	$m_1:m_i$	cost	%err	f_1/f_i	$m_1:m_i$
1	3.704;2.713	0.7455	2:1	0.7917	6.2	4.486	4:1	0.7917	6.2	4.394	4:1
2	1.0;1.0	2.4003	2:1	2.4667	2.8	3.886	4:1	2.4667	2.8	3.807	4:1

Table 1 Exhaustive/Limited-1

A two-queue case: influence of various parameters.

$\lambda_1=0.5, \lambda_2=0.1; \beta_1=\beta_2; s_1=s_2$.

		$\beta_1=\beta_2=1.0; s_1=s_2=0.05$					
	$c_1;c_2$	optimal pattern		mean delay approximation			
		cost	$m_1:m_2$	cost	%err	f_1/f_2	$m_1:m_2$
1	1.0;1.0	0.9682	1:1	0.9716	0.4	1.370	7:5
2	1.0;0.1	0.6707	3:1	0.6709	0.0	4.181	4:1
3	1.0;0.5	0.8046	1:1	0.8071	0.3	1.920	2:1
4	1.0;2.0	1.2810	1:2	1.2954	1.1	0.975	1:1

		$\beta_1=\beta_2=1.4; s_1=s_2=0.05$					
	$c_1;c_2$	optimal pattern		mean delay approximation			
		cost	$m_1:m_2$	cost	%err	f_1/f_2	$m_1:m_2$
1	1.0;1.0	4.4728	1:2	4.5011	0.6	0.843	1:1
2	1.0;0.1	2.2197	1:1	2.2334	0.6	2.500	5:2
3	1.0;0.5	3.2337	1:1	3.2337	0.0	1.180	1:1
4	1.0;2.0	6.5736	1:10	6.8813	4.7	0.600	3:5

		$\beta_1=\beta_2=1.0; s_1=s_2= 0.5$					
	$c_1;c_2$	optimal pattern		mean delay approximation			
		cost	$m_1:m_2$	cost	%err	f_1/f_2	$m_1:m_2$
1	1.0;1.0	1.6643	1:1	1.6643	0.0	1.08	1:1
2	1.0;0.1	1.0620	2:1	1.0687	0.6	2.63	5:2
3	1.0;0.5	1.3357	1:1	1.3736	2.8	1.44	3:2
4	1.0;2.0	2.2917	1:2	2.3032	0.5	0.80	4:5

8. Discussion

Below we discuss the operation rules derived in Section 5 and 6 and the numerical results reported in Section 7.

Discussion of the Operation Rules

The following observations can be made regarding the visit frequencies derived in (5.13a-c) and (6.8a-c):

1. For the exhaustive and the gated systems the rules derived by both analyses are identical and thus support each other.
2. For the limited-1 system, the rules derived by (5.13c) and (6.8c) somewhat differ from each other and thus require some discussion. Note that (5.13c) considers the second moment of the service time ($\beta_i^{(2)}$) while (6.8c) does not. To understand this examine the analysis leading to (5.13c) and observe that the mean waiting time $E\mathbf{W}_i^*$ expressed in (5.10) contains reference to $\beta_i^{(2)}$; however, (5.10) does not contain reference to $\beta_j^{(2)}$ $(j \neq i)$. This observation may suggest that (5.10) makes reference to second moment effects in an imbalanced way. Indeed, note that this imbalanced reference results from the assumption that the intervisit times $\mathbf{I}_i^*(j)$ are deterministically distributed. Thus, it seems that a more balanced analysis can be achieved by giving similar consideration to $\beta_i^{(2)}$, namely by "assuming" that $\beta_i^{(2)} = \beta_i^2$ (deterministic). This assumption will result in the following rule replacing (5.13c):

$$\text{limited}-1: \qquad m_i \leadsto \lambda_i + (1-\rho-\sum_{j=1}^{N}\lambda_j s_j)\cdot\frac{\left\{c_i\lambda_i\left[\frac{1}{1-\rho_i}\right]/s_i\right\}^{1/2}}{\sum_{j=1}^{N} s_j\left\{c_j\lambda_j\left[\frac{1}{1-\rho_j}\right]/s_j\right\}^{1/2}} \qquad (8.1)$$

   Note that now (8.1) is much more similar to (6.8c). As a matter of fact, for two queue systems, (6.8c) and (8.1) are *precisely identical.*
3. If we set $c_i = \beta_i$, then Equations (5.13a-b) and (6.8a-b) reduce to the rules derived in Boxma, Levy and Weststrate [1990a,b] for the minimization of $\sum_{i=1}^{N}\rho_i E\mathbf{W}_i$ (which is identical to the

minimization of the mean amount of work in the system).

4. In the light traffic situation $\lambda_i \to 0$, $i=1,...,N$, the visit frequencies for each service policy reduce to

$$f_i = \frac{\sqrt{c_i \lambda_i / s_i}}{\sum_{j=1}^{N} \sqrt{c_j \lambda_j / s_j}}; \tag{8.2}$$

indeed, note that (6.2a-c) now all reduce to the same approximation. On the other hand, in heavy traffic (in particular when $1-\rho-\sum \lambda_k s_k$ becomes small) the visit frequencies in the limited-1 system will be more or less linearly related to the arrival rates.

5. In the derivation of (6.2a-c) the assumption of Poisson arrivals plays a minor role. We conjecture that the visit frequencies given by (6.8a-c) give acceptable results even for non-Poisson arrival processes. This conjecture has not yet been investigated numerically. Results of Kruskal [1969] for a deterministic arrival (and service) process and exhaustive or gated service give the same visit rules as (6.8a-b) thus lending some support to the conjecture.

6. It is easily seen that the two approaches of Sections 5 and 6 can also be used for some other disciplines.

Discussion of the Numerical Results

1. The operation rules for determining the visit frequencies of the gated system seem to perform extremely well. In the wide examination reported in Section 7, the differences in the performance between the operation point predicted by these rules and the optimal point are extremely small. It seems that the only parameter for which the approximation reacts not as well is the mean service time, and when the differences in mean service times are extremely large (ratios of 1:10, 1:20) the approximation performs several percent worse than the optimal point.

2. The lower bound derived in Section 5 (see Remark 5.1) is not always tight, as revealed by the tests for the gated systems. The reader may observe that the difference between the lower bound and the "optimal" operation point is affected by the variability of the system. For example, in

Table Gated-1-Det in which both the service times and the switch-over periods are deterministic, the difference is very small. In Tables Gated-1, Gated-2 and Gated-3, in which the switch-over periods are exponential, the difference is larger (though, still quite small). Even larger differences are observed in small size systems (in which variability increases) and in systems with large variation between mean service times (Table Gated-4).

3. The two operation rules for determining the visit frequencies of the limited-1 systems (and mixtures of exhaustive and limited-1) lead to very similar predictions. These predictions are quite accurate in most cases. There is one exception: When traffic is heavy, with very asymmetric mean service times, both approaches find it difficult to give accurate predictions for (the ratios of) the mean waiting times. This is revealed by parts 2 and 3 of Table 2 Limited-1, and by Table 5 Limited-1. The worst results are contained in part 2 of Table 2. In some of these cases the objective function shows a very sensitive behaviour in the neighbourhood of the optimum. E.g., in case 5 the ratio 8:7 yields an error of 39.4%, while the ratio 6:5 results in an error of 100.9%! The less heavy traffic in part 3 reduces the errors significantly; and in Table 5 the presence of a third queue seems to lead to less sensitive behaviour of the objective function in the neighbourhood of the optimum.

4. In most other cases we have also observed a reasonably flat behaviour of the objective function in the neighbourhood of the optimum, with this function sharply rising at some distance of the optimum ratio. An extreme example is the last subcase of part 2 of Table 1 Exhaustive/Limited-1, where the minimal cost is found to be 6.5736 for a ratio of 1:10, but where the costs differ less than 5% from that value for a wide range of visit ratios.

5. The accuracy of the approximations does not appear to be very sensitive to the choice of cost parameters (see for example Table 3 Limited-1).

6. Figure 1 supports our belief that the optimal visit ratios are hardly sensitive to the choice of service time and switchover time distributions. If $c_i \equiv \beta_i$ and service is either exhaustive or gated, the pseudoconservation law for polling tables can be used to show (cf. Boxma et al. [1990b], Remark

3.3) that the optimal visit ratios are completely insensitive to the choice of service time distributions and (under certain conditions) of switchover time distributions.

Acknowledgments: We are grateful to J.P.C. Blanc and G. Choudhury for their assistance and for providing the numerical procedures and programs for the analysis of polling tables.

References

1. Baker, J.E., Rubin, I. [1987], Polling with a general-service order table. *IEEE Trans. Commun.,* Vol. COM-**35**, 283-288.
2. Blanc, J.P.C. [1990a], A numerical approach to cyclic-service queueing models, *Queueing Systems* **6**, 173-188.
3. Blanc, J.P.C. [1990b], The power-series algorithm applied to cyclic polling systems, *Report Tilburg University, FEW 445, Tilburg.*
4. Boxma, O.J. [1989], Workloads and waiting times in single-server systems with multiple customer classes, *Queueing Systems* **5**, 185-214.
5. Boxma, O.J., Groenendijk, W.P. [1987], Pseudo-conservation laws in cyclic-service systems. *J. Appl. Prob.* **24**, 949-964.
6. Boxma, O.J., Groenendijk, W.P., Weststrate, J.A. [1990], A pseudoconservation law for service systems with a polling table. *IEEE Trans. Commun.,* Vol. COM-**38**, 1865-1870.
7. Boxma, O.J., Levy, H., Weststrate, J.A. [1990a]. Optimization of polling systems. In: *Performance '90,* eds. P.J.B. King, I. Mitrani, R.J. Pooley, North-Holland Publ. Cy., Amsterdam, pp. 349-361.
8. Boxma, O.J., Levy, H., Weststrate, J.A. [1990b], Efficient visit orders for polling systems. *Report Centre for Mathematics and Computer Science, Amsterdam.*
9. Boxma, O.J., Meister, B. [1986], Waiting-time approximations for cyclic-service systems with switch-over times. *Performance Evaluation Review* **14**, 254-262.
10. Choudhury, G.L. [1989], Polling with a general service order table: gated service, *Report AT&T Bell Laboratories, Holmdel (NJ).*
11. Eisenberg, M. [1972], Queues with periodic service and changeover times. *Oper. Res.* **20**, 440-451.
12. Everitt, D.E. [1986], Simple approximations for token rings. *IEEE Trans. Commun.,* Vol. COM-**34**, 719-721.
13. Fuhrmann, S.W., Wang, Y.T. [1988], Mean waiting time approximations of cyclic service systems with limited service. In: *Performance' 87,* eds. P.-J. Courtois and G. Latouche, North Holland Publ. Cy., Amsterdam, 253-265.
14. Grillo, D. [1990], Polling mechanism models in communication systems - some application examples. In: *Stochastic Analysis of Computer and Communication Systems,* ed. H. Takagi, North-Holland Publ. Cy., Amsterdam, 659-698.
15. Groenendijk, W.P. [1989], Waiting-time approximations for cyclic-service systems with mixed service strategies. In: *Teletraffic Science for New Cost-Effective Systems, Networks and Services. Proc. 12th ITC,* ed. M. Bonatti, North-Holland Publ. Cy., Amsterdam, 1434-1441.
16. Kleinrock, L. [1976], *Queueing Systems, Vol. 2.* Wiley, New York.
17. 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.
18. Leung, K.K. [1990], Waiting time distributions for cyclic-service systems with probabilistically-limited service. *Report AT&T Bell Laboratories, Holmdel (NJ).*
19. Levy, H., Sidi, M. [1990], Polling systems: applications, modeling and optimization. *IEEE Trans. Commun.* **38**, 1750-1760.
20. Srinivasan, M.M. [1988], An approximation for mean waiting times in cyclic server systems with non-exhaustive service, *Performance Evaluation,* **8**, pp. 17-33.